THE UNIVERSITY OF CHICAGO

MAXIMIZING AND BORROWING INFORMATION IN RANDOMIZED TRIALS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PUBLIC HEALTH SCIENCES

BY
DANIEL EVAN SCHWARTZ

CHICAGO, ILLINOIS
AUGUST 2022

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my committee members Yuan Ji, Steve Raudenbush, and Jim Dignam for generously offering countless hours of their precious time to support my academic journey. I would also like to thank my friends and family, who know who they are, for making me smile and bearing with me all these years. I wouldn't be the same without you all.

ABSTRACT

To learn efficiently from randomized experiments, it is critical to understand how they may be designed and analyzed to best accumulate and interpret the statistical information that their data provide. To that end, this dissertation includes research on three important problems. In the first paper, we develop promising Bayesian uncertainty-directed (BUD) designs for faster and more informative dose-ranging clinical trials. The basic principle is to randomize new patients more often to doses that are expected to generate the most added information about the optimal dose, averaged over the posterior predictive distribution of their still unknown outcomes. This typically means assigning new patients to doses that are understudied relative to how strongly the data suggest they are optimal. We also use Bayesian model averaging of dose-response curves to robustly accelerate learning by letting each dose's effectiveness partially inform those of nearby doses. This butts against a computational challenge that has made BUDs with nontrivial data models impractical, so we develop an efficient Sequential Monte Carlo strategy to enable this appealing approach to multi-arm trial design. In the second paper, we propose a new model for borrowing from historical controls in efficacy trials. This model is called SPx ("synthetic prior with covariates") and uses carefully posed Bayesian model averaging to balance between competing philosophies about how the historical and new data are related. In simulations and a case study we show how SPx quickly distinguishes between historical data that are helpful and historical data that are misleading, leading to a smaller control group in the new trial to the extent reasonable. In the third paper, we consider the often overlooked problem that in multi-site efficacy trials there are often substantive grounds to believe that the effectiveness of each site may be related to its size or randomization ratio. We call this phenomenon endogeneity of design. We re-evaluate treatment effect estimators commonly used in practice and derive asymptotic and finite-sample results as well as run extensive simulations to characterize their performance under this more realistic assumption. In a detailed case study of a landmark

trial in education, we take a Bayesian viewpoint to evaluate the likely performance of the popular estimators in this specific setting. The implication is that endogeneity of design can significantly complicate analysis of multi-site trials, and existing methods are not well-equipped to handle this situation. For all three papers, code to reproduce the main analyses and simulations is included as supplementary files.

INTRODUCTION

This collection of papers is broadly focused on characterizing and making the most of the statistical information in randomized experiments: how it is accrued over time, how it may be supplemented by outside sources, and how well it is captured by standard methods. We primarily take a Bayesian approach to these analyses, which is the natural way to describe and quantify information about unknown quantities such as treatment effects given available data.

The setting of the first paper is a Phase 2b dose-ranging clinical trial, which seeks to identify which dose of a drug (if any) is effective enough to merit final testing in a large confirmatory trial. Here we develop a trial design using a Bayesian uncertainty-directed (BUD) randomization rule, which assigns new patients to doses in a way that seeks to quickly increase information about the optimal dose, given what we currently know about it from previous patients. This is combined with the use of Bayesian model averaging to flexibly incorporate pharmacological knowledge about the dose-response curve. In the literature, BUD trial design has not yet been used with nontrivial data models because of the inherent computational challenge, so we develop fast algorithms to overcome this barrier to making full use of the BUD principle.

In the second paper we turn our attention to early phase clinical trials of drug efficacy, where the control group has been studied in previous trials and we would like to use this information source to accelerate the new trial. The key statistical challenge is to appropriately control the degree of information borrowing so the historical data are relied upon when relevant but discounted when irrelevant. We propose the SPx model, standing for "synthetic prior with covariates," which extends existing approaches by using, again, Bayesian model averaging to account for different sources of heterogeneity between historical data and current trial data. Needless to say, we find Bayesian model averaging to be a convenient and tempting tool to balance between complexity and simplicity. We combine SPx with a simple

adaptive design and show strong performance in simulations and a case study.

The third paper considers multi-site randomized efficacy trials, especially but not exclusively in the social sciences, and reconsiders the performance of popular estimators under the more realistic assumption that site sizes and efficacy may be related. It is usually at least plausible that the size of a site (e.g. school or hospital) might be somewhat correlated with its effectiveness, which can dramatically change how accurate different estimators are. The first half of the paper presents new asymptotic theory and other analytic results about the estimators' performance, supplementing with simulations. While the estimators we compare in this paper are Frequentist (though some may be considered empirical Bayes), in the second half of the paper we compare them through a novel Bayesian lens. In the motivating case study, we apply a Bayesian model to quantify uncertainty about the key unknown parameters arising in our Frequentist analytic comparisons of the estimators (e.g. RMSE formulae). This then lets us directly describe, given the study's data, what the probable Frequentist performance (or if you prefer, Bayes risk) of the common estimators is in this setting.

CHAPTER 1

BAYESIAN UNCERTAINTY-DIRECTED DESIGNS WITH MODEL AVERAGING FOR MORE INFORMATIVE DOSE-RANGING TRIALS

1.1 Introduction

Phase 2b dose-ranging clinical trials are a critical milestone in drug development. They are when drug developers must decide if their drug is effective enough to merit running a Phase 3 trial for final approval and, if so, which dose of the drug should be used in the Phase 3 trial. Weak trial designs and analysis at Phase 2b can cause wrong and costly decisions for drug development and harm patients both in the trial and in the general population. Suboptimal designs not only make Phase 2b trials slower but also hinder decision-making about any subsequent Phase 3 trials. Undesirable outcomes include both wasting time and resources before learning that a poor drug is ineffective and ending development of drugs that are actually beneficial. This second error happens by either wrongly stopping effective and safe drugs before Phase 3 or by sending ineffective or toxic doses of otherwise good drugs to Phase 3. Recent work has shown in detail how suboptimal methods in Phase 2 can reduce the chance of running a successful Phase 3 trial because they are less likely to produce reliable evidence about if and how the Phase 3 trial should be run (Antonijevic et al. 2010). In fact, most drugs fail in the transition from Phase 2 to 3 (Hay et al. 2014), a sign of the need for stronger Phase 2 designs.

We find that, by and large, existing methods for Phase 2b trials either do not use adaptive randomization to maximize information about the trial goal or do not honestly leverage scientific beliefs about the dose-response curve. Yet these are two promising strategies to make Phase 2b trials more reliable and efficient, especially when combined.

1.1.1 Design goals and philosophy

In this work we propose a Bayesian uncertainty-directed (BUD) design, which is a design for multi-arm trials that adaptively randomizes patients while explicitly seeking to quickly increase evidence about the stated scientific goal of the trial (Ventz et al. 2018). In particular, when a new cohort of patients is enrolled in a BUD the probability that they will be randomized to each arm is proportional to the amount of information we expect to gain, given all current data, from treating them on that arm. Intuitively, we might gain more information from arms that have relatively few patients or that currently appear most promising (in trials where the goal is to find the "best" dose). However, in general which arms are most informative depends on the goal of the trial and the model, if any, that relates the arms to one another. The earliest example of this strategy we are aware of is Berry et al. (2002), but it has received little discussion until the approach was fleshed out more formally by Ventz et al. (2018) and Domenicano et al. (2019).

As opposed to quickly increasing information, most Bayesian adaptive designs for multi-arm trials are focused on maximizing benefit to the patients enrolled in the trial and explicitly randomize patients more often to the arms that currently appear most promising. These designs are often called Bayesian response-adaptive or adaptive randomization (BAR) designs, and have become increasingly common over the past two decades (Thall & Wathen 2007; Yuan & Yin 2011; Trippa et al. 2012; Yin et al. 2012; Wason & Trippa 2014).

At the same time, non-adaptive designs are still the de facto choice in Phase 2b trials, especially when coupled with the highly popular MCP-Mod approach to analysis (Bretz et al. 2005; Pinheiro et al. 2014). MCP-Mod is predominantly (though not always; Bornkamp et al. 2011) used in conjunction with a fixed design that uses equal randomization (ER) across arms. Its popularity is due in part to receiving regulatory blessing as an effective method from both the US Food and Drug Administration (LaVange & Zineh 2016) and the European Medicines Agency (CHMP 2014).

Earlier work in different settings has shown that BUD designs tend to be more efficient than BAR and ER designs, in some cases considerably so (Ventz et al. 2018). This means that BUD designs show promise to be faster (requiring fewer patients) or more informative (giving more reliable answers) than competitors, perhaps unsurprising because efficiently accruing information is a BUD design's organizing goal.

We have suggested that clinical trialists have an ethical responsibility to prioritize the scientific goals of a trial, but what about our ethical responsibility to the patients enrolled in the trial? For one, BUD designs can certainly be constrained to avoid apparently suboptimal doses, though this may somewhat slow down trials and make them less informative. Perhaps more to the point, BUD designs may actually avoid inferior doses more consistently than other designs intended for that purpose (e.g. BAR) by finding the optimal dose faster and more reliably. (See Table 2 in Ventz et al. (2018), where the BUD design tended to allocate similar numbers of patients to the best dose as did the BAR design, and the final allocation itself was less variable than with the BAR). In some sense, response-adaptive designs are greedy algorithms that may be worse at learning which doses are superior, thus harming their ability to assign patients to these doses.

1.1.2 Modeling the dose-response curve

For any Phase 2b trial design, an essential strategy for increasing efficiency is bringing in pharmacological background knowledge by modeling the dose-response curve. Intuitively this allows our inference about how effective each dose is to be informed not just by patient data from that dose, but also partially by patient data from other, and especially nearby, doses.

While we have some scientific understanding of how the true dose-response curve for any given drug may look, in all honesty we will be uncertain about the exact parametric form the curve should take (Bornkamp et al. 2011; Pinheiro et al. 2014). Our scientific expectations

about the dose-response curve (in the range of doses being studied) typically include the ideas that (1) it is continuous and smooth, (2) increases in the dose probably do not lead to reductions in the response rates (i.e. probably non-decreasing), and (3) the curve might show relatively diminished returns from small dose increases for both low doses (which may all be too low to have much effect) and high doses (which may all be close to achieving the maximum effectiveness of the drug). Beyond this, we cannot be certain that we know what parametric shape the dose-response curve follows. In truth, a simple dose-response curve with relatively few parameters is probably wrong, but it may still capture enough of the true relationship to enable a better design (compared to modeling each dose's response rate totally separately).

To this end we propose Bayesian model averaging (BMA) of simple parametric dose-response curves, which has the merit of relatively honestly describing the uncertainty in pharmacological beliefs about the shape of the dose-response curve while being relatively easy to communicate. Put colloquially, under the BMA our final estimate of the dose-response curve is a weighted average of the simple curves, weighting by how likely it is we think each simple curve is correct.

1.1.3 Computational demands

To use a BUD design along with non-conjugate data models raises computational complexity that has yet to be addressed in the existing BUD literature, so we develop fast Sequential Monte Carlo algorithms to enable this combination. This makes simulation studies of the method feasible, a necessity in a regulatory environment that requires checking Frequentist properties. The general Sequential Monte Carlo strategy we outline will likely be a helpful way forward to allow BUD designs in other settings to use non-conjugate models.

1.1.4 Specific setting and plan for the paper

To be more concrete, we consider a trial setting with a binary efficacy outcome and where the dose we would like to carry forward into Phase 3 is the minimum effective dose (MED), that is, the lowest dose with a response rate above some pre-specified threshold (deemed effective). In particular, we model the binary responses y_i for patient $i = 1, \ldots, n$ assigned to dose $d_i \in \mathcal{D}$ as

$$y_i | d_i, \theta \overset{ind}{\sim} Bern(\pi_\theta(d_i)),$$

where π_θ is the dose-response curve (with parameter θ). So the MED is defined as

$$d^* := \min\{d \in \mathcal{D} : \pi_\theta(d) \geq \pi_0\}$$

for the threshold $\pi_0 \in [0, 1]$. When no dose satisfies $\pi_\theta(d) \geq \pi_0$ we say $d^* =$ "does not exist". We consider trials with a moderate number of fixed doses (say 4-6), enough so that popular parametric dose-response models may reasonably be employed. For modeling convenience we scale the doses so that the minimum dose is 0 and the maximum dose is 1.

In this late Phase 2 setting the MED is commonly thought to be the appropriate dose to carry forward into Phase 3 because drug developers typically assume (1) that the doses included in the study all have tolerable toxicity levels (because of earlier clinical safety studies) and (2) that the dose-toxicity curve is monotone, so the lowest effective dose is also the safest. Typically the minimum dose being studied is either a placebo or rather low dose of the drug and the maximum dose is the largest dose confidently deemed safe given earlier clinical results.

The paper proceeds as follows. In Section 1.2 we describe the proposed BUD designs in substantive and mathematical detail. Section 1.3 outlines the Bayesian model averaging strategy to incorporate common pharmacological beliefs about the dose-response curve. Section 1.4 details our computational strategy using Sequential Monte Carlo to make BUDs with

non-trivial data models feasible to use and simulate in practice. In Section 1.5 we present results from an extensive simulation study comparing the design and modeling merits of our BUD-BMA approach to alternative methods, and in Section 1.6 we conclude with directions for future research.

1.2 Adaptive Trial Design

In plain English, a BUD design consists of the following steps:

1. Precisely define the scientific goal of the trial (the parameter of interest, in our case the "minimum effective dose" or MED) and a way to measure information about that goal given a set of data (the information criterion, which is some measure of the posterior certainty about this parameter given available data).

2. For each dose we could assign the next patient cohort to, calculate how much information we expect to gain about the scientific goal of the trial by assigning the cohort to that dose beyond what we already know from current data.

3. Randomize the next patient cohort to doses proportional to (a power of) the doses' information gains.

And the design continues until the trial's maximum sample size is reached or some adaptive stopping rule is met.

The basic premise of a BUD design for dose-ranging is that as the trial unfolds and we collect data, we start to learn about the response rates and about what doses are and are not likely to turn out to be the MED. Then we can often speed up a trial and get more information by concentrating future patients on the doses that will help us learn the most. To illustrate this point, Figure 1.2.1 shows 80% posterior credible intervals of the response rate for each dose during the middle of a trial, with the darkness of the interval indicating the posterior probability that each dose is the MED. Here the MED is defined as the lowest

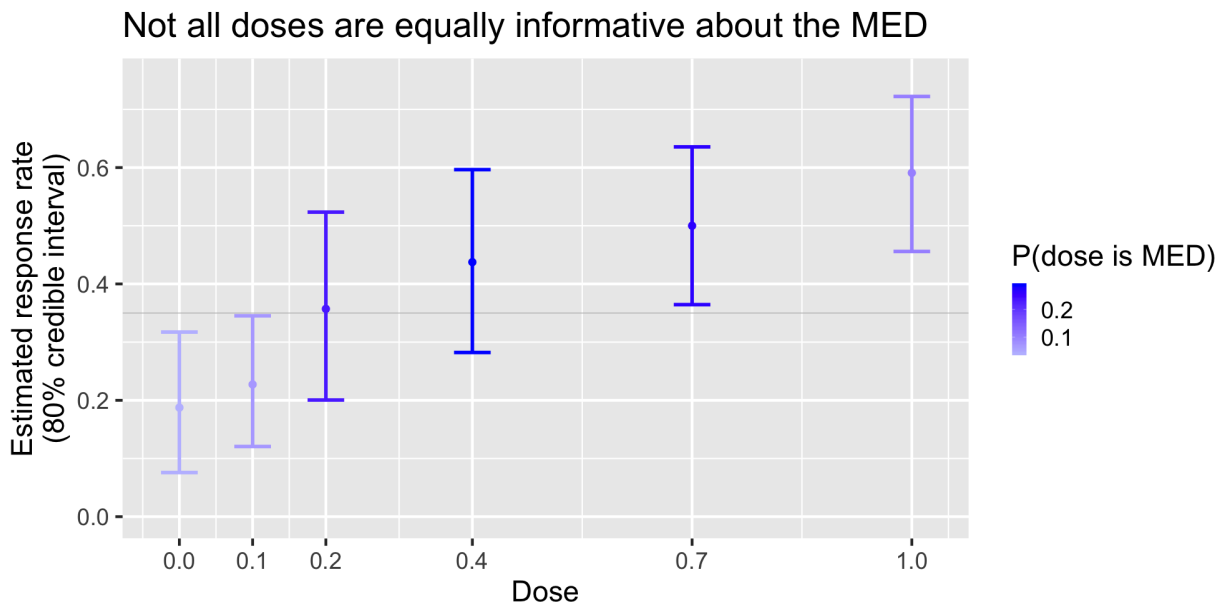

Figure 1.2.1: Illustration of which doses are likely to be most helpful to assign new patients to, given example data in the middle of a hypothetical trial. The vertical axis shows the posterior mean and 80% credible interval of each dose's response rate, and the color of each point/bar reflects the posterior probability. The extreme doses are relatively unlikely to be the MED (with a target rate of 35%, horizontal grey line) so putting the next cohort of patients on them is expected to help relatively little in narrowing down the MED.

dose (being studied) with a response rate greater than 35%. At this point in the trial we can see that the 0 and 0.1 doses probably have response rates that are too low to make them the MED and that the 1.0 dose is probably *too* effective, given that there are lower doses that seem quite possible as the MED. Because of this, it makes sense to focus new patients on the 0.2, 0.4, and 0.7 doses — the other doses are less likely to help us narrow down which dose is the MED.

Of course, the form of the model itself also impacts how informative we expect each dose to be. As a simple example, if early on in the trial (or at some later point) we are unsure if the drug is effective at *any* dose, and we have a model that says that the dose-response curve is monotone, then it makes the most sense to assign more patients to the highest (or at least higher) dose since if it is not effective enough then we know that none is (and the MED does not exist).

To explain the expected information gain, note that we can think of it as asking "if we assign the next cohort to this dose and saw these outcomes for them, what would our new posterior be and how much information (certainty) about the MED would it have?" Since we do not know what the outcomes will be, we consider how likely we think each different set of possible outcomes is for the next cohort (assigned to this dose) given our current data (i.e. the posterior predictive probability of the possible new data), and use that to average over the different amounts of information we would have for each possible new data for this next cohort. That average is the posterior predictive expectation of the information we will have after treating the next cohort at this dose, and the difference between it and the current information gives us the posterior predictive information gain from assigning the next cohort to this dose.

1.2.1 The BUD design, mathematically

We first introduce some notation. Let Σ_t be all the data (outcomes y_i and doses d_i for each patient i) up to patient cohort t. Let $\mathcal{D}$ be the set of doses being studied. Let $\Sigma_{t+1,d} := (\Sigma_t, y, d)$ be the full data up to time $t+1$ if cohort $t+1$ is assigned to dose d and has outcome vector y. Also let d^* be the minimum effective dose (MED), that is, the lowest dose among those being studied with a response rate greater than some predefined target rate. The MED does not exist if none of the doses being studied has a response rate greater than the target rate, so "does not exist" is a value d^* may take and thus the parameter space of d^* is $\mathcal{D}^+ := \mathcal{D} \cup \{\text{"does not exist"}\}$.

Then mathematically, the BUD design includes the following steps:

1. *Define the target of inference and the information metric.*

The target of inference is d^*, the MED. The current information is

$$\begin{aligned} I(\Sigma_t) &:= f(P(d^*|\Sigma_t)) \\ &= \sum_{d \in \mathcal{D}^+} P(d^* = d|\Sigma_t) \log P(d^* = d|\Sigma_t) \end{aligned}$$

where f is generically some functional of a distribution that measures how concentrated it is (i.e. close to degenerate) and $P(d^*|\Sigma_t)$ is the posterior of the MED given all the data up to cohort t. In our case we take f to be the negative entropy, partially since it is defined no matter the parameter space of d^* (which is not strictly numerical).

2. *Calculate the posterior predictive expected information gain if cohort $t+1$ is assigned to dose d.*

   Let the expected information gain be

$$\Delta_{t+1,d} := E_{y|\Sigma_t,d}\left[I(\Sigma_{t+1,d})|\Sigma_t\right] - I(\Sigma_t)$$

   where we take expectation over the posterior predictive distribution of the responses for cohort $t+1$, y, given that it is assigned to dose d and Σ_t.

3. *Randomize and treat the next cohort $t+1$.*

   Assign the patients in cohort $t+1$ to dose d with probability

$$P(d_{t+1} = d) \propto \Delta_{t+1,d}^h$$

   where h is some positive power. Typically we default to $h = 1$ (Ventz et al. 2018).

Patients can be assigned individually (i.e. a cohort size of 1), though assigning larger cohorts may be easier computationally and logistically.

1.2.2 Early stopping

The trial will continue until it reaches a pre-specified maximum sample size, or if early stopping is allowed, it will stop when the early stopping rule is first met. The common, intuitive Bayesian approach is to stop the trial when

$$P(d^* = d|\Sigma_t) \geq c$$

for some $d \in \mathcal{D}^+$ and threshold c (e.g. 95%). In other words, stop the trial when the posterior is sufficiently certain that it knows which dose is the MED.

In the BUD design, we may also want to stop trials early for purely computational reasons. As the posterior of the MED becomes very concentrated on a single dose, the information gains will become smaller and thus estimates of them can become too noisy, leading to unstable allocations. Stabilizing the allocations would require much more computation, and is probably not worthwhile if the data are really so blatant that the posterior is extremely confident. For this reason in our implementation of BUD designs we always stop trials when $P(d^* = d|\Sigma_t)$ exceeds 99% for some $d \in \mathcal{D}^+$.

1.3 Model

Our modeling strategy is based on Bayesian model averaging (BMA) of multiple parametric dose-response curves. This model suggests that the we believe the data to have been generated by one of a finite set of candidate curves but that we are uncertain as to which one, so inferences get averaged over each of the models. Formally, we

1. Specify a set of possible models $\mathcal{M}$ and suppose that the data Σ_t were generated by some unknown model $M \in \mathcal{M}$. In particular, these models are the dose-response curves $\pi^m_{\theta_m}$ (where θ_m is the parameter, possibly multivariate, for model m) for the Bernoulli outcome. This gives us the model-specific likelihoods $p(\Sigma_t|M = m, \theta_m)$.

2. Specify priors over both the models, $p(M = m)$, and the parameters in each model, $p(\theta_M|M = m)$.

3. Find the posterior probability that the data came from each model $m \in \mathcal{M}$,

$$p(M = m|\Sigma_t) \propto p(\Sigma_t|M = m)p(M = m),$$

where $p(\Sigma_t|M = m) = \int p(\Sigma_t|\theta_m, M = m)p(\theta_m|M = m)d\theta_m$ is the marginal likelihood (or evidence) of model m.

4. Then BMA inferences come from averaging all model-specific inferences (such as estimates of the response rates for each dose) over the posterior probability of each model. For example, the BMA posterior of the MED d^* is

$$p(d^* = d|\Sigma_t) = \sum_{m \in \mathcal{M}} p(M = m|\Sigma_t)p(d^* = d|M = m, \Sigma_t).$$

BMA is an attractive strategy to account for uncertainty in the shape of the dose-response curve for multiple reasons. First, it lets us directly include different scientific theories about how the drug affects patients, and because of this it is relatively straightforward to specify priors over the models and within each of them. Furthermore, statistical theory has long suggested that Bayesian model averaging gives superior predictions to Bayesian model selection (i.e., basing inferences entirely on the model with the highest posterior probability) (Wasserman 2000), which is closely tied to our problem — predicting the response rates at each dose to find the MED. Similar findings have also emerged in non-Bayesian comparisons of model averaging and model selection (Schorning et al. 2016), and notably in the dose-finding setting where the original and common usage of MCP-MOD relies on model selection (Bretz et al. 2005; Pinheiro et al. 2014). Finally, there may be little gain (or actual harm) from nonparametric strategies because although our setting has a decent number of doses

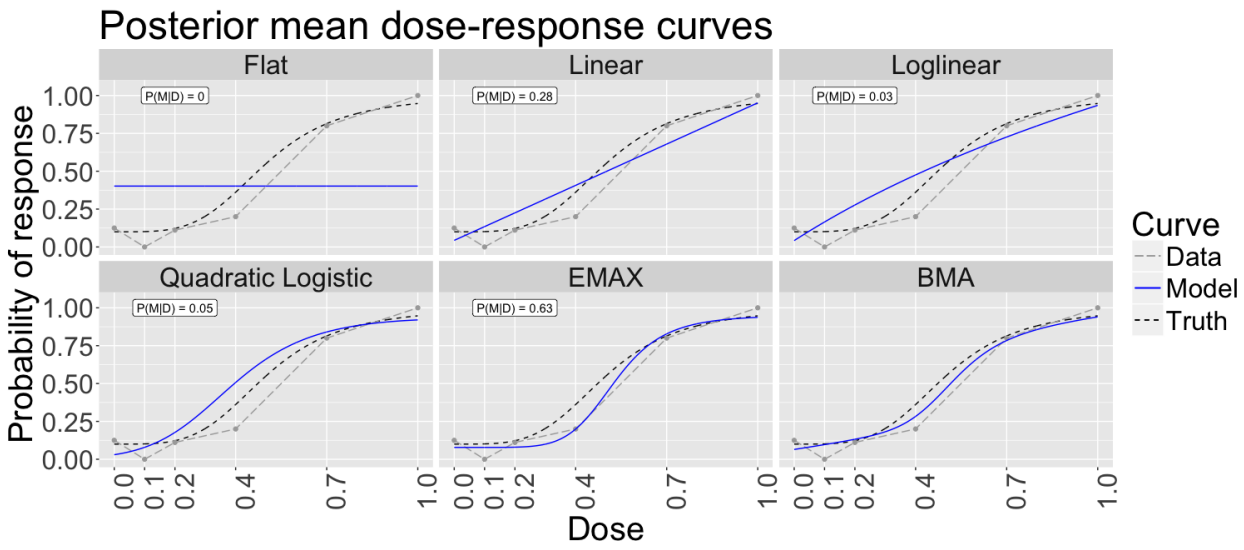

Figure 1.3.1: Example analysis under the Bayesian model averaging (BMA) of simple parametric dose-response curves. Each panel shows, for a different model, the posterior mean dose-response curve (blue) with doses along the x-axis and response rate along the x-axis. The BMA panel shows the average of the other five, weighted by the posterior probability that each is correct ($P(M|D)$). The observed data (long-dashed grey line) and true data-generating curve (short-dashed black curve) are the same in each panel and shown for reference.

compared to many clinical trials (6), we still do not have enough design points to estimate particularly nuanced curves.

In particular, we propose a set of 5 candidate dose-response curves and use these for the rest of the project, though of course this set should be tailored to each given setting. Figure 1.3.1 shows each of these curves, estimated from an example data set, as well as the BMA curve (averaging these 5 over the posterior distribution of the identity of the "true" curve).

The specific functional forms and priors (independent, except where specified condition-

ally):

$$
\begin{aligned}
\underline{\text{Flat}}: \ & \pi_\theta(d) = p \\
& p \sim Unif(0,1) \\
\underline{\text{Linear}}: \ & \pi_\theta(d) = \alpha + \beta d \\
& \alpha \sim Beta(0.1,3) \\
& \beta|\alpha \sim Unif(0,1-\alpha) \\
\underline{\text{Loglinear}}: \ & \pi_\theta(d) = \alpha + \beta \log_2(d+1) \\
& \alpha \sim Beta(0.1,3) \\
& \beta|\alpha \sim Unif(0,1-\alpha) \\
\underline{\text{Quadratic Logistic}}: \ & \pi_\theta(d) = \frac{1}{1+\exp(-(ad^2+bd+c))} \\
& \pi_{placebo} := \pi_\theta(0),\ \pi_{high} := \pi_\theta(1),\ x_{ext} := \frac{-b}{2a} \\
& \pi_{placebo} \sim Beta(0.1,3) \\
& \pi_{high} \sim Unif(0,1) \\
& x_{ext} \sim N(0.75, 0.29^2) \\
\underline{\text{EMAX}}: \ & \pi_\theta(d) = \alpha + E_{max}\frac{d^\lambda}{d^\lambda + ED_{50}^\lambda} \\
& \alpha \sim Beta(0.1,3) \\
& E_{max}|\alpha \sim Unif(0,1-\alpha) \\
& \lambda \sim Gamma(2,2) \\
& ED_{50} \sim TN(0.5, 3^2; (0,1))
\end{aligned}
$$

where $TN(x, y; S)$ denotes a normal distribution with mean x and variance y that has been truncated to the set S.

This set of models was chosen to accommodate a range of dose-response curves of varying

complexity (the number of parameters ranges from 1 to 4) and to include the possibility that the dose-response curve is non-monotone in the range of doses being studied, with the response rate peaking before the maximum dose of 1. This last feature is built into the quadratic logistic model, where we state the prior in terms of the transformed parameters $\pi_{placebo}$ (the response rate at the 0 dose, corresponding to α in the other models), π_{high} (the response rate at the high dose of 1), and x_{ext} (the dose at which the dose-response curve has its extremum) so that there is an 80% prior chance that this model is non-monotone in the range of doses being studied.

The Bayesian model averaging procedure is also desirable since it has a penalty for unnecessary model complexity naturally built in. In particular, the posterior probability of each model is mainly controlled by the marginal likelihood of the model, which will tend to be relatively smaller for more complex models that do not explain the data significantly better. This is because in the calculation of the marginal likelihood the likelihood function must be integrated over a higher-dimensional parameter space, so unless the additional parameter substantially improves the model fit (height of the likelihood) at some values, this integral will be over more small likelihood values, dragging down the average.

1.4 Computation

1.4.1 Computational needs of the method

BUD designs are conceptually attractive but can require considerably more computation than simpler designs. Even compared to Bayesian response-adaptive randomization (BAR) designs, BUD designs must compute many more posterior distributions, which is a challenge if these posteriors are not conjugate.

The piece of the BUD that requires more computation is the expected information gain, since direct calculation of this quantity requires that we learn what the information criterion

would be if we observed each possible new data set. Since the information criterion is a summary of the posterior of the MED, we apparently need to compute this posterior for each of these hypothetical data sets in order to decide how to randomize the new patients.

In particular, to run the adaptive design we need to compute $D \cdot (n_{\text{cohort}}+1) \cdot (n_{\text{per cohort}}+1) \cdot k$ posterior distributions, where D is the number of doses being studied, n_{cohort} is the number of cohorts, $n_{\text{per cohort}}$ is the number of patients per cohort, and k is the number of dose-response models included in the BMA. For example, with 6 doses, cohorts of 5 patients, 20 cohorts, and 4 dose-response curves, this comes to $6 \cdot 21 \cdot 6 \cdot 4 = 3024$ posteriors to adaptively randomize just 100 patients.

"Brute force" independent MCMC simulation for each of these posteriors would make running an actual trial cumbersome (even though in real life computation only happens for one cohort at a time) and importantly would make simulation studies for Frequentist properties of the design impractical or infeasible in most cases. To address this issue, we make use of the fact that many of these posteriors are closely related to each other since sequentially the data they condition on only differs by a few observations.

1.4.2 Sequential Monte Carlo

Our approach is based on a Sequential Monte Carlo (SMC) algorithm (Chopin & Papaspiliopoulos 2020). The basic idea is to sequentially use importance sampling to reweight the posterior at time t to reflect the posterior at time $t+1$.

In general, importance sampling is a method used to compute the mean (of a function h) for some "target" distribution f when it is only easy to sample directly from a "proposal" distribution g. Let $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$ denote a parameter vector for a model defined over a parameter space Θ (having dimension k). Importance sampling begins by taking a random

sample of parameter (vectors) $\theta^{(1)}, \ldots, \theta^{(J)} \overset{iid}{\sim} g$ and then using the unbiased estimator

$$\begin{aligned}\hat{\mu}^h &:= \frac{1}{J}\sum_{j=1}^{J} h(\theta^{(j)}) \cdot \frac{f(\theta^{(j)})}{g(\theta^{(j)})} \\ &= \frac{1}{J}\sum_{j=1}^{J} h(\theta^{(j)}) \cdot w_j\end{aligned}$$

where f and g are the densities of each distribution. This is just a weighted mean, where samples drawn from g are reweighted to resemble samples from f. The more similar g gets to f, the less variable the weights w_j and the more efficient $\hat{\mu}^h$ become, that is, a smaller J will be needed for a specified level of precision. The weighted sample $(\theta^{(j)}, w_j)_{j=1,\ldots,J}$ is often called a *particle system*.

It is intuitive how importance sampling might be used in our setting where the current posterior is changing step by step as new data gets added. This is the key idea, though even the relatively simple SMC approach we discuss has some further nuances. For example, we occasionally need to move the samples around so the weights don't get too extreme and variable (that is, a point that was unlikely under the posterior at time $t-5$ might be much likelier at time $t+1$ and now require an extreme weight).

1.4.3 SMC to update a single model's posterior

Here we give the SMC algorithm to update the posterior for a single dose-response curve, directly applied from Chopin (2002). It is the basic building block of our overall algorithm to compute the BMA of dose-response curves.

Let p_t be the posterior (for θ in dose-response model M, suppressed in the notation) after observing data for the first t cohorts. Let y_t be the binomial response of cohort t. Let $f(y_t|\theta^{(j)})$ be the likelihood of the new data y_t under parameter $\theta^{(j)}$. So say we have a particle system $(\theta^{(j)}, w_j)_{j=1,\ldots,J}$ for p_{t-1} — how do we update it to get a particle system

for p_t?

Algorithm 4.1 SMC to update a single curve's posterior

1. **Reweight**: Let $w_j \leftarrow w_j \cdot f(y_t|\theta^{(j)})$ for each particle $j = 1, \ldots, J$.

2. If $ESS < ESS_{threshold}$,
   (a) **Resample**: Resample $\theta^{(j)}$ w.p. $\propto w_j$ and let $w_j \leftarrow 1$ for each particle $j = 1, \ldots, J$.
   (b) **Move**: Move $\theta^{(j)}$ by some transition kernel K_t with stationary distribution p_t.

3. **Compute posterior**: Now $(\theta^{(j)}, w_j)_{1,\ldots,J}$ is a particle system for p_t, so we have unbiased estimates of $E_{p_t}(h(\theta))$ given by

$$\hat{\mu}_t^h := \sum_{j=1}^{J} w_j h(\theta^{(j)}).$$

The ESS, or effective sample size, of a particle system is the number of independent draws from p_t that would give us an estimator with the same variance as what we get from the particle system. It can be estimated as a simple function of the weights (Chopin 2002). By monitoring the ESS we can make sure that the particle system is always at least as efficient as an independent sample of the desired size ($ESS_{threshold}$).

Because the posteriors p_{t-1} and p_t have the same prior and because the data for each patient is independent in our model's likelihood, the weight for particle j only needs to change by a factor of $f(y_t|\theta^{(j)})$, i.e. the likelihood of the new data under $\theta^{(j)}$.

When the particle system is becoming impoverished (the ESS is too small), we would like to refresh it so that we can get more precise (but still unbiased) estimates for the new posterior (particularly, so our information criterion calculations are stable). The resampling step gives us the chance of discarding particles that are relatively unlikely in p_t (those with especially small weights), and after this the particle system $(\theta^{(j)}, w_j)_{j=1,\ldots,J}$, where the weights are now constant, is another valid particle system for p_t.

The move step gives us the chance to include new particles that were not previously in the system, especially those that are relatively likelier under p_t than the current particles. By definition of K_t having stationary distribution p_t, after the move step $(\theta^{(j)}, w_j)$ is still a valid particle system for p_t, and now one that should estimate its features more accurately. In practice, for K_t we use a single Metropolis-Hastings step where the proposal distribution is multivariate normal with mean vector and covariance matrix being the weighted sample mean and weighted sample covariance of the current particle system.

We stress here that the parameterization of the model can be critical to the accuracy and reliability of the posterior computation, especially after multiple updates, as Chopin (2002) mentioned briefly. In particular, given the normal proposals in the move step the SMC should be done on a parameter space Θ that will have relatively smooth posterior density over $\mathbb{R}^k$ (where k is the dimension of the model) and no tall spikes (especially if there are multiple modes). For example, parameterizing the quadratic logistic model's curvature parameter x_{ext} by $\theta_3 = (x_{ext} - 0.5)^3$ can produce a bimodal posterior with one mode being a spike. Unfortunately this causes the sequence of SMC estimated posteriors to degrade to biased representations of the actual posteriors after a moderate number of cohorts. In this case the fixed scale of the move step's normal proposals, which is determined by the overall variance of the current particle system, make the spiked region both hard to move to and move out of. This leads to unstable and worsening estimates of the size of the spike over time, meaning that the computed posterior for this parameter can be very wrong.

1.4.4 Algorithm to run a full trial

The SMC algorithm just given is the key tool required to get the necessary posterior inferences for the full BMA as the trial proceeds. We do this by maintaining a particle system for the current posterior of each of the dose-response curves and update each at each step of the trial. To find the BMA weights (the posterior probability of each dose-response curve) we

just need the marginal likelihood for each dose-response curve model, which can be estimated as a simple function of the particle system weights (Drovandi et al. 2014).

Each dose-response curve posterior, summarized by its current particle system, implies a posterior for the MED, that is, the posterior probability that the MED is equal to each dose being studied, given the type of dose-response curve. We can get each curve's implied MED posterior from its particle system (since we want to estimate the posterior expectation of the function $h(\theta) := \mathbb{1}\{\theta : \text{dose d is the MED}\}$ for each d). And the overall, BMA, posterior for the MED is simply the average of these posteriors, averaged by the BMA weights we have already calculated.

Then the overall algorithm for running a trial is as follows:

Algorithm 4.2 BUD-BMA dose-ranging trial

---

1. **Run-in**: Start trial with a short period of equal allocation.
2. **BUD**:
   (a) Calculate the next cohort's BUD allocation by finding each dose's expected information gain, for all the cohort's possible outcomes:
      i. do SMC to get the posterior for each dose-response curve in the BMA (Algorithm 1)
      ii. calculate the BMA posterior for the MED and its information

      and average these information criteria over the posterior predictive distribution.
   (b) Randomize next cohort to their dose.
   (c) After observing new outcomes, return to 1 to allocate the following cohort.
3. **Stop Trial**: When max sample size is reached or when there's high posterior certainty about the MED.

To avoid excessive computation, we approximate the information gains by ignoring the next cohort's possible outcomes that are extremely unlikely (such as 5 out of 5 responders when the current posterior predictive probability of response for this dose is 20%; such a scenario would account for 0.03% of the expected information gain). We have found that this

can cut down computation time by nearly half without noticeably changing the allocations chosen in the BUD.

1.5 Simulation Study

To better understand the performance of BUD-BMA and the underlying design and modeling choices, we conducted extensive simulations organized around four primary questions. First, how does the BUD strategy perform compared to other patient randomization schemes for dose-ranging trials? Second, how do different choices of the dose-response curve impact results? Third, what are the effects of adaptive trial stopping and the number of doses being studied? And fourth, how does BUD-BMA compare to other dose-ranging methods most popular in clinical trial practice? To measure performance, we focus primarily on Frequentist rates of identifying the MED and average total trial size but also consider average patient outcomes as a measure of benefit to trial participants.

In our simulations we consider trials run with each combination of four randomization rules and three modeling strategies, plus one popular Frequentist method, leading to 13 total combinations. The Frequentist method is MCP-Mod with equal number of patients randomized to each dose (ER-MCP-Mod).

For the randomization rules we consider BUD as described in Section 1.2, along with two more familiar Bayesian adaptive randomization rules as well as equal randomization. The first Bayesian alternative is response-adaptive randomization, where patients are randomized to doses with probability proportional to the posterior probability that each dose is effective (i.e. has response rate greater than the target rate for the MED), and we denote this by BAR-EFF. This will put more patients on doses perceived to have the highest response rates. The second Bayesian alternative is a bit more tailored, and randomizes patients with probability proportional to the posterior probability that each dose is the MED, and we denote this by BAR-MED. Finally in the equal randomization (ER) design we randomize an

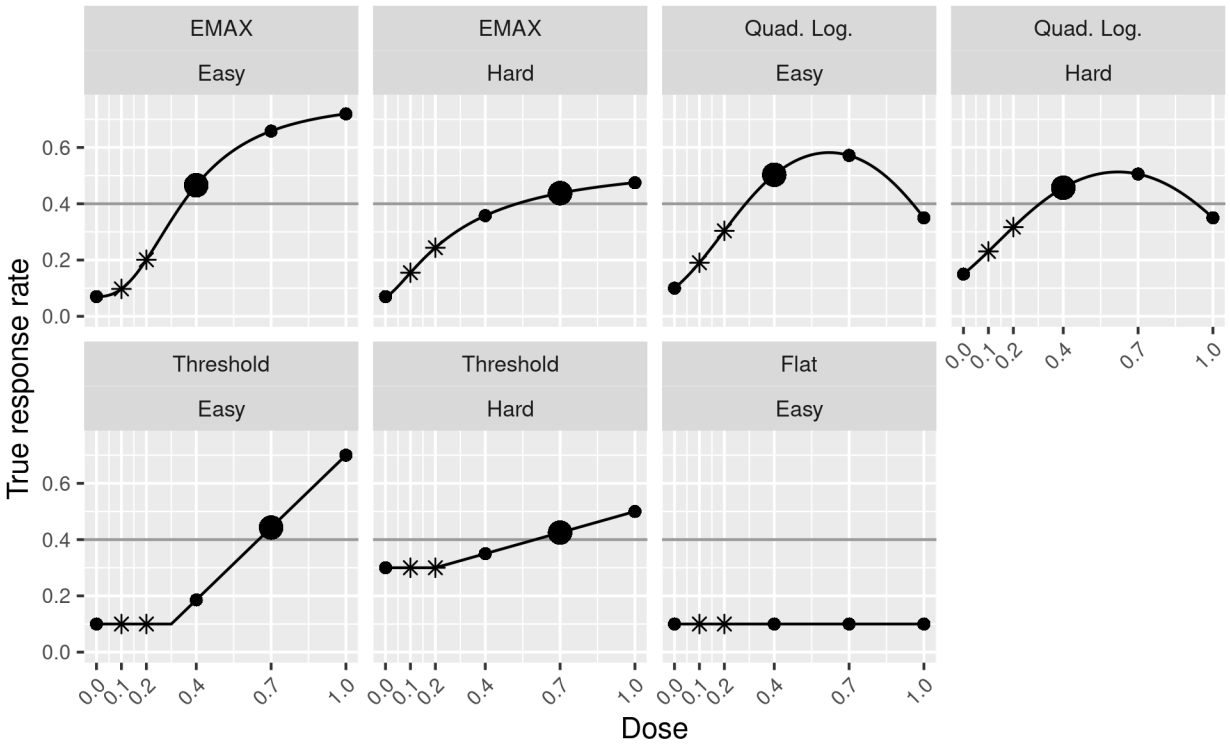

Figure 1.5.1: Data-generating (true) dose-response curves for the simulation study. The points indicate the doses being studied, with the starred points excluded in the cases with only 4 doses. The horizontal grey line denotes the 40% target rate defining the MED, and in each panel the large point denotes the MED.

equal number of patients to each dose.

In terms of specific design parameters, we also vary the total trial size and number of doses being studied. For all four designs we ran trials with a fixed total size of 250 patients, which also lets us record what would have happened had these trials stopped earlier at smaller fixed total sizes. For the designs with adaptive randomization (BUD, BAR-EFF, and BAR-MED) we also ran trials that stopped adaptively when there was 95% or higher posterior probability of knowing the MED. Across all designs we included both trials using 4 and 6 fixed doses.

The three modeling strategies include (1) the BMA of simple parametric dose-response curves described in Section 1.3, (2) the EMAX model that is used as a submodel in the BMA, and (3) an independent model that makes the doses' response rates totally unrelated

to one another by using independent Beta(1/2, 1/2) priors for each. The EMAX model, which describes four-parameter sigmoid curves, is very flexible and popular in pharmacology (Thomas 2006), but it cannot accomodate non-monotone curves and may be too large a model when there is little data and fewer doses.

The data-generating scenarios we simulate from consist of the 7 dose-response curves shown in Figure 1.5.1. Two each are EMAX curves, quadratic logistic curves, and linear threshold curves, with "easy" and "hard" cases, and there is also a flat scenario where the drug is ineffective and all doses have the same response rate. The linear threshold model, which is not included in the BMA ensemble, is flat at low doses and at some point transitions to increasing linearly over the remaining higher doses. For each scenario and method we simulated 1,000 trials.

In general we find that while no approach dominates universally across all scenarios, the BUD-BMA is overall a strong choice that is often best or close to best and that otherwise appears not to fall too far behind.

1.5.1 Choice of randomization rule

If the dose-response model is fixed, how do the different randomization rules compare to one another? An initial answer is given by Table 1.5.1, which reports method performance for trials with 6 doses and fixed trial size.

For a given modeling strategy, no matter the modeling strategy, BUD tends to handily beat or match ER and BAR-EFF at MED accuracy and average trial size. BUD also produces respectable patient outcomes compared to these designs despite that not being its stated goal, often greatly improving over ER and not falling too far of BAR-EFF (which has the main goal of maximizing patient outcomes within reason).

Compared to BAR-MED, BUD is often quite similar in MED accuracy, especially when BMA is used. At the same time it often produces smaller trials (sometimes by about 5-10

		BMA				EMAX				Independent				MCP-Mod			
Scenario	**Design**	MED	n	sd	y/n	MED	n	sd	y/n	MED	n	sd	y/n	MED	n	sd	y/n
EMAX																	
Easy	ER	0.70	250	0	37	0.83	250	0	37	0.83	250	0	37	0.67	250	0	37
	BAR-EFF	0.72	249	9	47	0.78	248	13	48	0.76	249	10	47				
	BAR-MED	0.83	245	21	43	0.85	243	22	43	0.79	246	18	41				
	BUD	0.85	241	28	42	0.93	234	39	41	0.92	228	44	35				
Hard	ER	0.78	250	0	29	0.70	250	0	29	0.49	250	0	29	0.63	250	0	29
	BAR-EFF	0.75	250	6	35	0.75	250	7	35	0.47	250	0	34				
	BAR-MED	0.79	250	0	34	0.77	250	0	34	0.53	250	2	32				
	BUD	0.79	250	6	36	0.81	250	0	36	0.56	249	9	32				
Flat																	
Easy	ER	1.00	250	0	10	1.00	250	0	10	1.00	250	0	10	1.00	250	0	10
	BAR-EFF	1.00	50	17	9	1.00	42	15	9	1.00	139	28	10				
	BAR-MED	1.00	50	17	9	1.00	43	15	9	1.00	138	29	10				
	BUD	1.00	49	15	9	1.00	40	12	9	1.00	129	24	10				
Quad. Log.																	
Easy	ER	0.88	250	0	34	0.64	250	0	30	0.80	250	0	34	0.95	250	0	33
	BAR-EFF	0.91	249	7	40	0.72	250	8	38	0.80	250	3	40				
	BAR-MED	0.95	248	10	40	0.90	250	3	40	0.87	249	8	38				
	BUD	0.96	248	10	38	0.87	249	8	38	0.95	240	30	34				
Hard	ER	0.63	250	0	34	0.53	250	0	34	0.66	250	0	34	0.74	250	0	34
	BAR-EFF	0.63	249	10	37	0.45	248	17	37	0.65	250	4	37				
	BAR-MED	0.68	249	10	38	0.64	249	12	38	0.74	250	5	36				
	BUD	0.71	249	10	37	0.56	249	14	37	0.81	248	12	34				
Threshold																	
Easy	ER	0.84	250	0	27	0.87	250	0	27	0.69	250	0	27	0.79	250	0	27
	BAR-EFF	0.86	247	13	40	0.87	246	19	40	0.67	250	3	39				
	BAR-MED	0.87	245	19	36	0.86	243	23	36	0.69	249	10	35				
	BUD	0.89	238	31	36	0.91	232	41	37	0.80	241	29	31				
Hard	ER	0.72	250	0	36	0.70	250	0	36	0.26	250	0	36	0.52	250	0	36
	BAR-EFF	0.76	250	0	40	0.74	250	0	40	0.30	250	6	38				
	BAR-MED	0.76	250	0	38	0.76	250	0	38	0.35	249	12	36				
	BUD	0.75	250	0	38	0.74	250	0	38	0.37	250	9	34				

Table 1.5.1: Performance metrics for the trial methods with fixed trial stopping at 250 patients and 6 doses under study. In particular, for each combination of design and modeling strategy we give the proportion of trials in which the MED is correctly identified (MED), mean trial size (n) and standard deviation of trial size (sd), and mean observed patient response rate as a percentage (y/n). The total trial size is not always 250 because the adaptive randomization designs were stopped early at 99% posterior confidence in the MED (see Section 1.2).

patients but other times by dozens). Figure 1.5.2 provides a broader comparison of BUD-BMA and BAR-MED-BMA. For the most part, BUD-BMA matches or moderately improves on BAR-MED in terms of MED accuracy and trial size. This is especially the case when there is a dose that is obviously *not* the MED but that is still essential to study to find the MED quickly because of the insights it provides about the underlying dose-response curve. This arises is in the easy quadratic logistic and threshold scenarios, and we return to this point in Section 1.5.4. An exception to BUD's better performance arises in the hard threshold scenario when only 4 doses are being studied. Here it puts relatively more patients at the lowest and highest dose, and slightly more often concludes that the 0.4 dose is the MED (despite in truth being ineffective).

1.5.2 Choice of dose-response model

If the randomization rule is fixed, which dose-response model performs best? Looking again to Table 1.5.1 we can see that here there is no universal winner, since model misspecification and inefficiency have different consequences depending on the underlying true dose-response curve. In general the BMA performs respectably, beating or nearly matching the next-best model (EMAX or independent) at MED accuracy and trial size in all but two scenarios.

The first case where it falters somewhat is (perhaps ironically) the "easy" EMAX scenario. Here it puts nontrivial mass on the loglinear submodel, which is biased downwards for the response rate of the true MED, the 0.4 dose, occasionally causing the BMA to lose just enough confidence in the 0.4 dose's effectiveness to instead claim that the next higher dose (0.7) is the MED. Unsurprisingly, simply using the EMAX model (which is correctly specified) does better here. The second case where the BMA struggles a bit is the hard quadratic logistic scenario, where it is more challenging to detect the non-monotonicity (since the EMAX model also fits relatively decently here) and the independent model does better.

In contrast to BMA, when the EMAX and independent models do poorly their perfor-

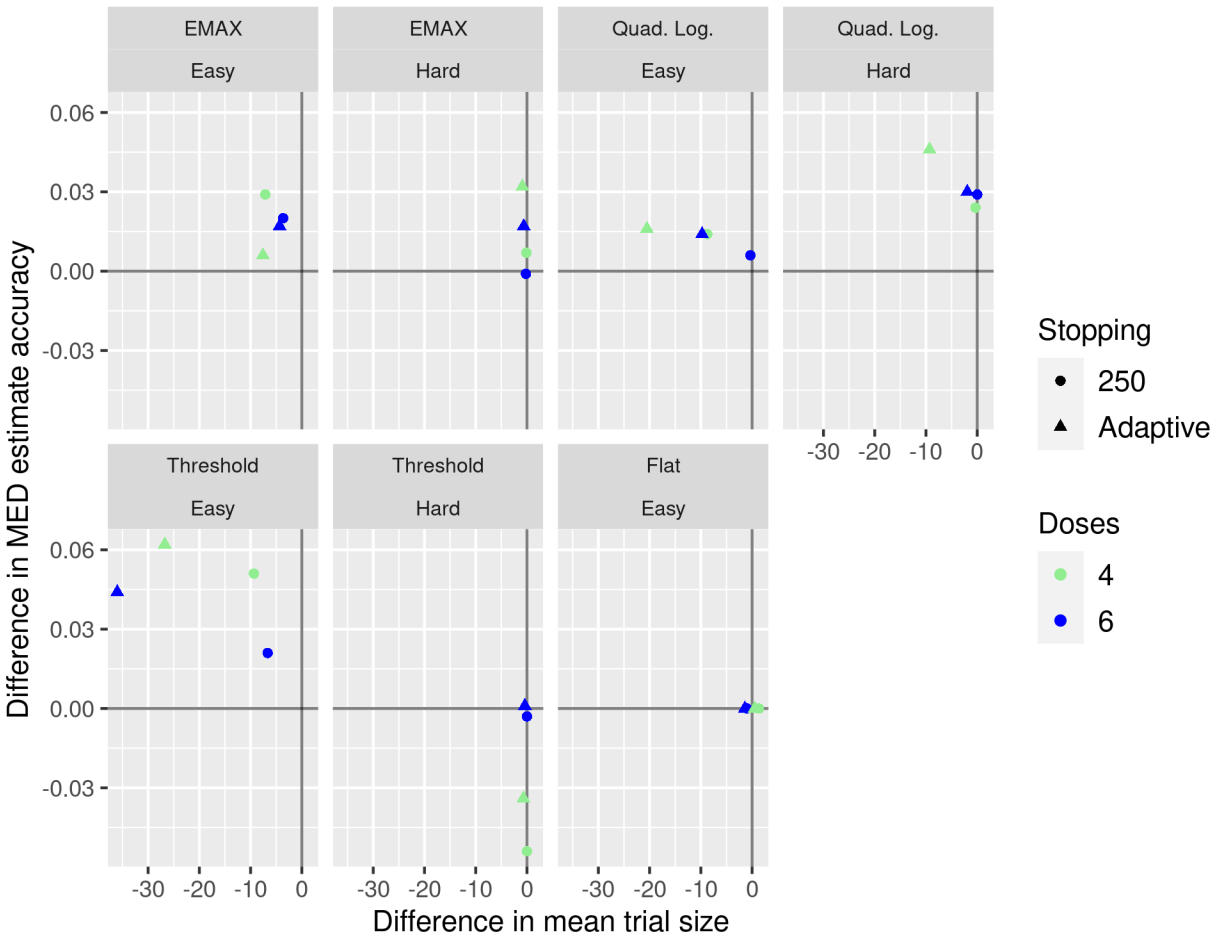

Figure 1.5.2: Difference in BUD-BMA and BAR-MED-BMA accuracy and trial size. For both accuracy and trial size the ordering is BUD minus BAR-MED.

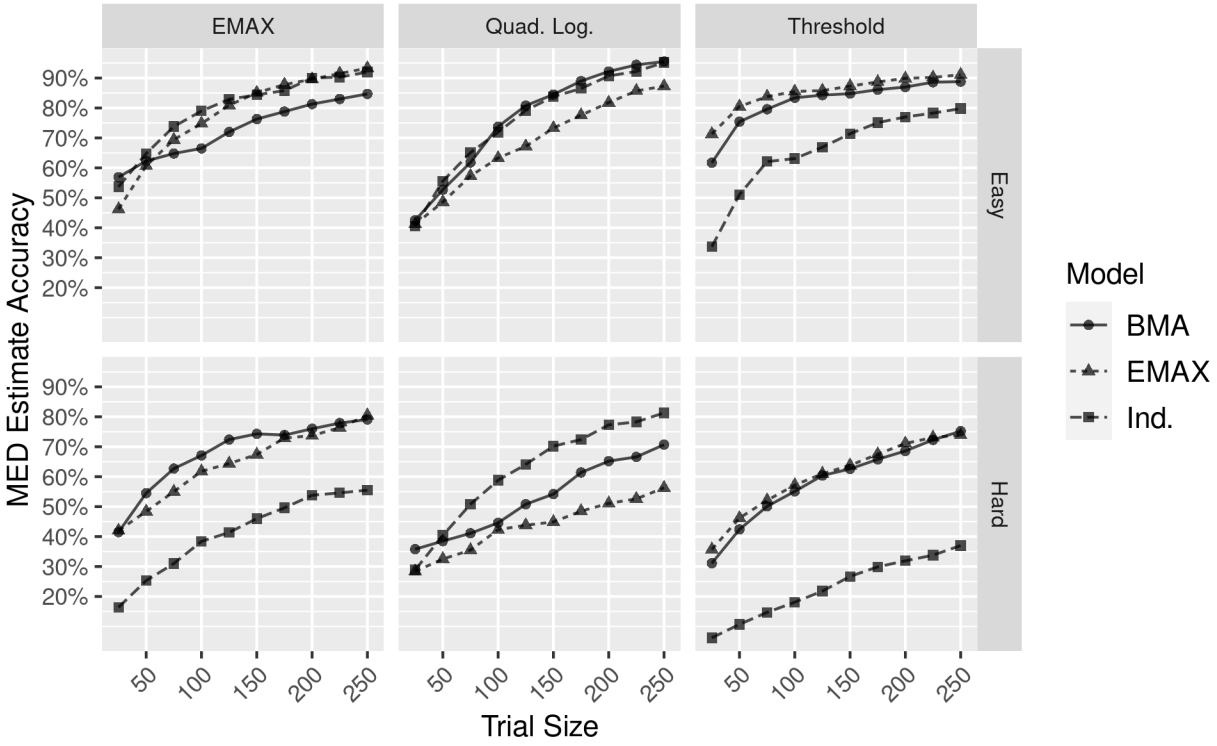

Figure 1.5.3: MED accuracy of BUD trials as a function of trial size and model. Rates computed based on the simulated trials with fixed size of 250 patients by considering what each trial would have concluded about the MED if it were stopped earlier (subject additionally to early stopping at 99% posterior confidence as described in section 1.2.

mance can be much worse. The EMAX model does very well in all scenarios but the quadratic logistic ones, where it is badly biased since it cannot accommodate non-monotonicity and depending on the design this can make it 15-20% (additively) less accurate at detecting the MED than the BMA or independent models. The independent model struggles very badly in cases where many of the doses have similar (or identical) response rates, which is common in the hard scenarios and the flat one, where it requires dozens of extra patients to be as accurate as the other models. These are essential situations where a method must perform well because it is typical for drugs being studied in Phase 2b to not be effective (regardless of dose) or at least to not be so wildly effective that the range of response rates across doses is very large.

Another important point here is that when misspecification and bias are at play the models show diverging performance as trial size grows (Figure 1.5.3). In other words, the loss in MED accuracy from using a badly biased dose-response model can be *worse* with a larger trial. This is evident in the quadratic logistic scenarios, where the EMAX model struggles relative to the others at larger trial sizes and the BMA is outpaced by the independent model in the hard case, where the flatter curve makes it harder to confidently identify the non-monotonicity (since the EMAX submodel still has a competitive overall fit here). Intuitively, the EMAX model's response rate estimates are biased upwards for the 1.0 dose (which isn't actually effective) but biased downwards for the 0.4 and 0.7 doses, which increases our uncertainty about if the MED even exists. This means a BUD will be more inclined to reduce this uncertainty, which under monotonicity it will do by putting more patients on the highest dose.

1.5.3 Adaptive trial stopping

What is the effect of stopping trials adaptively after a certain level of posterior confidence in the MED has been reached, as opposed to stopping at a fixed trial size? Figure 1.5.4 reveals

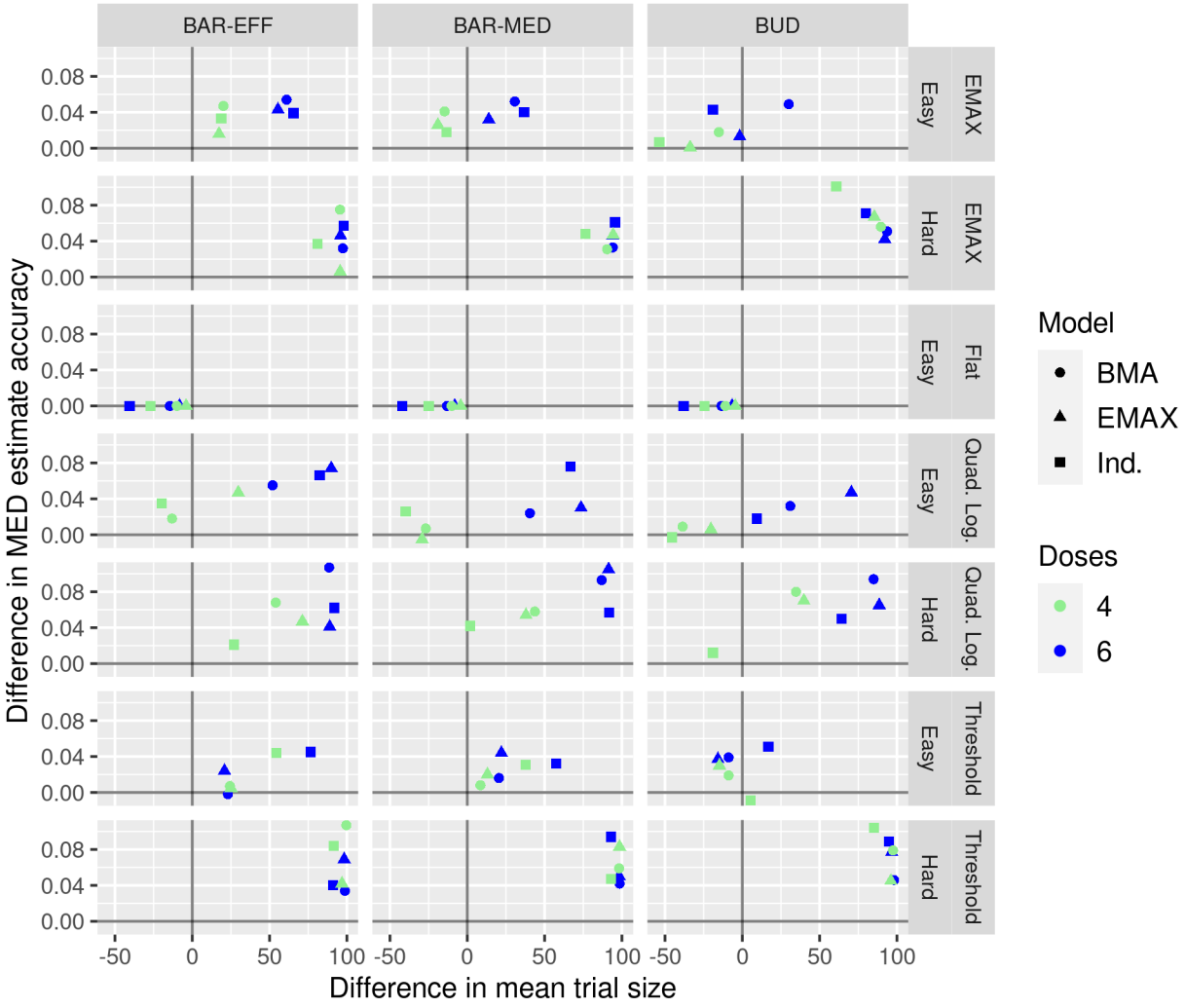

Figure 1.5.4: Difference in average accuracy and trial size between trials stopped adaptively vs. those with fixed total size. For both accuracy and trial size the ordering is adaptive minus fixed, so positive numbers mean the *adaptive* trial is more accurate/larger. Columns indicate randomization rule, rows indicate scenario, the shape of each point indicates the model, and the color of each point indicates the number of doses being studied.

how in many of the scenarios we consider 250 patients is not enough data to typically give 95% posterior certainty in the MED, so in these cases the adaptively stopped trials are larger and, unsurprisingly, more accurate due to the additional data (top right quadrant). This reflects how power to detect the MED is depends very sensitively on the a variety of features of the underlying dose-response curve, so it may not be sensible to follow the traditional strategy of choosing a *fixed* trial size based on a priori power analysis.

In some cases, such as when there are fewer doses being studied or the MED is obvious (e.g. the flat scenario), the adaptively stopped trials are on average much smaller and still as or more accurate. This is possible due to the definition of the stopping rule, which will stop trials early when the data are conclusive (perhaps most of the time, in these cases) *but* will make trials run longer when the data are inconclusive (and potentially misleading). If these trials with "unlucky" data are not too common (i.e. in scenarios where the MED can *typically* be learned quickly) then allowing to become large can increase the overall Frequentist accuracy rate while still keeping the average trial size fairly small.

Another effect of early stopping is on the Frequentist calibration of the posterior MED accuracy claims. This reflects the extent to which, when the BUD-BMA posterior is x% confident in the MED, if it is *also* correct x% in the Frequentist long-run of repeated sampling (Rubin 1984). Figure 1.5.5 shows that in general BUD-BMA is conservatively calibrated, in that its Frequentist accuracy is typically *greater* than or equal to its posterior certainty (of the maximum a posteriori MED estimate being correct). An exception can crop up in trials that are stopped adaptively with greater than or equal to 95% posterior certainty of knowing the MED (see, especially the quadratic logistic hard case). This is unsurprising, as conditioning on seeing more dramatic data (which the early stopping does) means the data are likely to be slightly noisier. This implies that more conservative or nuanced adaptive stopping rules may be useful in Bayesian trials that require strict Frequentist calibration (or conservative calibration).

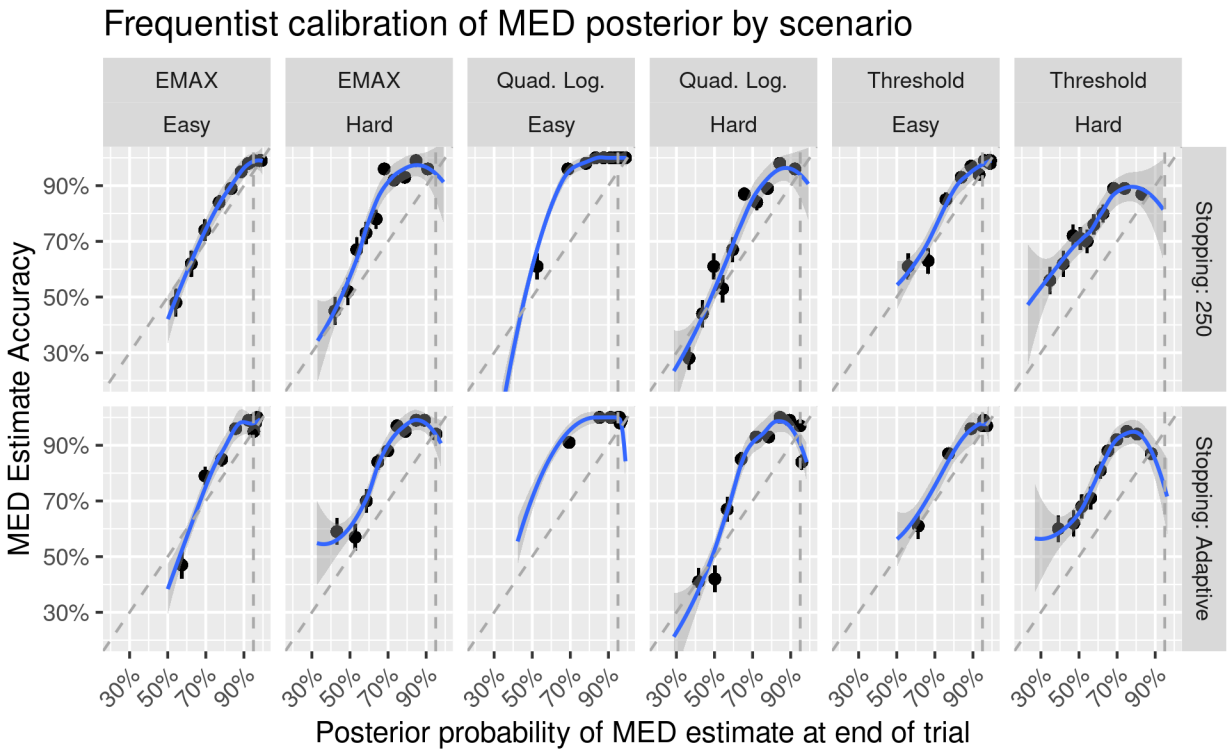

Figure 1.5.5: Frequentist calibration of BUD-BMA. The 1,000 simulated trials for each scenario are broken into deciles on the posterior credibility of trial's final MED estimate and plotted against the proportion among those 100 trials in which the MED estimate is actually correct. For instance, the leftmost point in each panel represents the 100 trials in each simulation in the lowest 10% of posterior confidences in the final MED guess. The location along the x-axis is the median posterior probability in each decile. The loess curves and confidence band reflect a regression of the underlying binary correctness of each trial's MED estimate on that trial's posterior credibility in its MED estimate.

1.5.4 BUD-BMA compared to popular alternatives

In Figure 1.5.6 we compare the BUD-BMA's patient allocations and accuracy to its likely competitors in practice. ER-MCP-Mod is standard in the field, BAR-EFF-EMAX is the most familiar Bayesian adaptive design and uses a simpler (to state) model of the dose-response curve, and BAR-MED-BMA has the most similar design philosophy and performance to BUD. The Bayesian designs shown here use adaptive trial stopping, while ER-MCP-Mod does not. We show the underlying true dose-response curves to make reasoning about the patient allocations easier.

In general the BUD puts fewer patients on low doses than do other designs when these doses are obviously not very effective and are not expected to be otherwise helpful for identifying the MED (by clarifying relevant features of the dose-response curve). The poorer performance of BAR-EFF compared to BUD and BAR-MED is due somewhat to its unabashed use of the EMAX model but more so to its insistence on putting patients on doses with higher response rates. In some cases this requires many more patients and still lead to lower accuracy.

Earlier we noted that BUD and BAR-MED can have quite difference performance when there is a dose that is obviously ineffective (and thus not the MED) but that is nevertheless important for learning about the MED under the model being used. The easy threshold scenario highlights one such case. Here the key dose is the 0.4 dose, which the BUD places more patients on (despite running a much smaller trial). Evidently it is important for learning about the slope of the curve and that the 0.7 dose is in fact above the 40% target rate defining the MED. In contrast the BAR-MED puts fewer patients on the 0.4 dose because it is well below the 40% target rate. A similar situation arises in the easy quadratic logistic scenario with the 0.2 dose.

Compared to ER-MCP-Mod, BUD-BMA is often much more accurate. This is due in part to its adaptive stopping, which allows larger trials. However, Table 1.5.1 reveals that

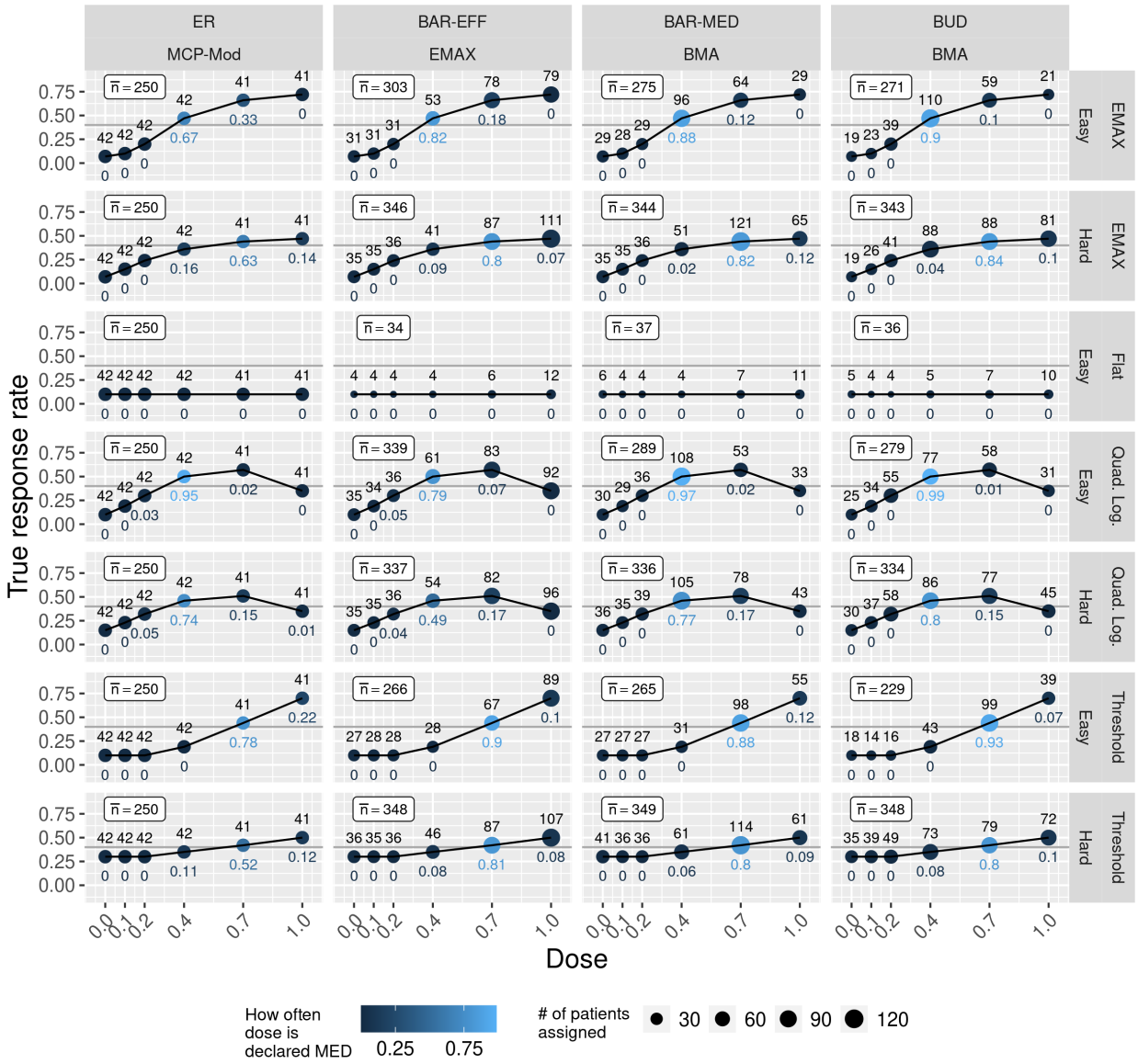

Figure 1.5.6: Patient allocations of BUD-BMA and competitors. We plot the true dose-response curve for each scenario (solid line), how many patients were assigned to each dose on average (number above each point and size of point), and in what proportion of trials each dose was declared the MED (number below each point and color of point). The box in the top left of each panel shows the average total sample size for each trial. The MED was defined as the lowest dose with response rate above 40% (horizontal gray line).

the BUD randomization rule and the BMA modeling are also major contributors to this difference in performance.

1.6 Discussion

In this paper we proposed a new approach to design and analysis of Phase 2b dose-ranging trials that combines a Bayesian uncertainty-directed (BUD) rule to randomize patients with Bayesian model averaging (BMA) of the dose-response curve. This is the first time in the literature that a BUD design has been used with a nontrivial data model, and we developed a fast Sequential Monte Carlo algorithm to handle the resulting increased computational burden. In general, we find that the BUD randomization strategy is well-suited to dose-ranging trials and can lead to more accurate or smaller trials than competing designs. While it is challenging to specify a model that performs optimally for a broad range of underlying dose-response curves in this setting, BMA is capable of striking a relatively attractive balance between efficiency and robustness.

There are a number of open problems for future work in this research area, both for design and analysis. An important practical problem is extending BUD designs to better accommodate outcomes that are slow to be observed. When new patients arrive before all previous patients have had their outcomes observed, a simple BUD can lose efficiency because the information gain calculation it uses for randomization does not account for the doses of these patients with pending outcomes, which we will soon learn more about regardless of how the next cohort is randomized. In ongoing work we are characterizing the extent and effects of this problem in practical settings, and to develop BUD rules and computational strategies that do not have this shortcoming.

Another possible direction is to consider more sophisticated and relevant forms of Bayesian model combination than model averaging. One issue with Bayesian model averaging is the fact that it assumes that one of the candidate models does in fact include the data-generating

distribution. Another is that it weighs the candidate models by their overall fit, which is not necessarily desirable in settings like this one where our goal is inference for a specific target parameter. In truth, if our goal is to identify the MED, then we likely want to draw most from the models that have good fit for the doses surrounding the MED (rather than for those that are far from it).

References

Antonijevic, Z., Pinheiro, J., Fardipour, P. and Lewis, R.J. (2010) Impact of Dose Selection Strategies Used in Phase II on the Probability of Success in Phase III. Statistics in Biopharmaceutical Research, 2, 469–486.

Berry, D.A., Müller, P., Grieve, A.P., Smith, M., Parke, T., Blazek, R., et al. (2002) Adaptive Bayesian Designs for Dose-Ranging Drug Trials. Case Studies in Bayesian Statistics Volume V (eds C. Gatsonis),, R.E. Kass),, B. Carlin),, A. Carriquiry),, A. Gelman),, I. Verdinelli), et al.), pp. 99–181. Springer New York, New York, NY.

Bornkamp, B., Bretz, F., Dette, H. and Pinheiro, J. (2011) Response-adaptive dose-finding under model uncertainty. The Annals of Applied Statistics, 5, 1611–1631.

Bretz, F., Pinheiro, J.C. and Branson, M. (2005) Combining Multiple Comparisons and Modeling Techniques in Dose-Response Studies. Biometrics, 61, 738–748.

Chopin, N. (2002) A sequential particle filter method for static models. Biometrika, 89, 539–552.

Chopin, N. and Papaspiliopoulos, O. (2020) An Introduction to Sequential Monte Carlo. Springer, Cham, Switzerland.

Committee for Medicinal Products for Human Use (CHMP). (2014) Qualification Opinion of MCP-Mod as an Efficient Statistical Methodology for Model-Based Design and Analysis of Phase II Dose Finding Studies under Model Uncertainty. Euorepean Medicines Agency.

Domenicano, I., Ventz, S., Cellamare, M., Mak, R.H. and Trippa, L. (2019) Bayesian uncertainty-directed dose finding designs. Journal of the Royal Statistical Society: Series C (Applied Statistics).

Drovandi, C.C., McGree, J.M. and Pettitt, A.N. (2014) A Sequential Monte Carlo Algorithm to Incorporate Model Uncertainty in Bayesian Sequential Design. Journal of Computational and Graphical Statistics, 23, 3–24. Hay, M., Thomas, D.W., Craighead, J.L., Economides,

C. and Rosenthal, J. (2014) Clinical development success rates for investigational drugs. Nature Biotechnology, 32, 40–51.

LaVange, L. and Zineh, I. (2016) FDA Fit-For-Purpose Determination of MCP-Mod.

Pinheiro, J., Bornkamp, B., Glimm, E. and Bretz, F. (2014) Model-based dose finding under model uncertainty using general parametric models: J. PINHEIRO ET AL . Statistics in Medicine, 33, 1646–1661.

Rubin, D.B. (1984) Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. The Annals of Statistics, 12.

Schorning, K., Bornkamp, B., Bretz, F. and Dette, H. (2016) Model selection versus model averaging in dose finding studies. Statistics in Medicine, 35, 4021–4040.

Thall, P.F. and Wathen, J.K. (2007) Practical Bayesian adaptive randomisation in clinical trials. European Journal of Cancer, 43, 859–866.

Thomas, N. (2006) Hypothesis Testing and Bayesian Estimation using a Sigmoid E max Model Applied to Sparse Dose-Response Designs. Journal of Biopharmaceutical Statistics, 16, 657–677.

Trippa, L., Lee, E.Q., Wen, P.Y., Batchelor, T.T., Cloughesy, T., Parmigiani, G., et al. (2012) Bayesian Adaptive Randomized Trial Design for Patients With Recurrent Glioblastoma. Journal of Clinical Oncology, 30, 3258–3263.

Ventz, S., Cellamare, M., Bacallado, S. and Trippa, L. (2018) Bayesian Uncertainty Directed Trial Designs. Journal of the American Statistical Association, 1–13.

Wason, J.M.S. and Trippa, L. (2014) A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. Statistics in Medicine, 33, 2206–2221.

Wasserman, L. (2000) Bayesian Model Selection and Model Averaging. Journal of Mathematical Psychology, 44, 92–107.

Yin, G., Chen, N. and Jack Lee, J. (2012) Phase II trial design with Bayesian adaptive randomization and predictive probability: Phase II Trial Design. Journal of the Royal Statistical Society: Series C (Applied Statistics), 61, 219–235.

Yuan, Y. and Yin, G. (2011) Bayesian phase I/II adaptively randomized oncology trials with combined drugs. The Annals of Applied Statistics, 5, 924–942.

CHAPTER 2

DYNAMIC BORROWING FROM HISTORICAL CONTROLS VIA THE SYNTHETIC PRIOR WITH COVARIATES IN RANDOMIZED CLINICAL TRIALS

2.1 Introduction

There is an explosive interest in utilizing historical control data to improve the design and analysis of a future trial, both in terms of methodological research and clinical trial conduct (Viele et al. 2014). From a regulatory perspective, historical controls have been permitted in *confirmatory* trials primarily in rare and pediatric diseases as well as in devices (Ghadessi et al. 2020). However, the regulatory threshold for their use is lower in non-confirmatory settings where there is less demand for conservative Type I error guarantees (U.S. Food and Drug Administration 2019). Statistical models are critical due to the challenge in reconciling historical data and concurrent control data. Ideally, historical data that are more "similar" should be borrowed from more to aid statistical inference. The main questions are how to measure the similarity and how to "borrow" based on this measure.

Popular statistical methods for leveraging historical controls include propensity score approaches, which typically match or weight historical and current patients based on covariates (Lim et al. 2018, Lin et al. 2018), as well as Bayesian modeling strategies including meta-analytic priors such as MAP and RMAP (Neuenschwander et al. 2010; Schmidli et al. 2014), power priors (Chen & Ibrahim 2000), commensurate priors (Hobbs et al. 2012), and multisource exchangeability models (Kaizer et al. 2018), which tend to borrow based primarily on similarities in response rates between the historical and new data. A barrier to using propensity score methods is that they require rich patient-level data, which is often unavailable (e.g. when such data are owned by a competing developer), and they must both have *and* select a sufficient set of covariates to control bias. While the Bayesian methods

are attractive strategies to dynamically adapt the degree of historical borrowing, our current work suggests that in some cases there is room in this framework to do so more efficiently.

We propose a new model called SPx, standing for "synthetic prior with covariates", to sharpen inferences about a new trial's control group response rate by borrowing from historical data. It relies on a Bayesian expert system (Spiegelhalter et al. 1993) that merges different philosophies for how the historical data should be treated. This model can be used simply for the analysis of a completed trial, but we also discuss how it can be embedded in an adaptive trial design to reduce the needed control arm sample size.

To be very clear, SPx requires only summary statistics, not patient-level data, from the historical trials. Such trial-level information is routinely reported in publications and press releases, and includes sample size, response rate, eligibility criteria, and average demographics or pre-trial clinical measures. This means that researchers using SPx may potentially draw from many more historical trials than if they were to use methods that rely on patient-level data, which are often inaccessible due to ethical and confidentiality reasons.

The paper proceeds as follows. In Section 2.2 we introduce the specific statistical setting, related methods, and the SPx model. Section 2.3 describes a two-stage adaptive design that leverages SPx to reduce control group sizes. In Section 2.4 we discuss results from an extensive simulation study that benchmarks the method's performance and sheds light on its approach to dynamic borrowing. In Section 2.5 we apply the SPx approach to the design and analysis of a trial in rheumatoid arthritis, and in Section 2.6 we conclude.

2.2 The SPx Model

2.2.1 Basic Data Setting and Related Models

Data structure. Following Schmidli et al. (2014), we consider trial-level historical data, which are summary statistics for the control arm. Specifically, denote the data by $(y_h, n_h, \mathbf{x}_h)$ for

trials $h = 1, \ldots, H, H+1$ where trials $1, \ldots, H$ are the historical trials and trial $H+1$ is the new trial, y_h is the number of responders in trial h's control group (from a binary endpoint), n_h is the number of control patients in trial h, and $\mathbf{x}_h$ is a $(p+1)$-dimensional vector containing p group-level covariates for trial h's control group as well as an intercept.

The sampling model of the historical and new trial control data is simply

$$y_h | \mathbf{x}_h, \psi_h \overset{iid}{\sim} Bin(n_h, \psi_h), \quad h = 1, \ldots, H, H+1,$$

where Bin denotes a binomial distribution, and ψ_h, potentially a function of covariates $\mathbf{x}_h$, is the true response rate of the control arm from trial h. While we say "control" arm here, in the historical trials this arm may have been a "treatment" that is now the standard of care. The key modeling questions are how to specify a joint prior on the true response rates ψ_h across the historical and new trials and how to take advantage of the covariates.

MAP and RMAP priors. The MAP, or meta-analytic predictive prior (Neuenschwander et al. 2010) is exchangeable, and models the logits $\theta_h := \text{logit}(\psi_h)$ of the historical trials *and* the new one with a common normal prior distribution:

$$\theta_1, \ldots, \theta_H, \theta_{H+1} | \mu, \tau^2 \overset{iid}{\sim} N(\mu, \tau^2).$$

The prior mean μ and variance τ^2 are given hyperpriors to complete the hierarchical model. Although the authors emphasize the predictive interpretation of the MAP prior in (Neuenschwander et al. 2010), here we focus on its hierarchical nature to emphasize how it induces borrowing through shrinkage. In MAP the variance parameter τ^2 controls the degree of historical borrowing: if it is small then the posterior will shrink θ_{H+1} strongly towards the historical data, and if it is large then historical borrowing will be curtailed. Thus the hyperprior for τ^2 is crucial to the performance of the MAP method, and in the original work Neuenschwander et al. (2010) recommended sensitivity analysis for this part of the prior.

In contrast, MAP's performance does not depend as heavily on the hyperprior for μ as long as it is not unreasonably concentrated, and often a noninformative uniform prior is used. The MAP model's popularity stems from its simple, familiar random effects model and good performance when the historical control rates are largely similar to the new trial's control rate. However, when the historical data are misleading the MAP approach often continues to borrow too heavily and thus lacks robustness.

Due to the MAP approach's strong reliance on the historical data's relevance to avoid bias and inflated Type I error, a Robust MAP (RMAP) model was also proposed (Schmidli et al. 2014). This extension is a mixture of the MAP prior and an independent prior that involves no borrowing from historical data, given by

$$\begin{aligned} \theta_1, \ldots, \theta_H | \mu, \tau^2 &\overset{iid}{\sim} N(\mu, \tau^2), \quad \text{and} \\ \theta_{H+1} | \mu, \tau^2 &\overset{ind.}{\sim} \pi N(\mu, \tau^2) + (1-\pi) Logistic(0,1). \end{aligned}$$

The componenent $Logistic(0,1)$ in the mixture prior for θ_{H+1} gives rise to robustness as it does not allow shrinkage of θ_{H+1}. Back on the probability scale, the $Logistic(0,1)$ component is equivalent to a $Unif(0,1)$ prior. In contrast to MAP, in RMAP the hyperprior on τ^2 is less essential for limiting bias because of the inclusion of the independent component. However, the prior weight π on the historical borrowing component is pre-specified and must be tuned through simulation to reliably control bias and Type I error. Based on that work $\pi = 0.5$ may be a reasonable default value in some settings.

Commensurate priors. Another approach to historical borrowing is a commensurate prior model (Hobbs et al. 2012), which explicitly models the difference between the historical and new trials (Pocock 1976). In our setting it first assumes that all historical rates are equal and then specifies the joint prior on the historical and new rates through a marginal prior

on the historical rate and a "commensurate" prior on the new rate given the historical one:

$$\theta_1 = \cdots = \theta_H = \theta | \mu, \tau^2 \sim N(\mu, \tau^2), \quad \text{and}$$
$$\theta_{H+1} | \theta \sim N(\theta, \sigma^2).$$

Thus the (logit) difference between the new trial's rate and the historical rate is modeled as $N(0, \sigma^2)$, where the variance σ^2 is the key parameter controlling the amount of historical borrowing. Despite that assuming homogeneity of the response rates in the historical trials may be an oversimplification, this conditional specification provides a different mechanism to control borrowing than in the meta-analytic approaches. In past work (Hobbs et al. 2012) commensurate priors have been designed to work with patient-level data and covariates.

2.2.2 The SPx Prior

SPx uses Bayesian model averaging (BMA) to combine elements from both the RMAP and commensurate prior methods with a regression on covariates. In particular, we model the historical rates as conditionally exchangeable given covariates. Then the prior for the new rate, θ_{H+1}, is a mixture of three alternative points of view, or "experts":

Expert 1. a commensurate model, a prediction based directly on the historical rates (centered on their weighted average);

Expert 2. a regression model, a prediction based on the new trial's covariates $\mathbf{x}_{H+1}$ and the same covariate-response relationship in the historical data (centered on this regression); and

Expert 3. an independent no-borrowing model, in which the new rate is unrelated to the historical rates and no historical borrowing occurs.

A key distinction with RMAP is that the SPx priors within the experts are well-separated, leading to decisions about the degree of historical borrowing that adapt more quickly to the data as we later discuss. We now introduce the SPx model step-by-step.

2.2.2.0.1 Priors for θ_h $(h = 1, \ldots, H)$ and θ_{H+1}. Below we assume that the historical response rate θ_h for trial $h = 1, \ldots, H$ can be modeled by a regression on covariates, and the new response rate θ_{H+1}'s prior is a mixture of three experts each using a different borrowing strategy: direct historical borrowing, regression (on covariates), and no borrowing. Specifically,

$$\theta_h | \boldsymbol{\beta}, \tau^2, \mathbf{x}_h \overset{ind.}{\sim} N(\boldsymbol{\beta}^T \mathbf{x}_h, \tau^2), \quad h = 1, \ldots, H$$

$$\theta_{H+1} = z_{hist} \underbrace{\theta_{hist}}_{\text{Expert 1}} + z_{reg} \underbrace{\theta_{reg}}_{\text{Expert 2}} + z_{ind} \underbrace{\theta_{ind}}_{\text{Expert 3}} .$$

Here the mixture, or model averaging, prior is obtained by assuming a Categorical prior for $(z_{hist}, z_{reg}, z_{ind})$, given by $(z_{hist}, z_{reg}, z_{ind}) \sim Cat(p_{hist}, p_{reg}, p_{ind})$, meaning that the z_k are binary and θ_{H+1} is drawn from model $k \in \{hist, reg, ind\}$ with prior probability p_k. Priors for each expert, i.e. $\theta_{hist}, \theta_{ref}$, and θ_{ind}, are discussed next.

2.2.2.0.2 Priors for $\{\theta_{hist}, \theta_{reg}, \theta_{ind}\}$. Below we use the terms "expert" and "submodel" interchangeably. The three prior submodels for θ_{H+1} are the key construction in SPx:

Expert 1 (direct historical): $\theta_{hist} | \mu_{hist}, \sigma^2 \sim N(\mu_{hist}, \sigma^2), \quad \mu_{hist} := \sum_{h=1}^{H} w_h \theta_h;$

Expert 2 (regression): $\theta_{reg} | \boldsymbol{\beta}, \tau^2, \mathbf{x}_{H+1} \sim N(\boldsymbol{\beta}^T \mathbf{x}_{H+1}, c\tau^2);$

Expert 3 (no-borrowing): $\text{logit}^{-1}(\theta_{ind}) \sim Beta(0.5, 0.5).$

The prior for θ_{hist}, the direct historical borrowing expert, is similar to the ideal of the commensurate prior without assuming the historical rates to be all equal; it instead assumes that θ_{hist} is centered at a weighted average of the historical rates. We define the weights w_h to sum to 1 and be proportional to a distance metric between the regression's predicted

response rates for trials h and $H+1$:

$$w_h \propto (0.5)^{\frac{|\tilde{\pi}_h - \tilde{\pi}_{H+1}|}{0.05}}$$

with $\tilde{\pi}_h := \text{logit}^{-1}(\boldsymbol{\beta}^T \mathbf{x}_h)$. Although the form of the weights typically does not have an outsize effect on the posterior (especially when H is not very small), this choice encourages more borrowing from historical trials with more similar covariates, to the extent that the covariates are thought to predict response rates.

The prior on θ_{reg}, the regression expert, accounts for the noise from regression on covariates. Note that this prior uses the same $\boldsymbol{\beta}$ coefficients as in (2.2.1), which "borrows" the information through covariates as the impact of the covariates on response rate is assumed to be the same ($\boldsymbol{\beta}$) in all the trials, historical or new. A key point is that its scale depends on the same τ^2 indicating how successful the regression is at predicting the historical response rates, but modified by a constant $c = 1/25$ for reasons we discuss below along with the hyperpriors.

Lastly, θ_{ind} is the independent or no-borrowing expert. It is included for robustness, and uses a standard Jeffreys prior.

To summarize, the full hierarchical model so far is given by

$$
\begin{aligned}
y_h|\mathbf{x}_h, \psi_h &\overset{iid}{\sim} Bin(n_h, \psi_h), \quad h = 1, \ldots, H, H+1 \\
\theta_h &:= \text{logit}(\psi_h), \quad h = 1, \ldots, H, H+1 \\
\theta_h|\boldsymbol{\beta}, \tau^2, \mathbf{x}_h &\overset{ind.}{\sim} N(\boldsymbol{\beta}^T\mathbf{x}_h, \tau^2) \quad h = 1, \ldots, H \\
\theta_{H+1} &= z_{hist}\theta_{hist} + z_{reg}\theta_{reg} + z_{ind}\theta_{ind}, \\
\theta_{hist}|\mu_{hist}, \sigma^2 &\sim N(\mu_{hist}, \sigma^2), \quad \mu_{hist} := \sum_{h=1}^{H} w_h\theta_h, \\
w_h &\propto (0.5)^{\frac{|\tilde{\pi}_h - \tilde{\pi}_{H+1}|}{0.05}}, \quad \tilde{\pi}_h := \text{logit}^{-1}(\boldsymbol{\beta}^T\mathbf{x}_h) \\
\theta_{reg}|\boldsymbol{\beta}, \tau^2, \mathbf{x}_{H+1} &\sim N(\boldsymbol{\beta}^T\mathbf{x}_{H+1}, c\tau^2), \\
\text{logit}^{-1}(\theta_{ind}) &\sim Beta(0.5, 0.5), \\
(z_{hist}, z_{reg}, z_{ind}) &\sim Cat(p_{hist}, p_{reg}, p_{ind}).
\end{aligned}
$$

2.2.2.0.3 Prior model probabilities and model hyperpriors. The prior model probabilities $(p_{hist}, p_{reg}, p_{ind})$ and the hyperpriors for the variances σ^2 and τ^2 (and to a lesser degree $\boldsymbol{\beta}$) are important because they impact how SPx's model averaging balances between the different borrowing strategies. To see how, note that the posterior for θ_{H+1} is the average of the posteriors under each expert, weighting by the posterior probability that each is "correct":

$$
p(\theta_{H+1}|D) = \sum_{k \in \{hist, reg, ind\}} p(\theta_k|z_k = 1, D) \cdot p(z_k = 1|D). \tag{2.2.1}
$$

The posterior weights in (1), $p_k^* := p(z_k = 1|D)$ for $k \in \{hist, reg, ind\}$, may be written as

$$
p_k^* \propto p_k \cdot p(D|z_k = 1)
$$

where $p(D|z_k = 1) = \int_{\mathbb{R}^{H+1}} p(D|\theta_1, \ldots, \theta_{H+1}) \cdot p(\theta_1, \ldots, \theta_{H+1}|z_k = 1)\, \mathrm{d}^{H+1}(\theta_1, \ldots, \theta_{H+1})$ is the marginal likelihood, or evidence, of expert k. It is important to recognize that under each expert the hyperprior affects the marginal prior on the logit rates, $p(\theta_1, \ldots, \theta_{H+1}|z_k = 1)$, which in turn can greatly influence the expert's marginal likelihood and posterior model probability, changing the behavior of SPx overall.

In light of this we set the hyperpriors with the goal of making SPx adaptively *either* allow fairly aggressive historical borrowing *or* quickly transition to little historical borrowing, depending on the similarity between the current and historical trial data. To do so we both:

(a) make the priors in the *hist* and *reg* submodels relatively strongly concentrated (i.e. high prior probability of small σ^2 and τ^2), and

(b) give relatively high prior weight to the *ind* (no-borrowing) submodel (i.e. $p_{ind} > p_{hist}, p_{reg}$).

In particular, we take

$$\begin{aligned}(p_{hist}, p_{reg}, p_{ind}) &= \left(\frac{1}{8}, \frac{1}{8}, \frac{3}{4}\right), \\ \sigma &\sim TCauchy(0, 0.02, (0, \infty)), \\ \tau &\sim TCauchy(0, 2.5, (0, \infty)),\end{aligned}$$

where $TCauchy(m, s, I)$ represents a Cauchy distribution with location m and scale s that has been truncated to the interval I. While τ has a larger scale than σ, recall that the (conditional) prior variance of the regression submodel is $c\tau^2 = 1/25 \cdot \tau^2$, not τ^2, so the regression submodel makes similarly strong prior predictions as the direct historical borrowing one. The prior on the regression coefficients, which does not unduly affect inference, is

$$\beta_i \overset{iid}{\sim} Cauchy(0, 2.5), \quad i = 0, 1, \ldots, p,$$

following from Gelman et al. (2008).

This choice of prior underlies SPx's novel strategy to dynamically determine how much historical borrowing is appropriate. Combining (a) and (b) makes SPx more robust as needed in (1) *primarily* by increasing the posterior probability given to the no-borrowing submodel, p^*_{ind}, and to a much lesser degree by making the posterior inferences of the borrowing submodels more conservative. In further detail, (a) makes the borrowing submodels' marginal likelihoods decrease more quickly as disagreement between the new and historical data grows, and (b) accounts for the fact that the no-borrowing submodel makes less confident predictions overall and thus may have a relatively lower marginal likelihood even when it should be favored. While similar robustness might be sought by giving the borrowing submodels more diffuse priors, this would come at the cost of weaker borrowing when the historical data are actually relevant. We discuss these points further in the supporting information.

Effectively, through its well-separated BMA formulation SPx cleanly distinguishes between the questions of 1) *whether* we should borrow from the historical data at all and 2) *how strongly* we should borrow from those data, if we decide to. Standard commensurate prior models do not make this distinction since they have a unimodal prior on σ^2 and typical uses of RMAP do so less because of more diffuse priors for τ^2, meaning that the MAP component will be unable to suggest as strong historical borrowing when appropriate; see, e.g. the credible interval widths in Table 2.4.1. Although this approach may not be the only strategy to achieve refined dynamic borrowing, we find it to be a useful device.

Treatment effect estimation. To make inferences for the treatment effect we model the treatment group data (y_{trt}, n_{trt}) as independent from the historical and control data:

$$y_{trt}|\psi_{trt} \sim Bin(n_{trt}, \psi_{trt})$$
$$\psi_{trt} \sim Beta(0.5, 0.5).$$

Combined with the SPx prior for the new trial's control rate this induces a prior on the treat-

ment effect, which we define as the difference $\delta := \psi_{trt} - \psi_{H+1}$ (although other comparisons, such as relative risk, could just as easily be used).

Computation. Posterior computation for SPx can be done easily and efficiently using standard Bayesian MCMC tools. We implemented the model using JAGS (called from R) and it takes at most a few seconds to analyze a single data set on a personal computer. This produces draws from the posterior distribution of ψ_{H+1}, which can be combined directly with draws from the conjugate posterior for ψ_{trt} to get draws from the posterior for the treatment effect δ. For full details, see the JAGS and R code included in the supporting information.

2.3 Adaptive Design Based on Posterior Inference

We propose a two-stage adaptive design, largely following Schmidli et al. (2014), to potentially reduce the control group size to the extent that additional reliable information can be gained through historical borrowing. Intuitively, in the first stage the new trial enrolls a fixed and prespecified number of control group patients, then uses an interim check to assess how much historical borrowing is desirable and calibrates the remaining number of control patients to enroll in stage two accordingly. The interim check measures the degree of "compatibility" between the historical and new trial (stage one) data and expresses this as the number of control patients effectively gained by using the historical data. If the new trial's stage one control data is deemed less compatible with the historical data, little borrowing will happen and the stage two size will not be reduced much or at all. This limits the impact of prior assumptions about the relevance of the historical data on the overall trial size. Because we model the treatment group data independently from the control group data (having no shared parameters), the treatment group patients may be enrolled without reference to the control patients or the stages of the adaptive design (i.e. by simply changing the randomization probability after the interim check).

Formally, the adaptive design follows:

- Stage 1: Collect data on n_1^c control patients.
- Interim: Calculate the *prior effective sample size* n_{eff}^c of the SPx model for ψ_{H+1} given the Stage 1 data.
- Stage 2: Collect data on n_2^c more control patients, where $n_2^c = n_{max}^c - n_{eff}^c - n_1^c$ but truncated to be at least $p_{min} n_{max}^c$ but no more than $p_{max} n_{max}^c$ for some $p_{min} \in [0, 1]$ and $p_{max} \in [1, \infty)$ (e.g. $p_{min} = 0.75$ and $p_{max} = 1.25$).

The second stage is designed so that the total sample size of the control group is not intolerably lower or higher than the target size n_{max}^c.

Crucial to this design is the prior effective sample size of SPx at interim, n_{eff}^c, which is intended to assess how much information the historical data contribute about the new trial's control rate and how many fewer patients the new control group needs in exchange for this additional information (Hobbs et al. 2013, Schmidli et al. 2014). We use a simple moment matching definition of effective sample size (Weber 2020), which finds the Beta distribution with the same mean and variance as the SPx posterior, takes the effective sample size of this Beta distribution (the sum of its parameters; Morita et al. 2008), and subtracts the current sample size at interim, n_1^c. In truth, defining an effective sample size that has desirable properties for complex non-conjugate models such as SPx is an open research area (Morita et al. 2008, Morita et al. 2012, Neuenschwander et al. 2020). Partially to this end, the sample size thresholds used in Stage 2 of the design somewhat limit the impact of the choice of definition. The simplistic definition we use has the benefit of producing more conservative (smaller) effective sample sizes for SPx than other definitions, which sometimes produce implausibly large effective sample sizes for the SPx model. In any case, by measuring the effective sample size during, and not before, the trial we can measure the extent to which the new data diverge from the historical data, potentially safeguarding against inappropriate

borrowing from irrelevant historical data.

At the end of the trial, the decision rule to detect a treatment effect may take a variety of forms. Depending on the disease and regulatory setting, interest may focus on detecting nonzero or clinically significant effects. To detect nonzero effects, we may use the rule

$$P(\delta \neq 0|D) \geq 1 - q_{nonzero} \tag{2.3.1}$$

for the treatment effect $\delta := \psi_{trt} - \psi_{H+1}$ where D is all of the historical and new trial data and $q_{nonzero} \in (0, 1)$ is the posterior Type I error threshold. Alternatively, to detect clinically significant effects, we may use the rule

$$P(\delta > \delta_0|D) \geq 1 - q_{clinical} \tag{2.3.2}$$

for some minimal threshold $\delta_0 > 0$ and some $q_{clinical} \in (0, 1)$.

2.4 Simulation Study

How accurately does the SPx model estimate the new trial's control response rate, and does it perform respectably when the historical data are misleading and borrowing would be detrimental? Further, when used with an adaptive design does SPx successfully reduce the trial's control group size while maintaining power and Type I error?

To answer these questions we simulated clinical trials from scenarios defined by two factors: (a) whether or not the historical control rates are misleading (i.e. on average notably different from the new control rate), (b) whether or not the group-level covariates are associated with response rates. This yields four basic scenarios: Scenario 1 [ideal], where historical control rates are not misleading and covariates are predictive; Scenario 2 [covr], where historical rates *are* misleading but covariates are still predictive; Scenario 3 [hist], where historical rates are not misleading but covariates are *not* predictive; and Scenario 4

[worst], where not only are historical rates misleading but also covariates are not predictive. In all cases the historical trials were loosely based on the real historical trials we analyze in Section 2.5.

The full details of data generation for all four scenarios are provided in the supporting information. For each scenario, we generated a historical data set where the number of historical trials ranged from 48 to 50 (see supporting information for details). We fixed that single data set, and repeatedly generated data for a new trial, 1,000 times for each method. This reflects the type of Frequentist repeated sampling we expect drug developers and regulators to be concerned with, since at the point of trial planning or analysis it is reasonable to imagine replicating the new trial but not also all of the historical ones. We also varied the target maximum sample size for the new trial, and in the scenarios where covariates are predictive (1 and 2) we varied whether SPx would include all the useful covariates. In all scenarios we simulated treatment group data for the new trial independently from the control data, at rates higher by both 0 and 30 percentage points.

The historical data for Scenarios 3 and 4 are identical to those from Scenarios 1 and 2, respectively, except that the covariate data have been permuted so they are no longer associated with the response rates. In our "non-misleading" Scenarios (1 and 3) the observed historical control rates ranged from roughly 18% to 37% and the new trial's true control response rate was roughly 26%; in our "misleading" Scenarios (2 and 4) the observed historical rates only ranged from roughly 24% to 37% while the new trial's true control rate was lower at roughly 18%. This means that direct historical borrowing in Scenarios 2 and 4 will bias estimates of the new trial's control rate upwards and thus estimates of the treatment effect estimates downwards, reducing both the Type I error and power of effect testing. The opposite would happen if the historical control response rates were lower than the new trial's response rate. Which situation is more likely in practice depends on how a variety of factors such as standard of care, lifestyle, and demographics have changed over time.

We compared several methods including SPx, RMAP (Schmidli et al. 2014), and an independent model (Ind.) with no historical borrowing (i.e. $\theta_{H+1} \sim Logistic(0, 1)$). For SPx and RMAP we included both versions where the new trial's control group size was fixed and where the two-stage adaptive design described in Section 2.3 was used (with $n_1^c = n_{max}^c/2$, $p_{max} = 1.25$, and $p_{min} = 0.75$). The two versions gave a contrast on the potential gain in the adaptive design in reducing the control group sample size. The value n_{max}^c is fixed at 150 or 80, which are the maximum control sample size. For SPx our standard implementation used only 2 of 6 covariates associated with response rates to mimic imperfect knowledge or data collection, but we also include cases where SPx uses all 6 covariates to show the effectiveness of the method when the regression is strong and not misspecified. For RMAP we used a 50-50 mixture between the borrowing (MAP) component and the non-borrowing component, since this specification performed well in Schmidli et al. (2014). The 90-10 mixture they consider was too aggressive and extremely biased in many of our scenarios.

2.4.1 Estimation Accuracy for the New Trial's Control Rate

Table 2.4.1 shows performance of the models and designs for the task of reliably and efficiently estimating the new trial's control response rate. Although the overall goal of the trial is treatment effect testing, performance of the modeling strategies can be understood with more nuance by first considering control response rate estimation.

The results in Scenarios 1 and 3 reveal that SPx performs strongly when the historical controls are directly relevant (i.e. their average response rate is close to the new trial's rate), as shown by much smaller control group size in the adaptive design and uniformly better RMSE and interval width while maintaining good coverage. Interestingly, the benefit of covariate adjustment is limited mainly to better credible intervals (in length and coverage) when the historical control rates happen to be similar to the new rate; the RMSE and average control group size of SPx are not distinguishable comparing Scenario 1 to 3.

	$n^c_{max} = 150$					$n^c_{max} = 80$				
	SPx		RMAP		Ind.	SPx		RMAP		Ind.
	Fixed	Adapt.	Fixed	Adapt.	Fixed	Fixed	Adapt.	Fixed	Adapt.	Fixed
Scenario 1										
Size	150	123.4	150	116.1	150	80	64.9	80	61.1	80
RMSE	0.025	0.026	0.025	0.027	0.032	0.024	0.026	0.022	0.027	0.032
Coverage	94.2	96.0	98.2	98.5	96.7	99.0	98.7	99.8	99.7	99.6
Width	0.104	0.113	0.123	0.136	0.140	0.135	0.153	0.155	0.173	0.190
Scenario 2										
Size	150	127.7	150	147.6	150	80	65.0	80	71.7	80
RMSE	0.031	0.034	0.041	0.045	0.030	0.033	0.038	0.050	0.059	0.030
Coverage	91.4	92.0	91.8	88.2	97.1	97.2	95.6	95.1	89.8	99.6
Width	0.114	0.123	0.138	0.138	0.125	0.153	0.165	0.183	0.187	0.171
Scenario 3										
Size	150	122.7	150	115.9	150	80	64.1	80	60.9	80
RMSE	0.022	0.023	0.025	0.027	0.033	0.021	0.023	0.022	0.026	0.032
Coverage	96.6	96.8	98.6	98.5	97.3	99.8	99.6	99.9	99.9	99.7
Width	0.106	0.116	0.123	0.136	0.140	0.138	0.157	0.155	0.173	0.190
Scenario 4										
Size	150	156.4	150	146.9	150	80	78.6	80	72.3	80
RMSE	0.040	0.043	0.042	0.046	0.030	0.046	0.052	0.050	0.057	0.029
Coverage	93.5	92.1	91.3	87.2	96.4	98.1	96.0	95.7	92.1	99.7
Width	0.140	0.135	0.138	0.139	0.125	0.189	0.187	0.183	0.187	0.170

Table 2.4.1: Control group size and Frequentist estimation accuracy for the new trial's control response rate, averaged over 1,000 simulated trials. Two of six predictive covariates are used. Metrics are defined as follows: size is the mean control group size; RMSE is the root mean square error of the posterior mean of the control group rate; coverage is the proportion of trials in which the 95% quantile-based posterior credible interval for the control group rate contains the true rate; width is the mean width of these credible intervals. Note that the credible intervals are Bayesian and not designed or calibrated to give exactly 95% Frequentist coverage.

Notably, SPx can still perform well when the historical control rates are misleading as long as group-level covariates are moderately predictive of the rates (Scenario 2); in this case SPx saw modest dips in RMSE and coverage compared to the no-borrowing Ind approach, unlike RMAP, while still allowing a considerably smaller control group in the adaptive design. Beyond this, when regression is not just moderately but instead strongly predictive (Table 2.4.2, where all 6 covariates are known to the analyst) SPx performs even better, either further surpassing other methods (when the historical rates are directly relevant as in Scenario 1) or now slightly edging them out (when historical rates are misleading as in

	Scenario 1					Scenario 2				
	SPx		RMAP		Ind.	SPx		RMAP		Ind.
	Fixed	Adapt.	Fixed	Adapt.	Fixed	Fixed	Adapt.	Fixed	Adapt.	Fixed
$\mathbf{n^c_{max} = 150}$										
Size	150	121.1	150	115.7	150	150	128.7	150	148.6	150
RMSE	0.022	0.023	0.025	0.027	0.032	0.029	0.032	0.042	0.046	0.030
Coverage	97.0	96.2	98.2	98.9	96.8	94.9	96.1	91.4	86.4	96.9
Width	0.102	0.113	0.123	0.136	0.140	0.115	0.123	0.138	0.138	0.125

Table 2.4.2: Control group size and estimation accuracy for the new trial's control response rate when all six predictive covariates are used, averaged over 1,000 simulated trials. See Table 2.4.1 for metric definitions.

Scenario 2).

Unsurprisingly, the performance of SPx is less rosy in the challenging setting of Scenario 4 where the historical data are entirely misleading, though a silver lining is that its Frequentist coverage only degrades slightly. Its RMSE also slightly edges out that of RMAP, but the no-borrowing approach is clearly much preferred here. This highlights the point that if there is significant concern about the relevance of the historical data then priors should be made more conservative to protect against the greater likelihood of bias. The straightforward way to achieve this in SPx would be to change the prior submodel probabilities to further favor the independent component. We conducted a small sensitivity analysis to demonstrate this point (see supporting information). As one might predict, bias is mitigated in Scenarios 2 and 4 at the cost of smaller efficiency gains from borrowing in Scenarios 1 and 3.

2.4.2 Power and Type I Error for the New Trial's Treatment Effect

Results for testing treatment effects largely follow from those on control rate estimation. We report the Type I error rate of declaring a *nonzero* treatment effect when the effect is zero, and the power to declare a *clinically significant* effect (a response rate difference of greater than 20%) when the effect is moderately large (a difference of 30%).

Tables 2.4.3 and 2.4.4 examine the statistical power of the adaptive designs assuming 2 and 6 covariates were included in SPx, respectively. Type I error is well controlled by

all methods, with the exception of SPx having slightly inflated error in Scenario 1 in the larger trial setting. Of course, a drug developer or regulator requiring that Frequentist Type I error be more strictly controlled may calibrate the model or decision rule by simulation under the specific scenarios they are concerned about. This is a reality of all Bayesian trial methods and some Frequentist ones as well (e.g. Lewis et al. 2007), a point we revisit in the Discussion.

The power of SPx tends to be stronger than that of RMAP except in Scenario 3, where it is more or less matched. Compared to the no-borrowing approach, SPx has better (Scenario 1), similar (Scenario 3), or slightly lower (Scenario 2) power except in Scenario 4, all while substantially reducing the control group size when allowed. Naturally, using a stronger collection of covariates for the regression (Table 2.4.4) makes SPx even more powerful. Scenario 4 is where SPx has more substantial power loss compared to not borrowing since neither the historical data nor covariates are informative of the new trial's control rate. In reality, if there is a strong belief that historical data are not reliable and covariates are not informative, SPx should not be considered and more importantly, one might not want to try borrowing information from historical data in the first place.

2.4.3 SPx's Adaptive Weighting of Borrowing Strategies

To illustrate how SPx automatically adjusts the type of borrowing it performs depending on the historical and new trial data, Figure 2.4.1 plots the simulation distribution of SPx's posterior submodel weights p_k^* in Scenarios 1 through 4. When the historical data are entirely misleading (Scenario 4) SPx strongly favors its no-borrowing submodel, especially when the new trial is larger. Otherwise SPx puts most posterior mass on its two borrowing submodels, appropriately favoring the regression submodel over the historical one when the historical rates are misleading but covariates are useful (Scenario 2). In the reversed setting where the historical rates are not misleading but covariates are not predictive (Scenario 3) SPx

		$\mathbf{n^c_{max} = 150}$					$\mathbf{n^c_{max} = 80}$				
		SPx		RMAP		Ind	SPx		RMAP		Ind
		Fixed	Adapt.	Fixed	Adapt.	Fixed	Fixed	Adapt.	Fixed	Adapt.	Fixed
δ	**Scenario 1**										
	Size	150	123.4	150	116.1	150	80	64.9	80	61.1	80
0	$P(\delta \neq 0)$	7.8	6.9	3.0	2.6	4.3	4.5	3.7	1.7	1.5	2.1
0.3	$P(\delta > 0.2)$	70.9	70.4	56.0	52.5	55.5	54.2	51.7	37.5	33.8	34.4
	Scenario 2										
	Size	150	127.7	150	147.6	150	80	65.0	80	71.7	80
0	$P(\delta \neq 0)$	0.9	1.2	1.1	1.3	3.3	0.7	0.6	0.2	0.0	1.9
0.3	$P(\delta > 0.2)$	55.7	53.1	42.8	42.9	61.5	31.4	27.4	16.9	14.6	38.3
	Scenario 3										
	Size	150	122.7	150	115.9	150	80	64.1	80	60.9	80
0	$P(\delta \neq 0)$	1.6	1.4	2.2	2.2	3.3	2.0	2.0	1.9	2.0	2.0
0.3	$P(\delta > 0.2)$	60.0	56.8	60.3	57.0	58.0	36.8	34.1	36.1	33.2	32.6
	Scenario 4										
	Size	150	156.4	150	146.9	150	80	78.6	80	72.3	80
0	$P(\delta \neq 0)$	2.5	2.2	1.2	1.4	2.8	0.9	0.9	0.5	0.4	2.5
0.3	$P(\delta > 0.2)$	48.4	49.8	41.0	39.3	60.7	21.3	21.2	16.5	14.6	36.8

Table 2.4.3: Type I error and power for the new trial's treatment effect when two of six predictive covariates are used.

		Scenario 1					**Scenario 2**				
		SPx		RMAP		Ind.	SPx		RMAP		Ind.
		Fixed	Adapt.	Fixed	Adapt.	Fixed	Fixed	Adapt.	Fixed	Adapt.	Fixed
δ	$\mathbf{n^c_{max} = 150}$										
0	$P(\delta \neq 0)$	5.9	6.4	2.4	1.6	3.3	1.6	1.5	1.1	0.9	3.1
0.3	$P(\delta > 0.2)$	74.6	72.5	59.1	55.0	57.9	60.3	57.0	43.5	42.3	61.0

Table 2.4.4: Type I error and power for the new trial's treatment effect when all six predictive covariates are used.

still gives moderate weight to the regression model because its inclusion of covariates adds some noise but not substantial bias. This can be verified from small values of the estimated regression coefficients (results not shown).

2.5 Case Study

We discuss the application of SPx to the development of novel treatments for rheumatoid arthritis (RA). RA is an auto-immune disease that affects more than 1.3 millions of patients in the United States (Hunter et al. 2017). The symptoms of this disease include pain and

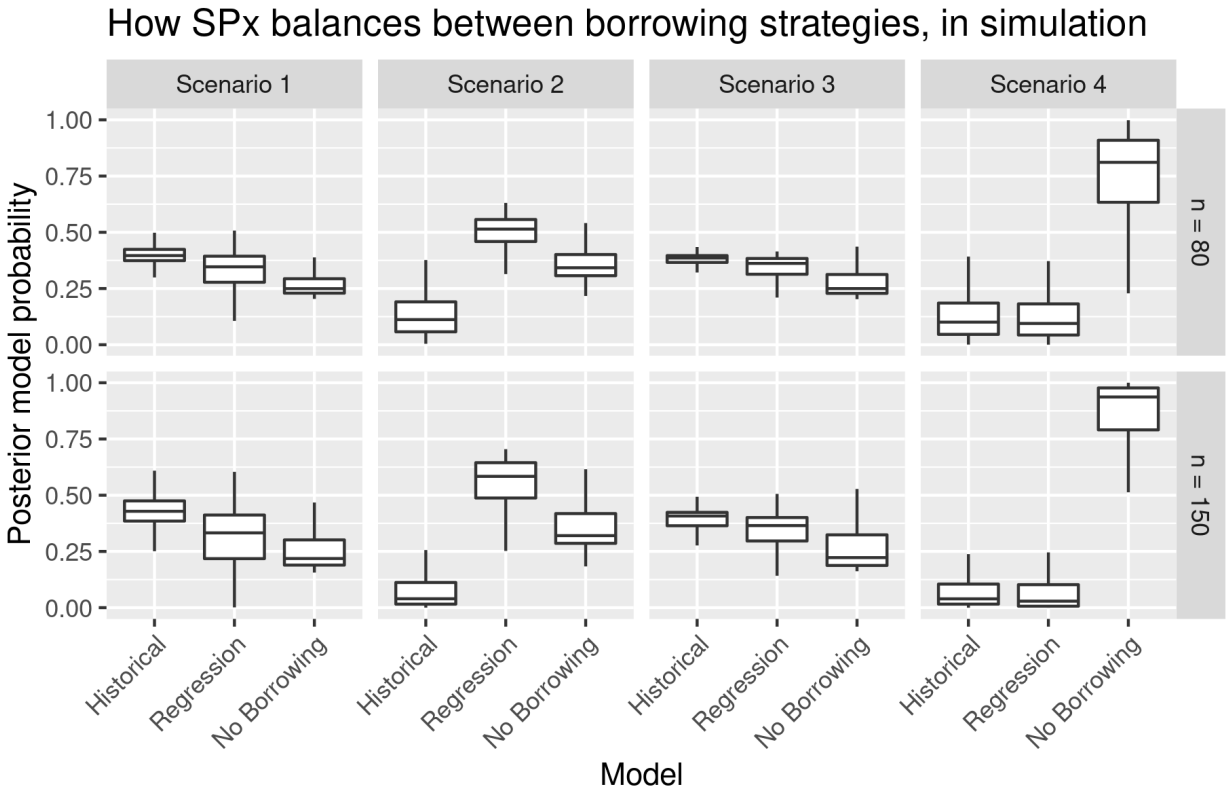

Figure 2.4.1: Boxplots of SPx's posterior submodel weights over 1,000 simulated trials in each scenario (columns) and trial size (rows).

swelling joints in hands and feet as well as morning stiffness lasting longer than 30 minutes. Although there is no cure for RA, several treatments are available to slow down the disease progression and alleviate symptoms. Adalimumab is a well-established, standard biologic therapy in RA and has been approved for almost 20 years. Because of the competitive landscape in RA, it is necessary to show that the new therapies are not only better than placebo but also better than adalimumab. We apply our SPx methodology in adalimumab trials conducted in the past two decades to illustrate the impact of including covariates in the model. Since adalimumab's own development was extensive, there are many past trials including an adalimumab arm that could be used as a rich source of historical data to potentially accelerate trials of prospective RA drugs.

We have collected group-level data from 11 past adalimumab trials, as shown in Table

Trial	Reference	Previous Treatment	Average Age	Size	Response Rate (%)
ALTARA	Kennedy et al. (2014)	MTX	48.8	43	39.5
ARMADA	Weinblatt et al. (2003)	MTX	56.0	62	21.0
DE019	Keystone et al. (2004)	MTX	56.1	200	24.0
IM133-001	Weinblatt et al. (2015)	MTX	51.4	61	39.3
ORAL-Standard	van Vollenhoven et al. (2012)	MTX	53.7	106	26.4
RA-BEAM	Taylor et al. (2017)	MTX	53.0	488	40.2
STAR	Furst et al. (2003)	MTX	55.8	315	29.5
A3921035	Fleischmann et al. (2012)	none	53.0	59	22.0
CHANGE	Miyasaka et al. (2008)	none	53.4	87	12.6
DE007	van de Putte et al. (2003)	none	50.2	70	10.0
DE011	van de Putte et al. (2004)	none	53.5	110	18.2

Table 2.5.1: Trials included in the adalimumab case study. Previous treatment, average age, size, and response rate refer to those of the control group in each trial. MTX is an abbreviation for methotrexate. The response rate is the proportion of patients experiencing a 20% improvement in joint health at 12 or 13 weeks on the American College of Rheumatology criterion (ACR20).

2.5.1. We note that Lim et al. (2018) provided a framework of objectively selecting historical trials to avoid cherry picking, and their strategy could be used to expand the data set to include even more historical adalimumab trials. We illustrate the use of SPx by borrowing from the placebo arms of these trials, though in practice the same could be done using the adalimumab arms. The primary endpoint we use is the popular, regulatory approved binary endpoint of ACR20, which is whether or not a patient has a 20% or greater improvement in joint health on the American College of Rheumatology criterion (ACR20) 12 or 13 weeks after treatment. The first trial-level covariate we use is whether patients in the trial had previous or ongoing treatment with methotrexate (MTX), a common first line therapy for rheumatoid arthritis. Unsurprisingly, Table 2.5.1 suggests that MTX use is a very strong predictor of ACR20 rates: the MTX trials have rates ranging from roughly 20-40% while the no-MTX trials have rates ranging from roughly 10-20%. This suggests that the regression strategy embedded in SPx may be useful despite the modest number of trials. We also include average age, which may be a proxy for disease progression or otherwise relate to ACR20 rates, though evidently in a weaker fashion here.

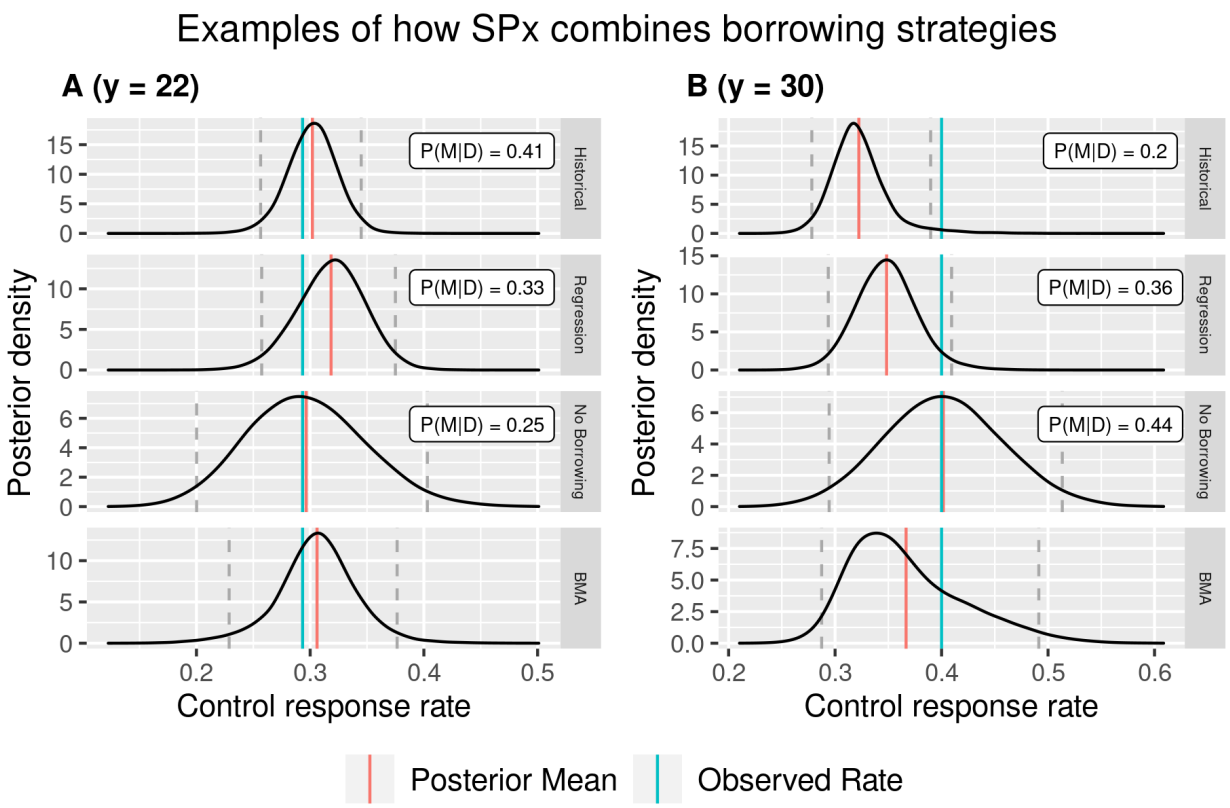

Figure 2.5.1: Posterior distributions of the new trial's control response rate in two examples. In A, 22 of 75 control patients in the new trial are responders, and in B 30 of 75 are responders. The top 3 panels in each show the posterior distribution for each of SPx's submodel, along with the submodel's posterior model probability $P(M|D)$. The bottom panel shows the model averaged posterior, which is the SPx inference.

We now show how SPx would borrow information from this historical data set in designing and analyzing the control arm in a new RA trial. We suppose that the new trial enrolls 75 control arm patients who have all had previous treatment with MTX and have an average age of 53. From Figure 2.5.1 we can see what the the Bayesian model averaging inference in SPx would be when the new control group's observed response is similar (A, at 29.3%) and quite different (B, at 40%) from the average historical response rates of 25.7% overall and of 31.4% among prior MTX trials. In A, we can see that SPx borrows fairly heavily, giving 75% of the posterior mass to to the relatively concentrated borrowing submodels. On the other hand, in B SPx is more conservative; while it still gives substantial

posterior mass to the borrowing submodels and its mean shrinks away from the observed rate, its 95% credible interval more or less reproduces that of the no-borrowing submodel.

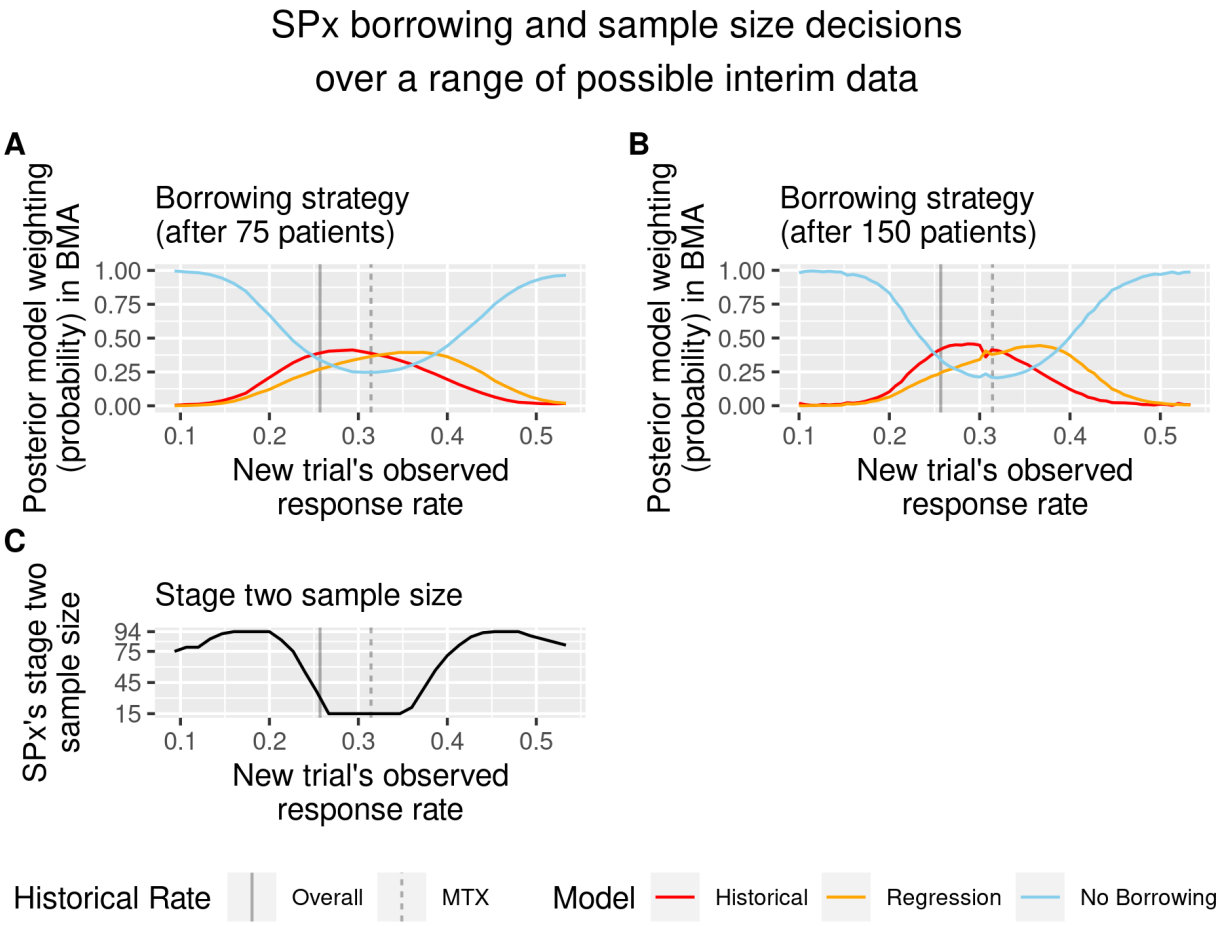

Figure 2.5.2: A and B plot the range of posterior submodel probabilities in SPx as the new trial's observed response rate varies, after 75 and 150 patients respectively. C plots the range of interim sample size decisions as the new trial's observed response rate varies, after 75 patients. The vertical grey lines mark the sample mean of the historical response rates over all 11 trials (solid) and among just the prior MTX trials (dashed).

Figure 2.5.2 illustrates how SPx would adjudicate between its three borrowing strategies over a continuum of possible new trial data, with the observed response rate ranging from (roughly) 10% to 50%. Like in Figure 2.5.1, the new trial's control arm has had previous MTX treatment and has an average age of 53. In panels A and B we can see that the degree of borrowing is largely controlled by how close the new trial's rate is to the overall

historical rate and the MTX historical rate. The borrowing submodels receive their maximum posterior weighting near these rates. As the observed rate diverges from these historical rates the no-borrowing submodel quickly gains posterior weight, especially when the new trial's control group size is larger. Panel C shows how the degree of borrowing affects the Stage 2 sample size calculation, which allows a much smaller control group when the observed rate at interim is close to the predictions of the *hist* and *reg* submodels. When the observed rate at interim differs from these historical predictions by a moderate but not large amount, the design requires a *larger* Stage 2 control group than the target size of 75. Within this range of observed response rates the modest historical borrowing that occurs actually *increases* uncertainty about the new trial's rate compared to not borrowing at all, and so the design calls for collecting more data than otherwise planned.

2.6 Discussion

The SPx method allows flexible borrowing from historical data using a novel Bayesian model averaging approach that balances between direct borrowing from the historical response rates, regression prediction using group-level covariates, and no borrowing at all. The key methodological insight to improve the model averaging is to make the borrowing submodels give strong prior predictions that are well-separated from the no-borrowing submodel while giving high prior weight to the no-borrowing submodel. This strategy offers multiple avenues to not only take advantage of the historical data but also to avoid over-using it when the data suggest this may be unwise. If the group level historical data are relevant to the new trial then under a simple two-stage adaptive design SPx can considerably reduce its control group size, reducing the trial duration and cost.

Our simulation results are in line with the intuitive appeal of the SPx approach and its performance is strong except when its prior assumptions are badly violated and the historical data are by no means useful. Aside from this exception, it produces more or

similarly accurate estimates of the new trial's control group rate and treatment effect while substantially reducing the control group size. We recommend that trialists concerned with dramatically unfavorable scenarios include these in their design simulations and tune the prior or decision rules accordingly as is standard practice.

More generally, trialists wanting to prospectively calibrate the borrowing/no-borrowing tradeoff (i.e. power/Type I error tradeoff) for a specific planned trial may find the best success in tuning one of two method parameters. First, they may experiment with increasing or decreasing p_{ind}, the prior probability of the no-borrowing submodel, as we do in Table A.1. Alternatively, they may experiment with changing the decision rule threshold $q_{nonzero}$ or $q_{clinical}$ for the trial's final treatment effect inference in equations (2) or (3), which would not require fresh simulation for each value if MCMC output has been saved. Both reducing p_{ind} and $q_{clinical}$ or $q_{nonzero}$ would reduce Type I error at the cost of also reducing power.

In general, a variety of factors impact whether a trial setting is likely to benefit and safely reduce control group size by using SPx. The greatest benefit will come when the historical and new trials share data on group-level covariates that are strongly predictive of response rates and when there are enough historical trials to estimate this regression relationship with reasonable accuracy. This may include heavily studied disease areas and drugs (e.g. pembrolizumab) or those where treatments and outcomes have been slow to change (e.g. newly diagnosed glioblastoma). Because the method only requires group-level data, it may be possible to include trials where the patient-level data would not be available due to privacy or proprietary concerns. The group-level covariates might ideally be believed to be strong predictors based on solid theoretical or past empirical evidence, though in settings with enough historical trials it may be possible to incorporate higher-dimensional covariates through the use of sparsity-inducing priors on the regression coefficients. While not the focus of the present work, using the SPx mixture modeling strategy with patient-level data and covariates would likely reduce trial sizes and increase accuracy even more

substantially.

References

Chen, M.-H. and Ibrahim, J.G. (2000) Power prior distributions for regression models. Statistical Science, 15.

Fleischmann, R., Cutolo, M., Genovese, M.C., Lee, E.B., Kanik, K.S., Sadis, S., et al. (2012) Phase IIb dose-ranging study of the oral JAK inhibitor tofacitinib (CP-690,550) or adalimumab monotherapy versus placebo in patients with active rheumatoid arthritis with an inadequate response to disease-modifying antirheumatic drugs: Tofacitinib in Patients with Active RA. Arthritis & Rheumatism, 64, 617–629.

Furst, D.E., Schiff, M.H., Fleischmann, R.M., Strand, V., Birbara, C.A., Compagnone, D., et al. (2003) Adalimumab, a fully human anti tumor necrosis factor-alpha monoclonal antibody, and concomitant standard antirheumatic therapy for the treatment of rheumatoid arthritis: results of STAR (Safety Trial of Adalimumab in Rheumatoid Arthritis). The Journal of Rheumatology, 30, 2563–2571.

Galwey, N.W. (2017) Supplementation of a clinical trial by historical control data: is the prospect of dynamic borrowing an illusion?: Historical control data: is dynamic borrowing an illusion? Statistics in Medicine, 36, 899–916.

Gelman, A., Jakulin, A., Pittau, M.G. and Su, Y.-S. (2008) A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics, 2.

Ghadessi, M., Tang, R., Zhou, J., Liu, R., Wang, C., Toyoizumi, K., et al. (2020) A roadmap to using historical controls in clinical trials – by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). Orphanet Journal of Rare Diseases, 15, 69.

Gravestock, I. and Held, L. (2019) Power priors based on multiple historical studies for binary outcomes. Biometrical Journal, 61, 1201–1218.

Hobbs, B.P., Carlin, B.P. and Sargent, D.J. (2013) Adaptive adjustment of the randomization ratio using historical control data. Clinical Trials: Journal of the Society for Clinical Trials, 10, 430–440.

Hobbs, B.P., Sargent, D.J. and Carlin, B.P. (2012) Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. Bayesian Analysis, 7.

Hunter, T.M., Boytsov, N.N., Zhang, X., Schroeder, K., Michaud, K. and Araujo, A.B. (2017) Prevalence of rheumatoid arthritis in the United States adult population in healthcare claims databases, 2004–2014. Rheumatology International, 37, 1551–1557.

Kaizer, A.M., Koopmeiners, J.S. and Hobbs, B.P. (2018) Bayesian hierarchical modeling based on multisource exchangeability. Biostatistics, 19, 169–184.

Kennedy, W.P., Simon, J.A., Offutt, C., Horn, P., Herman, A., Townsend, M.J., et al. (2014) Efficacy and safety of pateclizumab (anti-lymphotoxin-α) compared to adalimumab in rheumatoid arthritis: a head-to-head phase 2 randomized controlled study (The ALTARA Study). Arthritis Research & Therapy, 16, 467.

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

Lim, J., Walley, R., Yuan, J., Liu, J., Dabral, A., Best, N., et al. (2018) Minimizing Patient Burden Through the Use of Historical Subject-Level Data in Innovative Confirmatory Clinical Trials: Review of Methods and Opportunities. Therapeutic Innovation & Regulatory Science, 52, 546–559.

Lin, J., Gamalo-Siebers, M. and Tiwari, R. (2018) Propensity score matched augmented controls in randomized clinical trials: A case study: Clinical Trial Data Augmentation through Propensity Scores. Pharmaceutical Statistics.

Miyasaka, N. and The CHANGE Study Investigators. (2008) Clinical investigation in highly disease-affected rheumatoid arthritis patients in Japan with adalimumab applying standard and general evaluation: the CHANGE study. Modern Rheumatology, 18, 252–262.

Morita, S., Thall, P.F. and Müller, P. (2008) Determining the Effective Sample Size of a Parametric Prior. Biometrics, 64, 595–602.

Morita, S., Thall, P.F. and Müller, P. (2012) Prior Effective Sample Size in Conditionally Independent Hierarchical Models. Bayesian Analysis, 7.

Murray, T.A., Hobbs, B.P. and Carlin, B.P. (2015) Combining nonexchangeable functional or survival data sources in oncology using generalized mixture commensurate priors. The Annals of Applied Statistics, 9, 1549–1570.

Murray, T.A., Hobbs, B.P., Lystig, T.C. and Carlin, B.P. (2014) Semiparametric Bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data: Semiparametric Bayesian Commensurate Survival Model. Biometrics, 70, 185–191.

Murray, T.A., Thall, P.F., Schortgen, F., Asfar, P., Zohar, S. and Katsahian, S. (2020)

Robust Adaptive Incorporation of Historical Control Data in a Randomized Trial of External Cooling to Treat Septic Shock. Bayesian Analysis.

Neuenschwander, B., Capkun-Niggli, G., Branson, M. and Spiegelhalter, D.J. (2010) Summarizing historical information on controls in clinical trials. Clinical Trials, 7, 5–18.

Neuenschwander, B., Weber, S., Schmidli, H. and O'Hagan, A. (2020) Predictively consistent prior effective sample sizes. Biometrics, 76, 578–587.

Pocock, S.J. (1976) The combination of randomized and historical controls in clinical trials. Journal of Chronic Diseases, 29, 175–188.

Psioda, M.A. and Ibrahim, J.G. (2019) Bayesian clinical trial design using historical data that inform the treatment effect. Biostatistics, 20, 400–415.

van de Putte, L.B.A. (2003) Efficacy and safety of the fully human anti-tumour necrosis factor monoclonal antibody adalimumab (D2E7) in DMARD refractory patients with rheumatoid arthritis: a 12 week, phase II study. Annals of the Rheumatic Diseases, 62, 1168–1177.

van de Putte, L.B.A. (2004) Efficacy and safety of adalimumab as monotherapy in patients with rheumatoid arthritis for whom previous disease modifying antirheumatic drug treatment has failed. Annals of the Rheumatic Diseases, 63, 508–516.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D. and Neuenschwander, B. (2014) Robust meta-analytic-predictive priors in clinical trials with historical control information: Robust Meta-Analytic-Predictive Priors. Biometrics, 70, 1023–1032.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L. and Cowell, R.G. (1993) Bayesian Analysis in Expert Systems. Statistical Science, 8.

Taylor, P.C., Keystone, E.C., van der Heijde, D., Weinblatt, M.E., del Carmen Morales, L., Reyes Gonzaga, J., et al. (2017) Baricitinib versus Placebo or Adalimumab in Rheumatoid Arthritis. New England Journal of Medicine, 376, 652–662.

U.S. Food and Drug Administration. (2019) Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics. Draft Guidance.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., et al. (2014) Use of historical control data for assessing treatment effects in clinical trials. Pharmaceutical Statistics, 13, 41–54.

van Vollenhoven, R.F., Fleischmann, R., Cohen, S., Lee, E.B., García Meijide, J.A., Wagner, S., et al. (2012) Tofacitinib or Adalimumab versus Placebo in Rheumatoid Arthritis. New England Journal of Medicine, 367, 508–519.

Weber, S. (2020). RBesT: R Bayesian Evidence Synthesis Tools. R package version 1.6-1.

https://CRAN.R-project.org/package=RBesT

Weinblatt, M.E., Keystone, E.C., Furst, D.E., Moreland, L.W., Weisman, M.H., Birbara, C.A., et al. (2003) Adalimumab, a fully human anti–tumor necrosis factor α monoclonal antibody, for the treatment of rheumatoid arthritis in patients taking concomitant methotrexate: The ARMADA trial. Arthritis & Rheumatism, 48, 35–45.

Weinblatt, M.E., Mease, P., Mysler, E., Takeuchi, T., Drescher, E., Berman, A., et al. (2015) The Efficacy and Safety of Subcutaneous Clazakizumab in Patients With Moderate-to-Severe Rheumatoid Arthritis and an Inadequate Response to Methotrexate: Results From a Multinational, Phase IIb, Randomized, Double-Blind, Placebo/ Active-Controlled, Dose-Ranging Study. Arthritis & Rheumatology, 67, 2591–2600.

CHAPTER 3

TREATMENT EFFECT ESTIMATION IN MULTI-SITE TRIALS WITH ENDOGENOUS DESIGN: OLD ESTIMATORS, NEW RESULTS

3.1 Introduction

In a multisite randomized field trial, sites such as schools or hospitals are sampled; then, within those sites, individuals are assigned at random to treatments. These studies are now common in social welfare, medicine, and education (Raudenbush & Bloom 2015; Miratrix, Weiss, and Henderson 2021; Spybrook, 2013). It is often useful to regard the multisite trial as a fleet of planned experiments that differ in setting, implementation of the treatment, compliance, and subject demographics (Bloom et al. 2017; Raudenbush & Schwartz 2020; Walters 2015; Weiss et al. 2017). In these cases, we can estimate an average treatment effect specific to each site and it makes sense to regard these treatment effects as varying randomly over sites. Naturally the distribution of treatment effects across sites is of great substantive interest.

In contrast to what the classical theory of randomized block designs assumes (Cochran & Cox 1982), in many large-scale trials it *also* makes sense to regard site-specific design features such as sample sizes and proportions treated as varying randomly over sites. In these trials – where the blocks are typically administrative or geographic units such as schools, hospitals, or neighborhoods – design features are not entirely controlled by the investigator. Instead, they are realizations of an idiosyncratic social process that differs from site to site. For statistical purposes, these design features are conveniently encapsulated in the sampling precision (inverse variance) with which each site-specific treatment effect is estimated (by the difference between the treatment and control group sample means).

Taking this stance also makes obvious the possibility that treatment effects and pre-

cisions covary or are otherwise dependent. In this case we say that the site sizes or the design are "endogenous." This is often plausible when site size is associated with local resources, staff expertise, subject demographics, or other potential moderators of treatment effectiveness. For example, a common practice is to hold a random lottery among applicants at each site to decide who should be offered admission to a novel program (Angrist et al. 2016). Our concern is that the number who apply to each site may reflect the popularity, and thus, indirectly, the effectiveness, of each program. More may apply when local alternatives are limited, reflecting anticipated potential treatment effects (Heckman and Vytlical, 1998). Even in studies that do not use lotteries, sites with different sizes, resources, or participant preferences may vary unexplainably with regard to site sample sizes or proportions treated.

In this paper, we re-evaluate and review how the most common estimators of treatment effects perform under such endogeneity. In particular, we study how such endogeneity changes the performance of four common estimators of the site-average treatment effect (the "site-ATE"). The site-ATE is important for learning about the distribution of treatment effects across sites – a crucial and initial task when there are effectively different implementations of the treatment across sites. This work suggests that endogeneity is *also* important for other multi-site trial estimands (see Raudenbush and Bloom, 2015; Miratrix, Weiss, and Henderson, 2021), where our general approach to comparing estimators can still be used. In general, the combination of heterogeneous treatment effects and an endogenous design can give rise to a challenging bias-variance tradeoff, at least among popular estimators.

3.1.1 Logic of the Evaluation

The first three estimators we discuss are an unweighted or design-based estimator that weights sites equally (UW), a fixed effects estimator that weights sites by their precisions (FE), and random effects estimator having fixed intercepts and random coefficients that weights sites by a stabilized transformation of their precisions (FIRC). Reviews by Kautz,

Schochet, and Tilley (2017) and Miratrix, Weiss, and Henderson (2021) find that these are the dominant estimators in multisite field trials in education, and our informal reading suggests that they are also popular in a wider range of multisite trials in social, economic, and health sciences.

The UW estimator is unbiased under very general conditions, but when the design is imbalanced (which is typical) it can be extremely inefficient. In contrast, the FE and FIRC estimators condition on the observed design in a way that can greatly reduce their variances but introduces bias if the design is endogenous. Our fourth estimator, FIRC+, is not actually common, but is a natural extension of FIRC that includes the log precision in the regression in an attempt to eliminate bias. While model-based like FE and FIRC, it often produces estimates that are more towards the UW end of the spectrum.

To compare these approaches, we need a model for the joint distribution of site-specific treatment effects and precisions. Then comparing the estimators amounts to comparing their marginal distributions, integrating out the precisions. Under a parametric version of our model, we show how these marginal distributions depend on three scale-free parameters that characterize (1) variation in treatment effects (treatment effect heterogeneity), (2) variation in precisions (design imbalance), and (3) the association between treatment effects and precisions (endogeneity of design). In our real data case study we integrate over these three scale-free parameters within a Bayesian framework to compute the posterior probability that each estimator would achieve a smaller mean squared error than would a competitor, given available data, and we study the sensitivity of these conclusions to changes in the prior distributions of those parameters. For the most part our analytic results depend on a parametric model, use asymptotic Laplace-type approximations, and do not directly address the validity of standard confidence intervals, so we use simulations to test and broaden these results. We also identify situations leading to extreme pathology of the conservative unweighted estimator by using Meng & Xie's (2014) concept of self-efficiency to check when this estimator is

improved by throwing out data; these situations are surprisingly plausible.

3.1.2 Motivating Example: Head Start Impact Study

We apply these estimators to the Head Start Impact Study ("HSIS;" Puma et al. 2010) and use our analytic approaches, simulations, and Bayesian framework to illustrate their performance in an important applied setting. Head Start is a federally-funded preschool program for low-income families in the United States since 1965 (Vinoskis 2005). The program now serves nearly 900,000 children and has an annual budget of more than $10 billion (US Department of Health and Human Services 2019). Its congressional mandate requires that Head Start promote the development and learning of low-income children, with a particular focus on the outcomes we analyze: cognitive development in reading, vocabulary, math, and listening comprehension, and socio-emotional development as indicated by disinclination to aggressive behavior. Many more parents seek admission to Head Start than the program can accommodate. From a list of all Head Start centers in the U.S., centers were randomly selected to be included in the study. Within each center, admission to the program was based on a random lottery. A basic first question about Head Start's success is how well the typical center met key program aims. At the same time, HSIS exhibits the conditions suggestive of endogenous sample size. Centers are highly variable, not just in resources and demography but also in curriculum and pedagogy. Centers also vary greatly in size, depending on the number of parents who apply for admission and the number of available places in the local program. These conditions incline us to regard sample sizes and fractions assigned as random variables plausibly correlated with site effectiveness.

3.1.3 Plan of the Paper

In Section 3.2 we propose a simple model for potential outcomes and causal effects in a large-scale multi-site field trial. We define the joint distribution of the treatment effects and

site sizes as a basis for critically examining the estimators of interest. Section 3.3 reviews the asymptotic behavior of the point estimators by using Laplace approximations to characterize their marginal distributions. Section 3.4 evaluates point estimators and confidence intervals by simulation under a broader array of assumptions and Section 3.5 uses the self-efficiency to further characterize the pathology of the unweighted estimator. Section 3.6 analyzes the Head Start data. We assess the self-efficiency of the unweighted estimator, and we propose and illustrate our approach to Bayesian sensitivity analyses that compare the mean squared errors and confidence interval coverage of commonly-used estimators. Section 3.7 concludes with broader implications for experimental design and statistical research on estimation in related trials.

3.2 Potential Outcomes, Causal Effects, Estimands, and Estimators

We envision a multi-site trial in which each person i within site j possesses two potential outcomes $Y_{ij}(t_j)$ for $t_j \in \{0, 1\}$, where $t_j = 1$ if this person is assigned to the potentially unique version of the treatment practiced in site j and $t_j = 0$ if not. We define the person-specific causal effect of assignment to treatment as $B_{ij} = Y_{ij}(1) - Y_{ij}(0)$, and the site-specific average causal effect of assignment, defined over a large super population of persons in site j, is $E(B_{ij}|site = j) =: \beta_j = \mu_{1j} - \mu_{0j}$, where μ_{1j} is the average outcome if all members of this site-specific super-population are assigned to treatment and μ_{0j} is the average outcome if all such members are assigned to control. We envision that a sample of size $n_j$persons in site j is recruited to participate in the experiment. The super-population of interest in our case study is thus composed of Head Start sites with families that actively seek participation in the program. The site-specific sample size may reflect the appeal of the new program to local persons or the lack of local alternative options within that site.

Sample members are then assigned at random to treatment ($T_{ij} = 1$) or control . The

fraction of persons assigned to treatment, $\bar{T}_j = \sum_{i=1}^{n_j} T_{ij}/n_j$, may depend on $n_j$as well the number of available slots in the new program. We will compute the site-specific causal effect estimate $\hat{\beta}_j = \bar{Y}_{1j} - \bar{Y}_{0j}$ where $\bar{Y}_{1j} = \sum_{i=1}^{n_j} T_{ij}Y_{ij}/\sum_{i=1}^{n_j} T_{ij}$ and $\bar{Y}_{0j} = \sum_{i=1}^{n_j}(1 - T_{ij})Y_{ij}/\sum_{i=1}^{n_j}(1 - T_{ij})$ are sample means for treated and control persons in sit j. The estimate $\hat{\beta}_j$ is unbiased for $\beta_j$having sampling variance $V_j := Var(\hat{\beta}_j|\beta_j, n_j, \bar{T}_j)$. Based on our experience with large-scale multi-site trials, we expect $V_j$to be estimated very precisely, and, for simplicity, we assume V_j to be known. Although known, we regard $V_j$as random because it depends on randomly varying sample sizes. Thus, the basic data for our inquiry are $\{\hat{\beta}_j, V_j : j = 1, \ldots, J\}$.

3.2.1 Causal Inference Assumptions

The following assumptions underlie our choice of potential outcomes. First, following Hong and Raudenbush (2006), we have

(A1) **Intact sites**. A person can be a member of one and only one site.

(A2) **No interference between sites**. The potential outcomes of a person in one site do not depend on the treatment assignment of any person in any other site.

(A3) **No interference within sites**. The potential outcomes of a person do not depend on the treatment assignment of other persons in the same site.

(A4) **All units exposable**. Each person is potentially exposable to the treatment and control treatment within that person's site.

(A1) ensures that a person has potential outcomes in only one site. Together, (A2-A4) imply that each person has two and only two potential outcomes within that site. Our assumptions are equivalent to the "Stable Unit Treatment Value Assumption" (Rubin 1986) strictly within each site. One could readily challenge (A3). For example, Head Start instruction occurs

within a classroom setting, and the composition of the classroom will tend to generate peer behaviors and teacher responses that can shape the outcomes of each child. However, we shall assume for simplicity that such effects are baked into the version of the treatment potentially experienced by each child within a given local site.

(A5) **Ignorable Treatment Assignment**. Potential outcomes are independent of treatment assignment, that is $Y_{ij}(0), Y_{ij}(1) \perp T_{ij}$.

Together with (A1-A4), (A5) implies that the sample mean difference $\hat{\beta}_j$ is unbiased for the site-specific mean difference β_j, an assumption that is fulfilled by random assignment within sites.

3.2.2 The Joint Distribution of the Treatment Effects and Site-specific Sample Sizes

We are interested in the joint distribution of treatment effects and sample sizes across a super population of sites. Note that the sample sizes affect inference through the site-specific sampling variance V_j , which we model as $V_j = \mu_V e^{-\eta_j}$, where μ_V is the geometric mean of these sampling variances and $\eta_j$is a zero-mean random variable that we refer to as the log precision of $\hat{\beta}_j$ (though it has also been centered). For deriving our asymptotic results, we will regard V_j as log normal, i.e., $\eta_j \overset{\perp}{\sim} N(0, \sigma_\eta^2)$, but our simulations will relax that assumption. For the treatment effects, we write a simple model $\beta_j = \beta + b_j$ where b_j is a zero mean random effect having variance σ_b^2. We regard the joint distribution of treatment effects and log precisions as

$$\begin{aligned} &\hat{\beta}_j | \, \beta_j, V_j \overset{ind}{\sim} f(\beta_j, \; V_j, \; \ldots) \\ &\begin{pmatrix} \beta_j \\ \eta_j \end{pmatrix} \overset{iid}{\sim} g\left(\begin{bmatrix} \beta \\ 0 \end{bmatrix}, \; \begin{bmatrix} \sigma_b^2 & \sigma_{\eta b} \\ \sigma_{\eta b} & \sigma_\eta^2 \end{bmatrix}, \; \ldots \right) \end{aligned} \tag{3.2.1}$$

Here f is the sampling distribution of $\hat{\beta}_j$, describing how, within each site, the estimate $\hat{\beta}_j$ varies around its mean β_j due to random measurement error and which people were sampled within the site. We avoid the incorporation of covariates for simplicity. The ellipses in f and g are included because conceptually they could depend on additional parameters not specified here. g is some bivariate distribution on the real plane with mean vector $(\beta, 0)^T$, and $\sigma_{\eta b} = Cov(b_j, \eta_j)$. We note that σ_η^2 is a scale-free measure of design imbalance, and $\sigma_{\eta b}$ is a measure of the endogeneity of the design.

3.2.3 Estimands

Looking across an infinite super-population of sites, our key estimand is the site-average treatment effect ("site-ATE"). Given a simple random sample of size J from a list of all sites in the population,

$$E(\beta_j) = \lim_{J\to\infty} \left(\sum_{j=1}^{J} \beta_j / J \right) =: \beta. \tag{3.2.2}$$

Another important quantity is the between-site variance of the site-specific treatment effects, that is

$$Var(\beta_j) = \lim_{J\to\infty} \left(\sum_{j=1}^{J} b_j^2 / J \right) =: \sigma_b^2 \tag{3.2.3}$$

where $b_j = \beta_j - \beta$is a site-specific, zero mean random effect.

3.2.4 Estimators to be Compared

The most commonly used estimators include a simple unbiased and consistent design-based or unweighted estimator and two alternatives that exploit variation in sampling variances V_j to reduce variance at the risk of incurring bias.

Unweighted (UW). The simplest estimator of the site-average mean treatment effect is

$$\hat{\beta}_{uw} = \sum_{j=1}^{J} \hat{\beta}_j / J \tag{3.2.4}$$

and it is unbiased under our assumptions and the definition (3.2.1) regardless of the value of $\sigma_{\eta b}$ and the shape of g. This is the design-consistent estimator if sites constitute a simple random sample.

Ordinary Least Squares with Site Fixed effects (FE). Perhaps the most popular estimator of the average treatment effect in multisite trials (Weiss et al. 2017; Bloom et al. 2017; Miratrix, Weiss, and Henderson, 2021) models the treatment effects as constant using ordinary least squares with site fixed effects. In (3.2.1) this amounts to assuming $\sigma_b^2 = \sigma_{\eta b} = 0$ and normality of f, and defining $P_j := V_j^{-1}$, MLE gives us the fixed effects (FE) estimator

$$\hat{\beta}_{FE} = \sum_{j=1}^{J} P_j \hat{\beta}_j \Big/ \sum_{j=1}^{J} P_j = \sum_{j=1}^{J} e_j^{\eta} \hat{\beta}_j \Big/ \sum_{j=1}^{J} e_j^{\eta}. \tag{3.2.5}$$

Under weaker assumptions this estimator arises from ordinary least squares with fixed effects such that $Y_{ij} = \beta T_{ij} + \mu_{0j} + e_{ij}$, where μ_{0j} is an unknown constant for each site j , and treatment effects are equal in every site. The error e_{ij} is often assumed to have constant variance σ^2 across persons and sites, though more complex models for $Var(e_{ij})$ may be desirable (Bloom et al. 2017).

Fixed intercepts, random coefficients (FIRC). Bloom et al. (2017) expand the fixed effects model with a random coefficient for treatment, writing $Y_{ij} = (\beta + b_j)T_{ij} + \mu_{0j} + e_{ij}$where $b_j$is a mean zero random effect, independent across sites, having variance σ_b^2. FIRC is a simplified linear mixed model that removes potential confounding of random intercepts and treatment effects. Linear mixed models (also known as hierarchical linear models), along with FE have been found to be very commonly used estimators in multi-site trials in

education and behavioral science (Kautz, Schochet, and Tilley 2017; Miratrix, Weiss, and Henderson, 2021). The FIRC estimator can be written as

$$\hat{\beta}_{FIRC} = \sum_{j=1}^{J} \hat{\lambda}_j \hat{\beta}_j \Big/ \sum_{j=1}^{J} \hat{\lambda}_j, \tag{3.2.6}$$

where $\hat{\lambda}_j := \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + V_j}$ and $\hat{\sigma}_b^2$is the MLE σ_b^2 under FIRC when site-specific sample sizes are assumed fixed (see Section B.1 of the Appendix). Note that the $\hat{\sigma}_b^2$ in the numerator of $\hat{\lambda}_j$cancel in (3.2.6) and when $\hat{\sigma}_b^2 = 0$ the estimator is defined as $\hat{\beta}_{FE}$. Thus $\hat{\beta}_{FIRC}$ converges to $\hat{\beta}_{FE}$ as $\hat{\sigma}_b^2 \to 0$ and to $\hat{\beta}_{UW}$ as $\hat{\sigma}_b^2 \to \infty$. In terms of model (3.2.1) this estimator is MLE assuming normality for f, $\sigma_{\eta b} = 0$, and bivariate normality of g. It is also equivalent to the traditional random effects estimator in meta-analysis (Raudenbush & Bryk 1985). In the multilevel modeling literature, $\hat{\lambda}_j$ is known as the estimated site-specific reliability of $\hat{\beta}_j$ as an estimate of the true site-specific effect β_j (Raudenbush & Bryk 2002, Chapter 3). Reliabilities are high when the site-specific treatment effects have large variance relative to the sampling variances, V_j. To see how FIRC relates to FE more explicitly, note also, when $\hat{\sigma}_b^2 > 0$, that $\hat{\beta}_{FIRC} = \sum_{j=1}^{J} e^{\eta_j}(1 - \hat{\lambda}_j)\hat{\beta}_j \Big/ \sum_{j=1}^{J} e^{\eta_j}(1 - \hat{\lambda}_j)$. Like FE, FIRC is biased whenever $\sigma_{\eta b}$ is nonzero.

Fixed intercepts, random coefficients plus (FIRC+). Although not used to date in the literature, we consider a fourth estimator that is a natural modification of FIRC to handle endogeneity by including a linear adjustment for the log precision, η_j. If $\beta_j$and η_j are linearly related, we can remove the bias by estimating the model $\hat{\beta}_j = \beta + \alpha\eta_j + \varepsilon_j + e_j$, where α is a regression coefficient, $\varepsilon_j := \beta_j - (\beta + \alpha\eta_j)$, η_j, and e_j are independently normal with mean zero and constant variance; we call this model FIRC+. In terms of (3.2.1), assuming that both f and g are normal imposes this linear conditional relationship between site-specific ATEs and log precisions. For $\hat{\sigma}_b^2 > 0$, the maximum likelihood estimate (MLE) of the

site-average treatment effect is

$$\hat{\beta}_{FIRC+} = \sum_{j=1}^{J} \hat{\lambda}_j^+ (\hat{\beta}_j - \hat{\alpha}\eta_j) \Big/ \sum_{j=1}^{J} \hat{\lambda}_j^+ \tag{3.2.7}$$

where $\hat{\lambda}_j^+ = \hat{\sigma}_{b|\eta}^2/(\hat{\sigma}_{b|\eta}^2 + V_j)$ with $\hat{\sigma}_{b|\eta}^2$ being the MLE of $\sigma_{b|\eta}^2 = \sigma_b^2(1-\rho_{\eta b}^2)$ with $\rho_{\eta b} = \sigma_{\eta b}/(\sigma_\eta \sigma_b)$ and $\hat{\alpha}$ is the MLE of α. See the Section B.1 of the Appendix for further details. For $\sigma_b^2 = 0$, the MLE is the fixed effects estimator.

3.3 Asymptotic Behavior of the Point Estimators

In any multi-site trial, three main quantities control the comparisons among our four estimators. First, the role of heterogeneity of treatment effects operates largely through the reliability $\tilde{\lambda} = \sigma_b^2/(\sigma_b^2 + \mu_{\tilde{V}})$, where $0 \leq \tilde{\lambda} \leq 1$, which measures in aggregate how accurately the site-specific estimates $\hat{\beta}_j$ estimate the true impact β_j relative to how much these impacts vary. This reliability will be high when site impacts vary greatly and/or when the typical sampling variance, $\mu_{\tilde{V}}$, is small. Second, the role of variation in precisions is captured by σ_η^2, the variance of the (centered) log precision of $\hat{\beta}_j$. Although σ_η^2 can exceed 1, a value approaching 1 would indicate exceptionally large variation in precision. And third, the contribution of endogeneity of precisions is captured by$\rho_{\eta b}$, the correlation between log precisions and effect sizes.

Analytic expressions for the biases and variances of the four estimators are given in Table 3.3.1. Except where noted otherwise, they are approximate in three senses: they are large-sample (as the number of sites $J \to \infty$); they rely on bivariate normality of $\beta_j$and η_j; and they are based on Laplace approximations to the actual asymptotic distributions. Our simulations show that the approximations are accurate under a fairly wide range of assumptions about the joint distribution of β_j and η_j, even for modest J. We provide two approximations for the variance of the FIRC estimator to facilitate clear comparisons to FE

	Estimator	Asymptotic Bias	Asymptotic Variance
UW	$\frac{1}{J}\sum_{j=1}^{J}\hat{\beta}_j$	0	$\sigma_b^2 + \mu_{\tilde{V}} e^{\frac{1}{2}\sigma_\eta^2}$
FE	$\frac{\sum_{j=1}^{J} e^{\eta_j}\hat{\beta}_j}{\sum_{j=1}^{J} e^{\eta_j}}$	$\sigma_{\eta b}$	$\sigma_b^2(1+\rho_{\eta b}^2\sigma_\eta^2)e^{\sigma_\eta^2} + \mu_{\tilde{V}} e^{-\frac{1}{2}\sigma_\eta^2}$
FIRC	$\frac{\sum_{j=1}^{J}\hat{\lambda}_j\hat{\beta}_j}{\sum_{j=1}^{J}\hat{\lambda}_j}$	$\frac{\sigma_{\eta b}(1-\tilde{\lambda}^*)}{1+\sigma_\eta^2\tilde{\lambda}(1-\tilde{\lambda}^*)}$	$(i):\quad \left[\sigma_b^2\left(1+\rho_{\eta b}^2 A\right)e^{v_2^*(1-\tilde{\lambda}^*)^2} + \mu_{\tilde{V}} e^{-\frac{1}{2}v_2^*(1-2\tilde{\lambda}^{*2})}\right]\left(1+\sigma_\eta^2 d_1^{*2} v_2^*\right)^{1/2}$ $(ii):\quad \frac{(\sigma_b^{*2}+\mu_{\tilde{V}})[1+\sigma_\eta^2\tilde{\lambda}^*(1-\tilde{\lambda}^*)]^{\frac{1}{2}}}{e^{\frac{1}{2}v_1^{*2}(1-\tilde{\lambda}^*)^2}}$
FIRC+	$\frac{\sum_{j=1}^{J}\hat{\lambda}_j^+(\hat{\beta}_j-\hat{\alpha}\eta_j)}{\sum_{j=1}^{J}\hat{\lambda}_j^+}$	0	$\frac{(\sigma_{b\|\eta}^2+\mu_{\tilde{V}})[1+\sigma_\eta^2\tilde{\lambda}^+(1-\tilde{\lambda}^+)]^{\frac{1}{2}}[1+v_1^+(1-\tilde{\lambda})^2]}{e^{\frac{1}{2}v_1^+(1-\tilde{\lambda}^+)^2}}$

Table 3.3.1: Asymptotic Bias and Variance of the estimators. For UW the results are finite-sample under (3.2.1), with the variance (normalized by J^{-1}) also requiring marginal normality of η_j. The FE, FIRC, and FIRC+ results are asymptotic under (3.2.1) with bivariate normality, and FIRC and FIRC+ also use the Laplace approximations. As an exception, the FIRC+ bias result is finite-sample and only requires that the conditional expectation of β_j is linear in η_j (as happens under bivariate normality). For full definitions of $\tilde{\lambda}^*$, $\tilde{\lambda}^+$, v_1^*, v_2^*, A, and v_1^+ see the Appendix. Briefly, $\tilde{\lambda}^*$ and $\tilde{\lambda}^+$ are the probability limits (as $J\to\infty$) of $\hat{\tilde{\lambda}}$ under FIRC and FIRC+ respectively, A is near 1, and $v_1^*, v_2^*, v_1^+ < \sigma_\eta^2$.

and FIRC+ respectively. For full details and derivations see the Appendix.

3.3.1 Comparing UW and FE

Constant treatment effects. When $\sigma_b^2 = 0$ and $\eta_j \sim N(0, \sigma_\eta^2)$, UW has a normalized variance of $Var(J^{\frac{1}{2}}\hat{\beta}_{UW}) = \tilde{\mu}_V e^{\frac{1}{2}\sigma_\eta^2}$. In contrast, FE, which is in this case unbiased, has an asymptotic normalized variance of $Var(J^{1/2}\hat{\beta}_{FE}) = \tilde{\mu}_V e^{-\frac{1}{2}\sigma_\eta^2}$. Thus, under these assumptions the asymptotic efficiency (ARE) of FE relative to UW is $e^{\sigma_\eta^2}$. In practice, this ratio can be quite large. For example, we have found that in the national Welfare to Work experiment, $\sigma_\eta^2 \approx 0.9$ (Bloom et al. 2017), yielding an ARE of 1.57. In our own case study of Head Start $\sigma_\eta^2 \approx 0.5$, leading to an ARE of 1.28. That FE dominates UW when treatment effects are constant is no surprise given that, under that assumption, FE is best linear unbiased, but

the loss of efficiency using UW can be substantial in practice under this assumption.

Variable treatment effects. When treatment effects vary and are correlated with sampling precisions, FE can be quite seriously biased. For example, when $\eta_j \sim N(0, \sigma_\eta^2)$, the large-sample bias of FE is $\sigma_{\eta b}$. We will not be able to rule out a non-negligible bias of $\hat{\beta}_{FE}$ in our case study.

Notwithstanding its potential bias, one might suspect that FE would always have a lower variance than UW. However, the opposite can occur. Even when $\rho_{\eta b} = 0$, our large-sample approximations indicate $Var(J^{1/2}\hat{\beta}_{FE}) \geq Var(J^{1/2}\hat{\beta}_{UW})$ if $\tilde{\lambda} \geq \left(1 + e^{\frac{1}{2}\sigma_\eta^2}\right)^{-1}$ (as long as $\sigma_\eta^2 > 0$), a condition that will be met sometimes in practice; for even an extremely imbalanced study with $\sigma_\eta^2 = 1$ this would require $\tilde{\lambda} \geq 0.38$ roughly, which is possible when sites are moderately heterogeneous or large, leading to relatively high precisions. Intuitively, when heterogeneity is high, FE has a higher chance of heavily up-weighting large sites whose true effects happen to be in the tails of g (the cross-site site-ATEs distribution), leading to a more variable estimator.

3.3.2 Comparing FE and FIRC

We generally prefer FIRC to FE for estimating the site-ATE. First, we will show that, under weak assumptions, $Bias(\hat{\beta}_{FIRC}) \leq Bias(\hat{\beta}_{FE})$. Moreover, based on our approximations, the asymptotic variance of FIRC is sometimes much smaller than that of FE.

Comparing the biases. By working directly with the estimators' expectations, it is fairly straightforward to show that in finite samples FIRC is never more biased than FE:

Theorem 3.3.1. *For any number of sites J,* $\left|Bias(\hat{\beta}_{FIRC}, \beta)\right| \leq \left|Bias(\hat{\beta}_{FE}, \beta)\right|$.

This result does not depend on parametric assumptions or asymptotics. Intuitively, the FIRC weights are transformations of the FE weights that are always shrunk towards the constant UW weights, which produce an unbiased estimator. The proof is in the Appendix.

To assess the magnitude of these biases, we see from Table 3.3.1 how the bias of FIRC

is a simple fraction of the bias of FE under our approximations. This fraction becomes smaller as the site-ATEs become more variable relative to their estimates' noise (increasing $\tilde{\lambda}$). The two estimators converge as $\tilde{\lambda} \to 0$ (vanishing treatment effect heterogeneity of effects across sites). As $\tilde{\lambda} \to 1$ (heterogeneity of effects across sites increases relative to$\mu_{\tilde{V}}$) FIRC converges to UW and is thus unbiased. The fraction thus has a maximum of $1 - \tilde{\lambda}$ and diminishes as σ_η^2 increases. In sum, as σ_b^2 and σ_η^2 increase, FIRC's bias reduction relative to FE increases.

Comparing the variances. Table 3.3.1 provides two approximations for $Var(J^{1/2}\hat{\beta}_{FIRC})$. Expression (i) gives us clear insight into how FIRC uses available information efficiently (putting aside bias). For $\tilde{\lambda}$ near 0 FIRC approximates FE. As $\tilde{\lambda}$ increases, indicating heterogeneity of impact, FIRC curbs the tendency of FE to put too much weight on site-treatment effect estimates that are precise but possibly far from the mean. As $\tilde{\lambda}$ approaches 1, FIRC converges to UW. Hence, the FIRC estimator ranges between UW and FE, and in fact in such a way that its asymptotic variance always beats both of theirs (see, e.g., Figure B.7 in the Appendix). Altogether, FIRC asymptotically dominates FE, at least when site effects β_j and log precisions η_j are bivariate normal.

3.3.3 Comparing UW and FIRC and FIRC+

Among the "bias-reducing" estimators, FIRC+ is unbiased and it dominates UW asymptotically (as the MLE) based on our approximations. However, its variance is typically close UW's (see also Figure B.7 in the Appendix); when both σ_η^2 and $\rho_{\eta b}$ are large it can be somewhat more efficient (up to about 15% smaller RMSE), but this small gain comes at the price of extra assumptions, particularly the assumption of a linear association between η and b.

We have seen that FIRC is always at least as good as FE (asymptotically). Also, UW tends to perform similarly to FIRC+ with fewer assumptions. Therefore the comparison

between FIRC and UW becomes particularly interesting.

While its variance is always smaller than that of UW, in endogenous designs FIRC will have a larger MSE than will UW when J grows large enough. This is because FIRC is inconsistent: its bias stays fixed, while all of their variances converge to 0. Given the imbalance in precision as indexed by σ_η^2 , the relative heterogeneity of effect as indexed by $\tilde{\lambda}$, and endogeneity as indexed by $\rho_{\eta b}$, at what "critical" value of Jdoes UW beat FIRC with respect to MSE? The answer, plotted in Figure 3.3.1, is given by the equation

$$J = \frac{Var(J^{1/2}\hat{\beta}_{UW}) - Var(J^{1/2}\hat{\beta}_{FIRC})}{Bias^2(\hat{\beta}_{FIRC}, \beta)}, \tag{3.3.1}$$

where the variances of each estimator are normalized to not depend on J. Below each solid curve FIRC beats UW. When $\tilde{\lambda}$ is small (in the leftmost column) the curves are flatter, so the amount of bias FIRC can bear and still beat the unbiased estimators decreases relatively slowly as J increases. For example, when heterogeneity is low and imbalance is high ($\tilde{\lambda} = 0.1$ and $\sigma_\eta^2 = 0.7$) , if $\rho_{\eta b} = 0.1$ then a study must have well over 350 sites (not plotted) before FIRC loses to UW. Yet when heterogeneity is very high and imbalance is low ($\tilde{\lambda} = 0.7$and $\sigma_\eta^2 = 0.4$), if $\rho_{\eta b} = 0.1$ then a study must only have 200 sites before it loses to UW. Figure 3.3.1 also illustrates where performance of the FIRC and UW estimators diverge significantly (the light gray lines and the color intensity). These regions are characterized by high endogeneity and large J, and the discrepancy enhanced when σ_η^2 is large and $\tilde{\lambda}$ is near 0.5. The Appendix provides a similar figure that includes FIRC+ as well as UW. FIRC+ plus tends to behave similarly to UW except when endogneity is very high.

To summarize the analytic results, first, FIRC is to be preferred over FE. Second, FIRC can be much more efficient than UW and FIRC+ when endogeneity is small or treatment effects are constant or only modestly variable. However, the bias of FIRC can be problematic, especially in trials with many sites. If J is small, the endogeneity must be quite large, perhaps uncommonly so, before FIRC loses to UW and FIRC+ with respect to mean squared error,

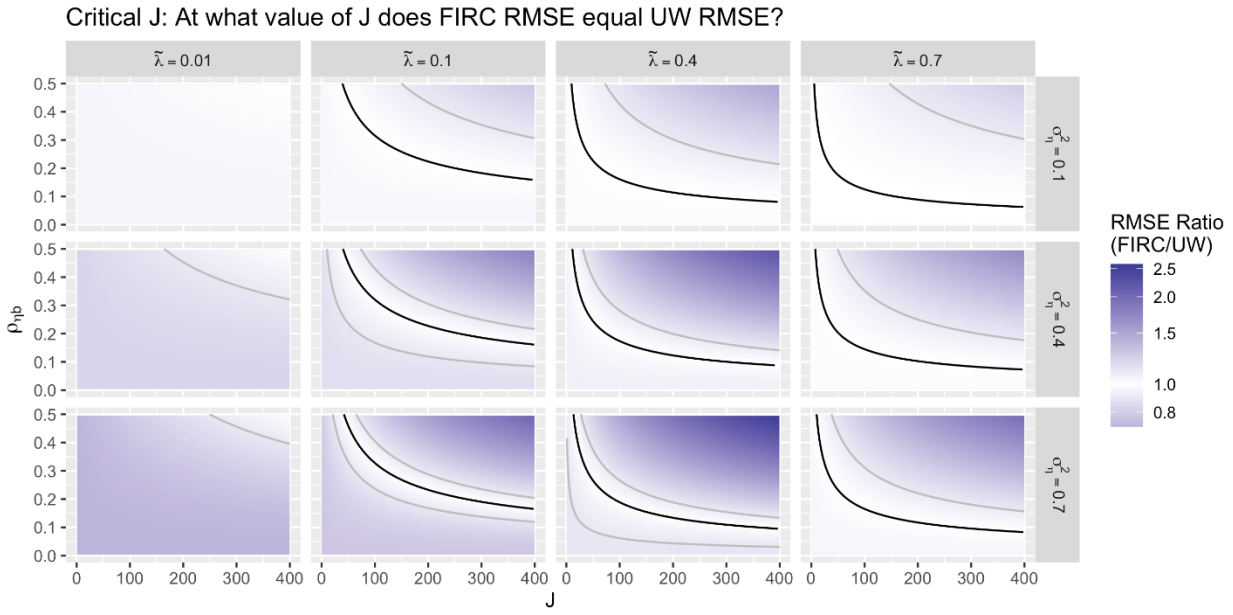

Figure 3.3.1: Curves in J and $\rho_{\eta b}$ where the approximate FIRC RMSE equals the RMSE of UW (solid lines) and FIRC+ (dashed lines). Curves are given for varying $\tilde{\lambda}$(columns) and σ_η^2(rows). For fixed$\tilde{\lambda}$, varying $\tilde{V}$ or $\sigma_b$does not affect the curves (since they cancel out in the ratios). Above the black curves UW beats FIRC in RMSE (as the FIRC bias increases), while below each curve FIRC beats UW. The gray curves show where the ratio of FIRC over UW RMSE are 0.9 (below black curve) and 1.1 (above black curve). The intensity of color shows how far the ratio is from 1 at each point.

and it never loses by much. But if J is large, FIRC can lose, by quite a lot, to UW and FIRC+ when the endogeneity is more moderate (but still appreciable) and heterogeneity is appreciable.

3.4 Simulation Study of Point Estimators and Confidence Intervals

How accurate are these approximations when the bivariate normality assumption fails? When the approximations are not sufficiently accurate, how do the estimators perform relative to one another? To address these questions we simulated trials coming from a wide range of specific cases of model (3.2.1) and design parameters.

Factor	g (joint dist. of site-ATEs and log precisions)	J (# of sites)	σ_b (variance of site-ATEs)	$\mu_{\tilde{V}}$ (geometric mean of sampling variances)	σ_η^2 (imbalance)	$\rho_{\eta b}$ (endogeneity of design)
Levels	Normal, heavy-tailed, nonlinear, heteroscedastic	30, 100, 350	0, 0.1, 0.2, 0.35	0.1^2, 0.3^2, 0.5^2	0.1, 0.4, 0.7, 1	0, 0.1, 0.3, 0.5

Table 3.4.1: Factors varied in the simulation study.

3.4.1 Simulation Setup

Table 3.4.1 gives the specific levels of each factor we vary. We consider studies with all combinations of: small, medium, or large number of sites (J); no, low, moderate, or high heterogeneity of site-average treatment effects (σ_b); low, moderate, or high precision of site-average treatment effect estimates ($\mu_{\tilde{V}}$, which depends mainly on the sample sizes n_j); low, moderate, high, or very high imbalance across sites (σ_η^2); no, low, moderate, and high endogeneity of design ($\rho_{\eta b}$); and four different plausible parametric joint distributions of site-ATEs and log precisions (g). We note that our values for σ_b and $\mu_{\tilde{V}}$ let $\tilde{\lambda}$ range fairly evenly between 0 and 1. Altogether this gives a total of $3 \cdot 4 \cdot 3 \cdot 4 \cdot 4 \cdot 4 = 2304$ distinct scenarios. For each scenario we simulated enough replicates (multisite trial data sets) needed to get bias and variance estimates accurate to two decimal places; this ranged from 500 to nearly 100,000 replicates across scenarios.

We specify g through the marginal distribution of η_j and the conditional distribution of β_j given η_j. For full details see Section B.5 of the Appendix, but briefly in all four cases we made $\eta_j \overset{iid}{\sim} N(0, \sigma_\eta^2)$ and then varied the conditionals (and hence the marginal distribution

of the site-ATEs β_j). The conditional distributions have unequal conditional variances, each scaled so the marginal variance σ_b^2 is the same (so estimators' MSEs are comparable across scenarios). In the normal case, the FIRC+ model is correct. In the heavy-tailed case, where $\beta_j = \beta + \alpha\eta_j + \kappa_j$ with ${}_j \overset{iid}{\sim} t_3$, especially effective/ineffective sites are more common than under normality. In the nonlinear case, where $\beta_j|\eta_j \overset{\perp}{\sim} N(\beta+\alpha\eta_j+\gamma(\eta_j^2-\sigma_\eta^2),\ \tau^2)$, the positive association between log precision and treatment effect grows as log precision increases. In the heteroscedastic case, where $\beta_j|\eta_j \overset{\perp}{\sim} N(\beta + \alpha\eta_j,\ \tau^2 e^{-\frac{1}{2}\eta_j})$, sites with higher precisions have less variable effects.

3.4.2 Root Mean Squared Error of Point Estimators

As detailed in Appendix Section B.6, the simulation reveals that all of the RMSE approximations are typically very good when $\beta_j|\eta_j$ is conditionally normal or heavy-tailed, even when there are relatively few sites ($J = 30$). Not surprisingly the FIRC and (especially) FE approximations deteriorate as nonlinearity or heteroscedasticity increase and when σ_b^2 is nonzero; the FIRC+ approximations also suffer in the nonlinear case (though not much in the heteroscedastic case). The Monte Carlo error is such that we can trust the simulated RMSEs to be off from the truth by up to about 5% in either direction.

Simulation results for the relative efficiency of the estimators are shown in Figure 3.4.1, which plots the ratio of each precision-weighted estimator's RMSE to that of UW as $\tilde{\lambda}$ (x-axis), σ_η^2 (column), and $\rho_{\eta b}$ (color and line) vary in quite strongly heteroscedastic and nonlinear cases when $J = 100$. The normal (of course) and heavy simulation echo the asymptotic approximations very closely, so we leave them to the Appendix. When $J = 30$ and $J = 350$ the estimator relationships are qualitatively similar to those shown here so we leave these figures to the Appendix as well. The heteroscedastic and nonlinear simulation results reveal the following nuances. As previously emphasized, these scenarios may not arise frequently in practice, but they are important departures from normality to be aware of.

FIRC vs. FE. FE can beat FIRC in the heteroscedastic case when $\rho_{\eta b}$ is low, especially when imbalance σ_η^2 and heterogeneity $\tilde{\lambda}$ are high. This happens because the heteroscedastic case we simulated gives an additional reward to precision weighting (since in this particular case, the more precise sites also have less variable effects β_j) and FE's precision weighting is more aggressive than FIRC's. However, when $\rho_{\eta b}$ is nonzero and moderate FE's bias can still be much larger than FIRC's and the extra variance reduction is dwarfed. In the nonlinear case FE's bias is exacerbated more than FIRC's is, so it loses by an even wider margin than in the normal case.

FIRC+ vs. UW. While our asymptotics suggest that FIRC+ universally dominates UW, of course this is not the case in the nonlinear setting, which violates the FIRC+ regression assumption and gives it a bias that stays fixed as the sample size grows. In particular, we can see that FIRC+ can become worse than UW (by up to 40% when $J = 100$, and up to 130% when J = 350) in the nonlinear case when $\tilde{\lambda}$ is not very small or very large (i.e. bounded away from 0 and 1) and there is at least moderate imbalance.

FIRC vs. UW and FIRC+. In the strongly nonlinear case, FIRC only beats UW substantially if $\tilde{\lambda}$ is 0 or very close, and otherwise it only beats UW if $\rho_{\eta b}$ is small and then by a fairly modest margin (10-20% in our scenarios). But FIRC is less biased than FIRC+ when $\rho_{\eta b}$ is not large. The heteroscedastic case reduces FIRC's and FE's variance, leading FIRC to beat UW and FIRC+ by a larger margin when there is little endogeneity. An antagonistic heteroscedastic case (where the more precise or larger sites have *more* variable effects) would not make FIRC look so good.

3.4.3 Coverage of Confidence Intervals

To construct confidence intervals for the ATE, the standard strategy most analysts would default to is to center the interval at the preferred method's point estimate and then find bounds by subtracting and adding the appropriate multiple of an estimated standard error.

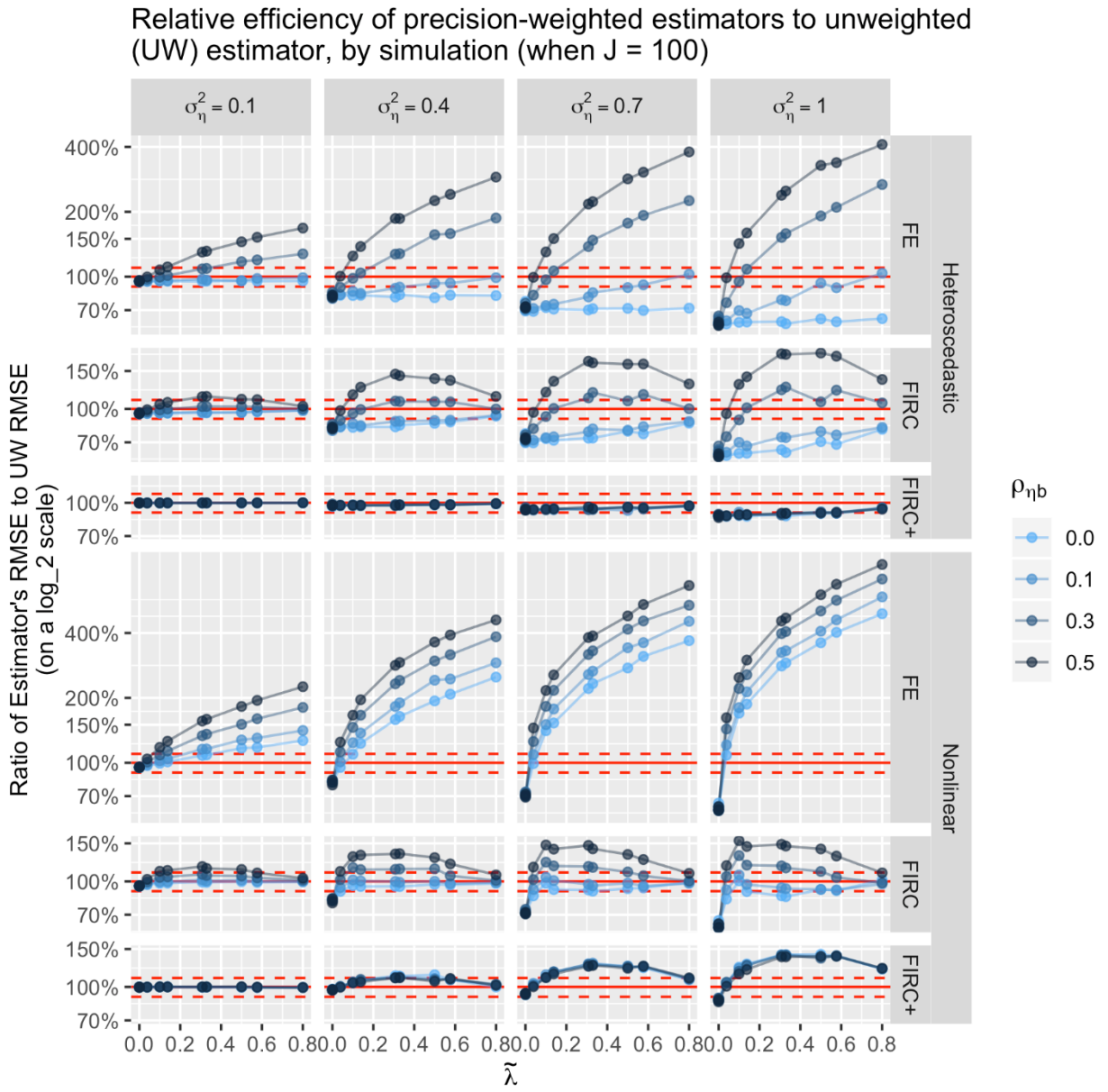

Figure 3.4.1: The ratio of each estimator's RMSE relative to the RMSE of UW, as $\widetilde{\lambda}$ varies on the x-axis, for $J = 100$ in the heteroscedastic and nonlinear cases. Separate plots are given for each σ^2_η(columns) and estimator/distribution (rows). A RMSE ratio of 100% is shown by the horizontal solid red lines, while the dashed red lines are at 110% and 90% to demarcate cases where estimators perform comparably to UW. The color of each point corresponds to the correlation $\rho_{\eta b}$. These ratios only depend on σ^2_b and $\mu_{\tilde{V}}$ (not shown) through $\tilde{\lambda}$.

However, centering a confidence interval at a biased point estimator is asking for trouble. Figure 3.4.2 illustrates this issue through simulation results from the normal case (see the Appendix Section B.6 for the others). When the bias of FIRC and FE grow, controlled largely by $\sigma_{\eta b}$ along the x-axis, coverage of the standard FIRC and FE confidence intervals plummets far below the nominal 95% level. The FIRC+ interval also suffers from poor coverage in the strongly nonlinear case, where it is also biased. The problem is exacerbated by sample size, since as J grows the standard errors shrink and these confidence intervals will converge onto the wrong (biased) ATE. Even if the bias is small, in any real study we will not know exactly where in this downward spiral the coverage will fall – it may be close to nominal or embarrassingly lower. The conclusion is simple: to accurately measure uncertainty using the simplest type of confidence interval, one must be conservative and use an unbiased estimator (UW, or FIRC+ if trusted).

What recourse do researchers have? One possibility is to report a different point estimate (e.g. FIRC) than the one used to construct the confidence interval (UW or FIRC+). This invites the chance that the point estimate is not in the confidence interval, though in practice this would be extremely rare because the four estimators are highly correlated and except in unrealistically large (or precise) trials the UW or FIRC+ confidence intervals will be more than wide enough to cover the FIRC and FE point estimates. Another is to construct more sophisticated intervals that are likely shorter but still guarantee their advertised level of coverage (though perhaps asymptotically). For example, one might adapt a FAB (Frequentist assisted by Bayes) method (Yu & Hoff 2018) to this setting or use bias-corrected bootstrap methods like BCa (Efron 1987).

3.5 Self-Inefficiency of the Unweighted Estimator

Our asymptotic and simulation results highlight that while UW's guaranteed unbiasedness may be important in some settings, the price is that it can use the trial data very inefficiently,

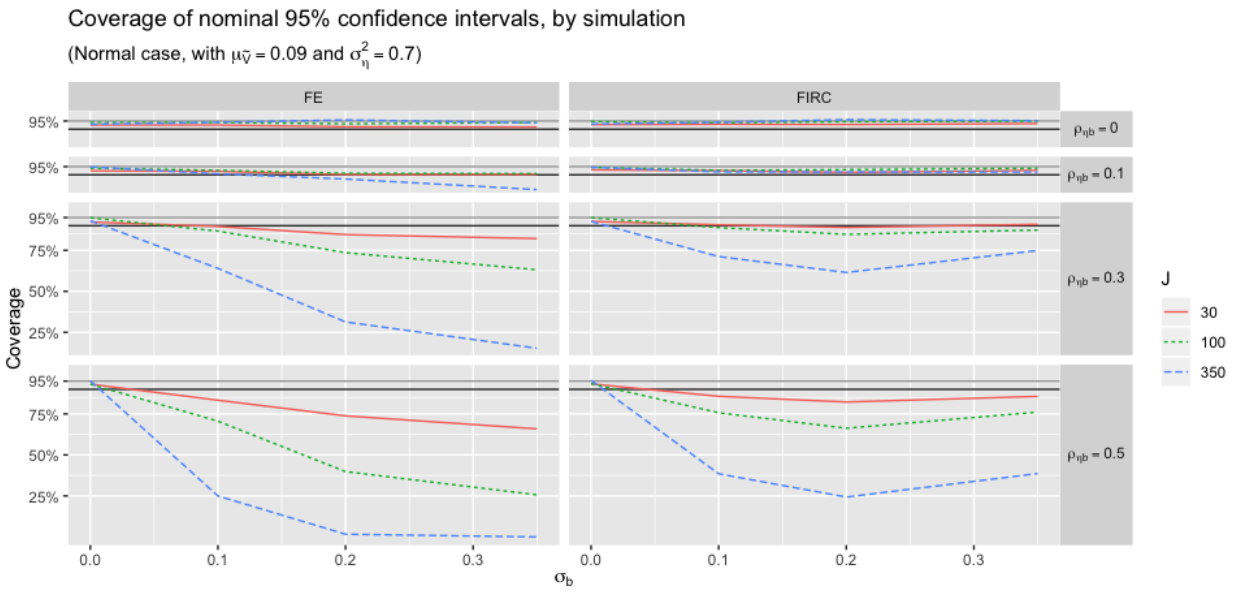

Figure 3.4.2: Actual coverage rates (by simulation) for the standard nominal 95% confidence intervals centered on the different point estimators, as a function of σ_b (x-axis). Plots are given for FIRC and FE (columns) and values of $\rho_{\eta b}$ (rows). Each colored line shows scenarios with a different sample size$J$. The horizontal light grey lines mark the nominal 95% rate, and the dark grey lines mark 90%. The data where simulated from the normal case (as in the simulation study in Sections 4), with $\mu_{\tilde{V}} = 0.09$ and σ^2_η, though changing these parameters does not change the results qualitatively (i.e. the coverage can still be bad).

especially if there is substantial imbalance in design across sites. In this Section we consider another perspective on UW's use of trial data by asking if, and when, UW can be improved by throwing out data. This is a form of the *self-inefficiency* property coined by Meng and Xie (2014), and we derive conditions describing when UW suffers from this odd problem.

Consider discarding some sites and recomputing the unweighted estimator on the remaining subset $S \subset \{1, \ldots, J\}$ to get the subset estimator $\hat{\beta}_S := s^{-1} \sum_{j \in S} \hat{\beta}_j$ where $s := \sum_{j=1}^{J} \mathbb{1}\{j \in S\}$. We refer to $\hat{\beta}_S$ as the subset estimator and focus on the case where S is the subset of sites with the s most precise site-specific effect estimates. If this subset estimator has smaller mean squared error than UW, then we say that the original estimator is *self-inefficient*.

Lack of self-efficiency indicates that a method makes poor use of available data. How-

ever, we stress that self-efficiency is likely a basic quality of reasonable estimators and not necessarily a sign that an estimator is optimal in apparently stronger senses. Self-inefficiency of the "obvious" unbiased estimator in a given setting, like the UW estimator of the site-ATE in multisite trials, is a sign that the problem may suffer from a difficult bias-variance tradeoff and may be easier to check than more fully characterizing the tradeoff over a class of estimators. We show later that, in the case of Head Start it is hard to deny that the unweighted estimator is self-inefficient, casting doubt on the value of "bias-free" answers despite that they are often prized (Schochet 2016).

The following theorem describes how $\hat{\beta}_{UW}$ may be self-inefficient if the study design is sufficiently imbalanced across sites. We define $\pi_S := s/J$ (the fraction of sites being kept for the subset estimator), $\sigma^2_{b\in S} := Var(\beta_j | j \in S)$ (the variance of site effects in the subset), $\Delta_S := E(\beta_j | j \in S) - E(\beta_j | j \notin S)$ (the difference in expected effects between kept and discarded sites), and $C_{\in S} := Cov(\beta_j, \beta_k | j \in S, k \in S)$ (the covariance of effects in different kept sites). Recall that we consider only subsets where we keep the s most precise sites, that is the sites with $\eta_j > \eta_{(J-s)}$ where $\eta_{(J-s)}$ is the $(J-s)$-th order statistic. Thus the subset is random (as a function of the η_j) but has fixed size.

Theorem 3.5.1. *Under* (3.2.1)*, suppose that for some* $s \in \{1, \ldots, J-1\}$ *the set of the* s *most precise sites,* $S := \{j \in \{1, \ldots, J\} : \eta_j > \eta_{(J-s)}\}$*, satisfies*

$$E(V_j | j \in S) \leq \pi_S E(V_j) - \left[\sigma^2_{b\in S} - \pi_S \sigma^2_b\right] - s(1-\pi_S)^2 \Delta^2_S - (s-1) C_{\in S} \tag{3.5.1}$$

or, equivalently,

$$\Delta^2_S \leq (1-\pi_S)^{-2} s^{-1} \left(\left[\pi_S E(V_j) - E(V_j | j \in S)\right] + \left[\pi_S \sigma^2_b - \sigma^2_{b\in S}\right] \right) - (1-\pi_S)^{-2} \frac{s-1}{s} C_{\in S}. \tag{3.5.2}$$

Then the unweighted estimator $\hat{\beta}_{UW}$ *is self-inefficient as an estimator of* β*.*

See Section B.7 of the Appendix for the proof, which is mostly straightforward com-

putation. Intuitively, condition (3.5.1) is met when the typical sampling variance in the subset (i.e. the average sampling variance excluding the $J - s$ largest ones), $E(V_j|j \in S)$, is small relative to the typical sampling variance over all sites, $E(V_j)$. (3.5.2) describes this condition in terms of how large the difference in average treatment effects in the kept and discarded sites, Δ_S, may be (controlled indirectly by the covariance $\sigma_{\eta b}$). Both show how self-inefficiency is less likely when the site effects are variable enough in the subset ($\sigma^2_{b\in S} > \pi_S \sigma^2_b$). The covariance term $C_{\in S}$, included here for completeness, is typically negligible and often provably zero (see Section B.7 of the Appendix) so we exclude it from the following discussions. Because Theorem 3.5.1 is a finite-sample result that requires weak assumptions it can be used for rather reliable sensitivity analysis to assess whether the "conservative" UW estimator can be demonstrably improved on in practice, as we do in Section 3.6.

3.6 Head Start Impact Study

The effectiveness of Head Start has long been subject to debate, and in 1998 the US Congress mandated a randomized evaluation, the Head Start Impact Study (Puma et al. 2010). After drawing a sample of 351 Head Start centers, the experimenters sought applications from eligible families. Because the number of places was limited, they conducted a lottery to make offers of admission. The resulting sample included 4,400 3-4 year-old children. Restricting our analysis to (a) centers with at least one experimental and one control child and (b) children with complete data on our outcomes, our analytic sample includes 3,392 students in 316 centers (sites). Although an optimal analysis would likely use multiple imputation, we prefer to avoid questions of missing data here in order to clarify key methodological issues in this study, which focus on alternative estimators. We analyze the first year of outcome data, collected in the spring of 2003.

Head Start is mandated to promote learning and socioemotional development among

low-income preschoolers. We therefore study early reading (Woodcock Johnson III Letter-Word Identification), math (Woodcock Johnson III Applied Problems), and vocabulary (Peabody Picture Vocabulary Test or PPVT) in the cognitive domain. For the socioemotional domain, we study a measure of aggressive or "externalizing" behavior (the Child Behavior Problems Index). Oral comprehension (Woodcock Johnson III Oral Comprehension) requires a mix of cognitive and attentional skill.

For each outcome, we estimate the site-average effect of treatment assignment. Head Start Centers vary substantially in program design, resources, and geographic context. The distribution of treatment effects across such heterogeneous programs is of interest and the mean of that distribution is the site-ATE. We also consider the variance of that distribution. Many others have focused on a population of sites, even if implicitly through their use of standard linear mixed models, including Feller et al. (2006), Bloom & Unterman (2014), and Walters (2015). Our focus on the effects of treatment assignment, often the called intent-to-treat analysis, reflects the fact that some admitted children went to another or no preschool, and some rejected children attended Head Start after getting in off a wait list. The effects of assignment are relevant to a policy environment in which not all who receive an offer will participate. HSIS also exhibits the conditions discussed in the introduction that invite the bias-variance tradeoff.

Heterogeneity of impact. Program features such as teacher-staff ratios, staff qualifications, and facilities vary significantly across program sites (Puma et al. 2010). Sites also vary in key aspects of program design, including curriculum, teacher evaluation and training, director qualifications, and the guiding educational philosophy of the 90 agencies that operate the program. Past studies have found significant variation – even within an agency – in how staff implement the program, their geographic settings, local labor market conditions, supply of teachers, and the social, ethnic, and cultural background of parents and children (Puma et al. 2010). So variation in treatment effects is highly plausible.

Design-estimand mismatch and endogeneity of design. Our target population is the population of sites, but the design is highly imbalanced across sites, so equal weighting will be inefficient. HSIS planners initially sought to study sites with an average of 27 children (with 16 assigned to treatment and 11 to control), but the sites ended up being smaller on average (with a mean site size $\bar{n} = 10.7$ in our analytic sample) and quite variable in size (with n_j ranging from 2 to 74). Proportions treated also varied across sites ($\bar{T}_j$ ranging from 20% to 89%, with a mean of 62%). This design reflects the dramatic variation across centers in the number of applicants in competitive lotteries. Varied sample sizes could represent variation in the effectiveness of local recruiters, in the prestige and perceived effectiveness of the local Head Start program, in the desirability of local child-care alternatives, or in the size of the center relative to the size of the local population. Because these factors may plausibly be related to the actual effectiveness of each center (β_j) we cannot ignore the possibility that the design is endogenous. Putting these three conditions together promotes uncertainty in how to analyze this study.

3.6.1 Empirical Results

Tests of significance. As indicated in Table 3.6.1, all estimators indicate significant positive effects of assignment to Head Start on early reading achievement and receptive vocabulary and no estimator suggests a non-zero effect on oral comprehension. However, the estimators for the other two outcomes produce ambiguity about impacts. In particular, FIRC and FE suggest statistically significant effects on math achievement and aggression while the UW analyses suggest rejecting the null hypothesis in neither case. FIRC+ would lead us to reject the null hypothesis for math but not aggression. These results are based on two-sided t-tests (with 315 degrees of freedom) of a null hypothesis of no average treatment effect. Emulating an analyst who would use a single method of estimation for all five outcomes, these results control for multiple testing using the popular Benjamini-Hochberg (Benjamini & Hochberg

1995) method to control the false discovery rate at 3% across outcomes (separately for each method).[1] Point estimates across the four methods are similar, and the apparent differences in hypothesis rejections reflect smaller standard errors associated with precision-weighted methods. This may seem to imply that precision-weighting is superior, but if FIRC and FE (and FIRC+) are biased, their tests will not achieve the nominal level. We will apply our tools for sensitivity analysis to explore these discrepancies in findings.

Estimated magnitude of treatment effects. For the significant outcomes, the estimators mostly agree on the magnitude of the site-ATE in terms of both point estimates and confidence intervals. Point estimates of the treatment group difference for reading are 0.18 to 0.19 standard deviation (sd) units of the outcome variable, suggesting that a typical child assigned to Head Start would outscore about 57 to 58 percent of the members of the control group. Compared to other important educational evaluations, this effect is large enough to be substantively important. It is, for example, very similar to the effects produced for reading in the famous Tennessee study of class size reduction (Finn & Achilles 1991). For vocabulary, the point estimates are 0.11-0.12 sd units, indicating that a typical child assigned to the experimental groups would outscore about 54 percent of the control group members. Point estimates of impact for math and aggression are a bit smaller (about 0.08 and -0.08 sd units, respectively). One way to gauge the practical significance of the reading impact is to note that assignment to Head Start produces a gain equivalent to about 14 weeks of learning, assuming a typical learning rate of 0.014 sd units per week. To obtain this estimated rate, we regressed the outcome on age among 3 and 4 year-olds, holding constant child fixed effects. For math, the gain would be 6-7 weeks. (Such a metric is not possible for vocabulary using PPVT, which is standardized by age).

1. We chose this level because under multiple testing it exacerbates differences between the estimators, an interesting problem that may crop up in other studies, especially with more outcomes. Marginally, the "conventional" 5% level leads to the same set of rejections as does 3%.

	UW	FIRC+	FIRC	FE
Reading	0.173*	0.180*	0.193*	0.194*
	(0.045)	(0.043)	(0.037)	(0.037)
Math	0.092	0.096*	0.080*	0.079*
	(0.048)	(0.043)	(0.032)	(0.032)
Oral	-0.030	-0.061	-0.026	-0.027
	(0.040)	(0.039)	(0.029)	(0.029)
Behavior	-0.080	-0.081	-0.076*	-0.076*
	(0.047)	(0.043)	(0.034)	(0.034)
Vocabulary	0.123*	0.125*	0.109*	0.109*
	(0.040)	(0.038)	(0.032)	(0.032)

Table 3.6.1: Point estimates of the site-ATE for the five basic outcomes in HSIS under each method. Standard errors are given in parentheses. The asterisks denote significance at a 3% FDR level using the Benjamini-Hochberg procedure within each method (over the 5 outcomes). The 3% level was chosen to highlight the potential for different testing decisions across methods.

3.6.2 Sensitivity Analyses

To better understand how the four basic estimators perform in HSIS, we consider how the results of Sections 3 through 5 may apply to this study. In particular we want to assess, given the Head Start data, (a) how self-inefficient UW could plausibly be, (b) what the analytic approximations suggest about FIRC's RMSE compared to UW, and (c) whether or not the FIRC confidence intervals may plausibly have unacceptable coverage. The framing of these sensitivity analysis questions is by definition Bayesian, so we take a Bayesian approach to answering them. We pose a fully Bayesian version of model (3.2.1), considering a range of priors, and compute the posterior distribution of the model parameters, which translates directly into the posterior evidence about each question. Throughout we focus on the reading and math outcomes as they provide an interesting contrast.

3.6.2.1 Bayesian Framework and Key Parameter Inferences

Putting priors on the parameters of model (3.2.1) is obviously in some sense a subjective exercise, as is any sensitivity analysis that makes judgments about what regions of a parameter space are worth considering. In general our priors rule out implausibly large parameter values but are informative about little else, aside from a bivariate normality assumption. In particular, our main sensitivity analyses are based on the following Bayesian version of (3.2.1), except where noted otherwise:

$$\begin{aligned}
\hat{\beta}_j|\beta_j, V_j &\overset{ind.}{\sim} N(\beta_j, V_j) \\
\eta_j &= \log \mu_{\tilde{V}} - \log V_j \\
\begin{bmatrix} \beta_j \\ \eta_j \end{bmatrix} |\beta, \sigma_b^2, \rho_{\eta b}, \sigma_\eta^2, \mu_{\tilde{V}} &\overset{iid}{\sim} N\left(\begin{bmatrix} \beta \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_b^2 & \rho_{\eta b}\sigma_b\sigma_\eta \\ \rho_{\eta b}\sigma_b\sigma_\eta & \sigma_\eta^2 \end{bmatrix} \right) \\
\sigma_b &\sim TCauchy(0, 0.065, (0, \infty)) \\
\sigma_\eta &\sim TCauchy(0, 0.24, (0, \infty)) \\
\rho_{\eta b} &\sim TN(0, 0.3, (-1, 1)) \\
\log \mu_{\tilde{V}} &\sim N(0, 10^2) \\
\beta &\sim N(0, 0.5^2)
\end{aligned} \tag{3.6.1}$$

where $N(m, v)$ indicates a normal distribution with mean (vector) m and (co)variance (matrix) v, $TCauchy(c, s, I)$ indicates a Cauchy distribution with location c and scale s that has been truncated to the interval I, and $TN(m, v, I)$ indicates a normal distribution with mean m and variance v that has been truncated to the interval I. The bivariate normality of β_j and η_j makes (3.6.1) akin to FIRC+ in its linearity and homoscedasticity assumptions, points we touch on at the end of this section. The variance components use half-Cauchy priors following Gelman (2006) and Polson & Scott (2012), which have slowly degrading tails to not hamper inference if the true variances are quite large. The hyperparameter values were chosen so that σ_b has a $\sim$10% prior probability of being greater than 0.4 (a very high

level of heterogeneity), σ_η has a $\sim$15% prior probability of being greater than 1 (extremely high imbalance), $\rho_{\eta b}$ has a $\sim$95% prior probability of being in (-0.6, 0.6). The prior for $\mu_{\tilde{V}}$ is more or less flat on the log scale, and the prior for β is weakly informative in that it rules out extremely large effect sizes (given that the outcome has been standardized).

Since σ^2_η and $\mu_{\tilde{V}}$ are estimated with fairly high information from the data, we only considered alternative priors for σ^2_b and $\rho_{\eta b}$. For σ^2_b, we tried both priors putting more mass on large levels of heterogeneity ($TCauchy(0, 0.4, (0, \infty))$, which has a 50% probability of $\sigma_b > 0.4$, or $TCauchy(0.75, 0.01, (0, \infty))$, which has a 99% prior probability of $\sigma_b > 0.4$) as well priors putting more mass on low levels of heterogeneity. Neither greatly affect our sensitivity analyses, so we do not report those results. On the other hand, the prior for $\rho_{\eta b}$ *does* matter, and we discuss results for a range of alternatives in Figure 3.6.2. At the end of the section we discuss the possibility of making larger departures from the underlying model.

To preview the results of this analysis, Table 3.6.2 gives posterior means and credible intervals for the key sensitivity parameters as well as some hints about the performance of the estimators across the different Head Start outcomes. The first major point here is that for all outcomes there appears to be very little evidence about the magnitude of any endogeneity of design given the very wide intervals for $\rho_{\eta b}$. This point is fleshed out further in Figure 3.6.2, which we will discuss shortly in the sensitivity analysis for the analytic approximations. Naturally σ^2_η is well estimated (given the large J) and inferences are essentially identical across outcomes because this parameter is scale-free. Finally, looking to the reading and math outcomes, we can see that for reading there is evidence of fairly substantial heterogeneity of site effects whereas for math this heterogeneity is likely more modest. This intuitively suggests (based on the results of Sections 3 and 4) that bias is more likely to be a concern for the reading outcome than the math outcome. We examine this intuition more formally in the next subsections.

	Reading	Math	Oral	Behavior	Vocabulary
$\rho_{\eta b}$	0.06 (-0.40, 0.52)	-0.04 (-0.59, 0.49)	-0.04 (-0.59,0.49)	0.00 (-0.56, 0.55)	-0.07 (-0.61, 0.47)
σ^2_η	0.58 (0.49, 0.67)	0.57 (0.49, 0.67)	0.58 (0.49, 0.67)	0.58 (0.49, 0.67)	0.58 (0.49, 0.67)
$\tilde{\lambda}$	0.04 (0.00, 0.12)	0.01 (0.00,0.04)	0.01 (0.00,0.05)	0.01 (0.00, 0.03)	0.01 (0.00,0.06)
σ_b	0.11 (0.00, 0.25)	0.05 (0.00, 0.13)	0.05 (0.00, 0.13)	0.05 (0.00, 0.13)	0.05 (0.00, 0.15)

Table 3.6.2: Posterior means and 95% credible intervals (in parentheses) for the key sensitivity parameters controlling the operating characteristics of the site-ATE estimators, for each of the Head Start outcomes. The credible intervals are highest posterior density, and note that for $\tilde{\lambda}$ and σ_b the lower bounds are only written as 0 after rounding.

3.6.2.2 Self-inefficiency of the UW Estimator

Figure 3.6.1 shows sensitivity analyses to check for self-inefficiency of the UW estimator in HSIS, as described in Section 3.5. We see strong evidence of self-inefficiency. When discarding only one site (right where the curves approach the y-axis), $|\Delta_S|$ must be larger than 0.68 or 0.64 in order to ensure self-efficiency in reading and math respectively; in other words the least precise site must have an site-specific ATE more than 0.64 (or more) better or worse than the ATE of all the others, an extreme proposition. Further, if precisions are uncorrelated with site-specific ATEs then we could discard 45-64% of sites (see the x-intercepts) and still achieve the same MSE using UW. We conclude that the UW analysis, which produces unbiased point estimates and honest confidence intervals under weak assumptions, is almost surely self-inefficient, and possibly quite seriously so. One would hope we can do better with a different estimator.

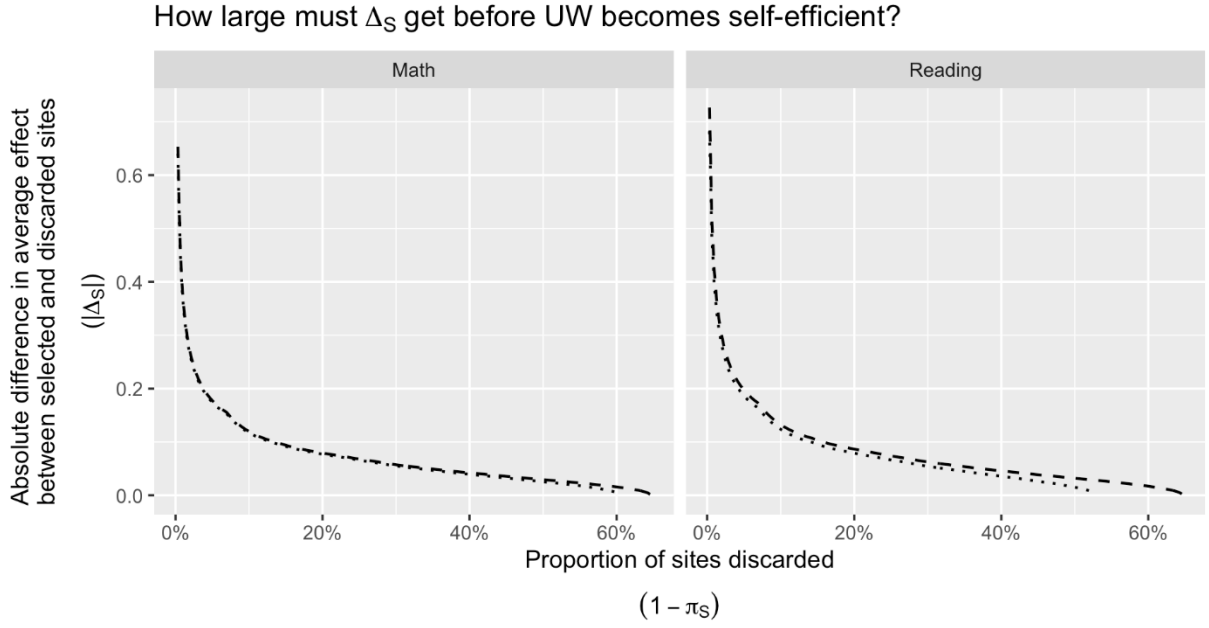

Figure 3.6.1: Sensitivity analyses for self-efficiency in the case studies. Each panel shows how large $|\Delta_S|$ can be before the UW estimator is self-efficient (from the RHS of (3.2)), as a function of the fraction of sites being discarded $(1 - \pi_S)$. Above the curves UW is self-efficient and below them it is self-inefficient. We hold σ_η^2 constant at its point estimate 0.57 since the curves are qualitatively similar across its 95% confidence intervals. The dotted and dashed curves are for σ_b at the bounds of its 95% CI from FIRC; $\sigma_{b \in S}$ is fixed at its conservative lower bound $\pi_S \sigma_b$ (harder to find self-inefficiency).

3.6.2.3 RMSE Approximations

As already alluded to in Table 3.6.2, there appears to be little evidence in the HSIS data about the magnitude of $\rho_{\eta b}$. The top portion of Figure 3.6.2 underlines this point by showing that posterior distribution for $\rho_{\eta b}$ essentially reproduces its prior, over a range of substantively diverse alternative prior distributions (top row). While dramatic, this is perhaps not surprising given the small site sizes and relatively high noise levels for $\hat{\beta}_j$ in HSIS.

The bottom portion of Figure 3.6.2 shows the induced posterior distributions of the RMSE ratio between FIRC and UW for the four alternative priors on $\rho_{\eta b}$. For math, essentially regardless of prior beliefs about $\rho_{\eta b}$, the data and model imply that bias is not plausibly big enough to make FIRC's RMSE worse than UW's. This is because the strong

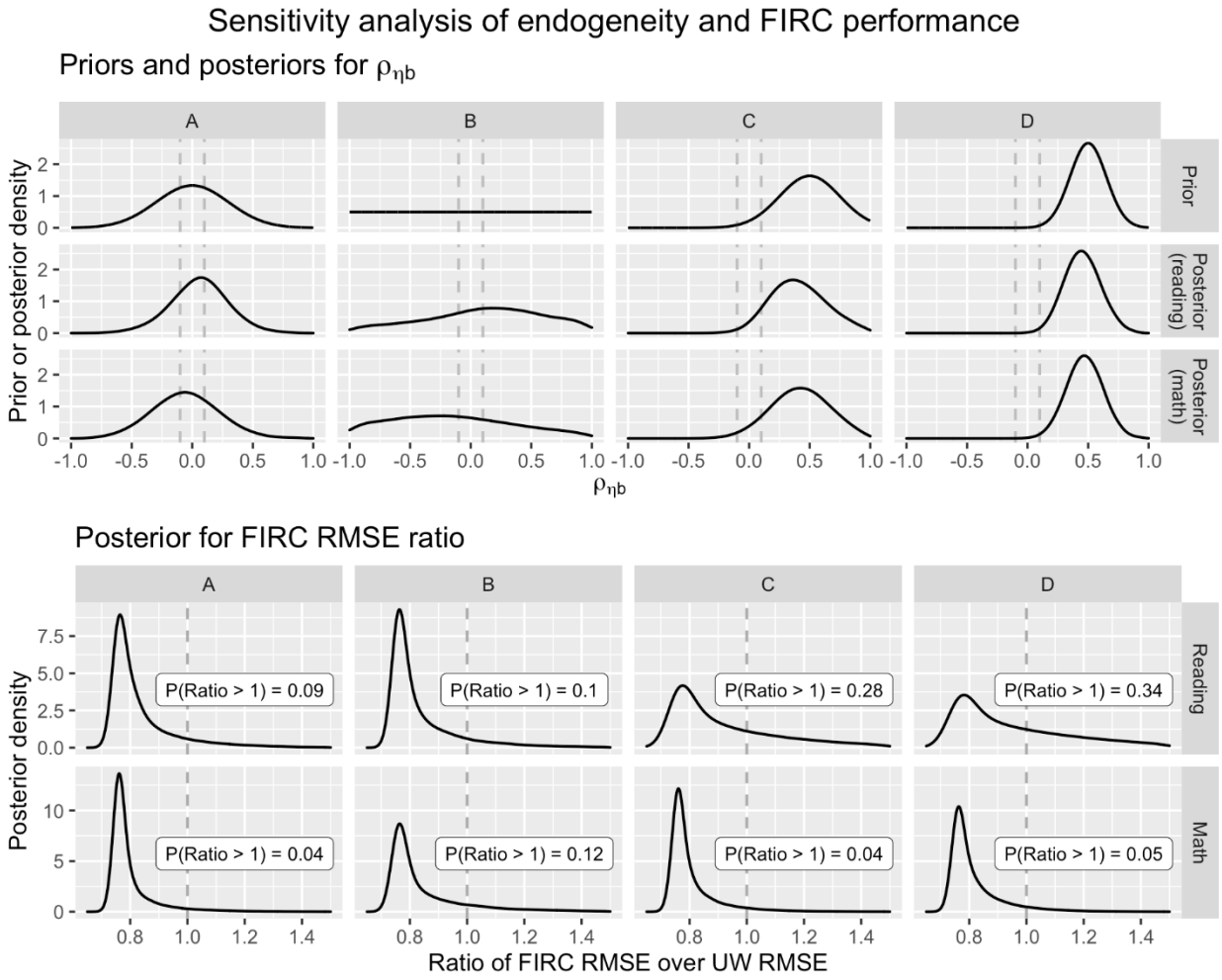

Figure 3.6.2: Bayesian inference for endogeneity and asymptotic RMSE expressions. The top grid of plots shows the four alternative prior densities for $\rho_{\eta b}$ (columns, with A corresponding to Table 3.6.2), as well as corresponding posteriors for this parameter given the reading and math data respectively. Column A corresponds exactly to (3.6.1), whereas B-D are (3.6.1) with the change to the $\rho_{\eta b}$ prior as shown in the first row. The bottom grid of plots shows the posterior densities of the ratio of the FIRC to UW RMSE approximations, which is simply a function of model parameters, under the four different priors.

evidence for very low heterogeneity of effects (see Table 3.6.2) in math limits the influence of even very large values of $\rho_{\eta b}$ (recall the expression for FIRC's asymptotic bias in Table 3.3.1). In contrast, for reading there is more evidence of cross-site treatment effect heterogeneity, so antagonistic priors that more or less rule out small $\rho_{\eta b}$ (panels C and D) suggest that FIRC's bias is more significant. That said these (probably substantively unrealistic) priors

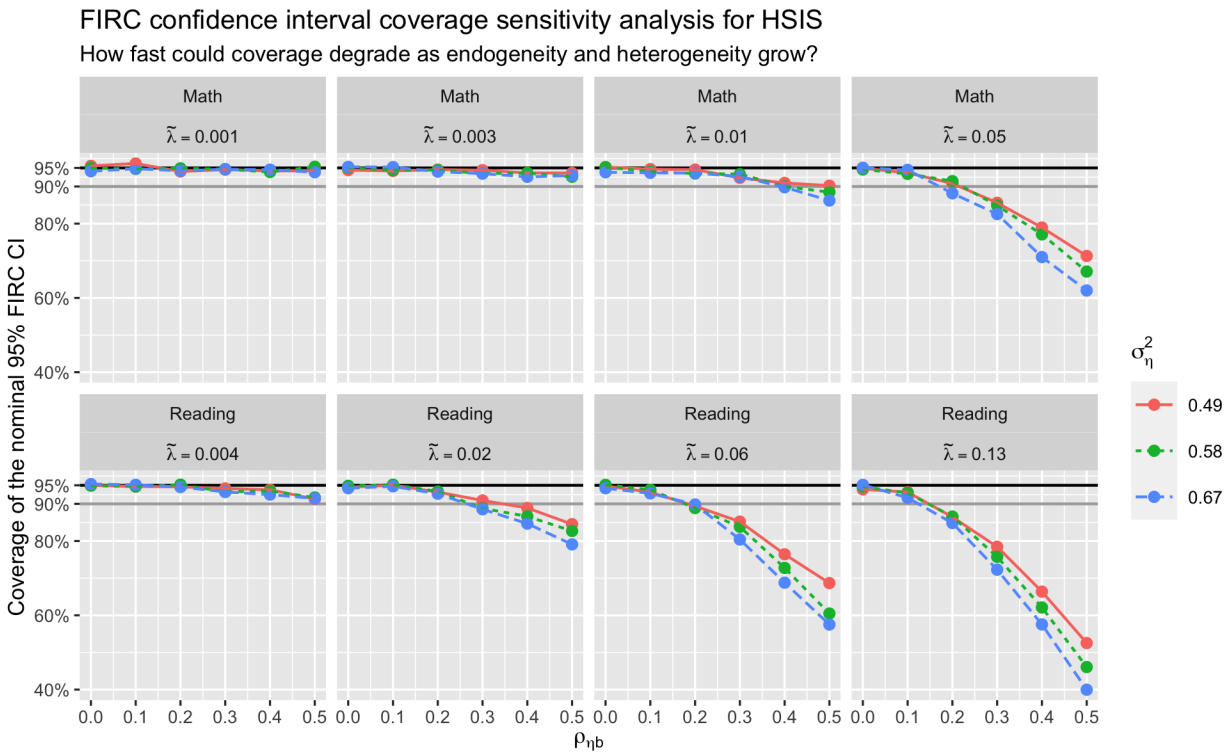

Figure 3.6.3: Simulated FIRC confidence interval coverage using parameter values plausible under the HSIS posteriors, as $\rho_{\eta b}$ (x-axis), $\tilde{\lambda}$ (columns), and σ^2_η (color and line shape) vary. Values for $\tilde{\lambda}$ correspond to the bounds for 95% and 50% highest posterior density intervals under (3.6.1) given the math (top row) and reading (bottom row) data. Values for σ^2_η correspond to the 95% highest posterior density interval and the posterior mean, which are virtually identical for both outcomes.

for $\rho_{\eta b}$ still only yield roughly 30% posterior probability that FIRC's RMSE is larger than UW's, and deep posterior concern about FIRC's RMSE would require extremely antagonistic priors for $\rho_{\eta b}$ even beyond the prior in panel D (i.e. saying that it is close to 1 with high probability). It is worth noting that these conclusions rely on the accuracy and relevance of the asymptotic expansions from Table 3.3.1, though the simulation results of Section 3.4 suggest that these are quite accurate except for fairly extreme departures from bivariate normality.

3.6.2.4 Confidence Interval Coverage

Figure 3.6.3 displays coverage rates of nominal 95% confidence intervals, evaluated at plausible values of σ_η^2 and $\tilde{\lambda}$ for the math and reading outcomes based on their posterior distributions under (3.6.1), with $\rho_{\eta b}$ ranging from 0 to 0.5 (not taken from any posterior). We focus on FIRC because it will have better coverage than FE since its bias is smaller (Theorem 2). When heterogeneity is very small ($\tilde{\lambda} < 0.005$ roughly), coverage rates for FIRC are better than 90% even for high endogeneity ($|\rho_{\eta b}| = 0.5$). However, FIRC's coverage declines below 90% as heterogeneity increases (at about $\tilde{\lambda} = 0.05$). When heterogeneity is rather higher ($\tilde{\lambda} = 0.13$, the upper limit of its 95% credible interval in reading), coverage of FIRC confidence interval deteriorates quite rapidly even for fairly moderate levels of endogeneity.

3.6.3 Conclusions and Caveats

The sensitivity analyses in this section show that (a) the unweighted estimator is very likely self-inefficient and may be extremely so, (b) FIRC is likely non-trivially more accurate as a point estimator than UW, and (c) the FIRC confidence intervals may have poor coverage for the reading outcome if endogeneity is nontrivial. For better or for worse, these analyses also indicate that the very small site sizes in HSIS mean that the data provide very little evidence about how endogenous the design may have been.

The bivariate normality assumption underlying our Bayesian inferences and the asymptotic approximations we evaluate here may be considered a limitation of these sensitivity analyses, although arguably not a challenging one to overcome. First, any nonlinearity or heteroscedasticity in the joint distribution of β_j and η_j is likely not detectable in the HSIS data given the high noise level (small $\tilde{\lambda}$). Further, because heterogeneity is small there is limited potential for these features to exist (since the magnitudes of both are constrained by the marginal variance of the site effects) and substantially affect our conclusions. However, in a study with larger heterogeneity they would be more important possibilities for analysts

to explore (e.g. through models and priors that allow and even expect them).

3.7 Implications and Open Research Areas

This paper is premised on the idea that the sample sizes in each site of a large multi-site randomized field trial often arise from a random social process. The precision with which a site's treatment effect can be estimated may then plausibly be endogenous. The initial statistical consequence is that the joint distribution of these site-specific precisions and treatment effects then becomes an important object of study. Our aim has been to shed new light on familiar estimators of the site-average treatment effect through asymptotics, simulations, and Bayesian analysis of key sensitivity parameters in an illustrative case study.

The basic conclusion is that when endogeneity of design is a possibility none of the common estimators we consider strikes a *generally* attractive balance between bias and variance. Three scale-free parameters largely control the biases and variances of these estimators: (1) $\tilde{\lambda}$, the reliability with which site-specific treatment effects vary (measuring treatment effect heterogeneity), (2) σ_η^2, the variance of the centered log precisions (measuring design imbalance), and (3) $\rho_{\eta b}$, the correlation between treatment effects and log precisions (measuring endogeneity of design).

The design-based approach, while producing unbiased point estimates and valid confidence intervals, may in some cases be self-inefficient, meaning that it makes such poor use of data that its variance and even mean squared error can be improved by discarding data. Yet a precision-based estimator (FIRC), which in many cases will plausibly produce a better point estimator, will be inconsistent under endogeneity, meaning that confidence interval coverage will inevitably deteriorate as the number of sites, J, increases. The bias-variance trade-off that arises from a forced choice among the most common estimators strikes us as intolerable. We can readily envision a class of estimators that tolerate more bias when J is small and less as J increases.

Better estimators of the site-ATE under endogeneity. For these reasons we recommend research on new estimators that would perform better when endogeneity of design is a concern. One estimation strategy that comes to mind is weight-smoothing, potentially of the "penalized spline of propensity" variety (Little 2004), where smoothing would be greater for smaller values of J, given the scale-free parameters. Another possible direction is Bayesian approaches that explicitly try to model the endogeneity of design while perhaps being careful to not take it so seriously as to discount the finite-sample efficiency gains from borrowing more heavily from the more precise sites. A key methodological challenge here is the fact that, for most multi-site trials, noise levels make the level of endogeneity hard to estimate accurately.

Other features of the site-specific ATE distribution. While our inquiry has focused on the site-average impact, any representation of the distribution of average treatment effect across sites would require estimation of the treatment effect variance and/or quantiles. Also of interest is the association between the site-specific control-group means and treatment effects, which is a measure of how much the treatment reduces or increases inequality. Here design-based versus model-based estimators may differ markedly.

Other target populations. We have focused on inferences regarding an unweighted population of sites, for example, schools, school districts, hospitals, or neighborhoods within which random assignment occurs. A population that is typically also of interest is the populations of persons nested within those sites. Miratrix, Weiss, and Henderson (2021) found that, in 12 multi-site trials they reviewed, inferences for the site-average mean were more sensitive to choice of estimator than were inferences regarding the average treatment effect defined over a population or persons. In general, however, design-based estimators of the person-average mean will be fairly efficient when sites are sampled proportional to their size. Suppose that, instead, an experimenter has opted for a constant sample size for each site despite the fact that the site-specific population sizes vary substantially. Then the

bias-variance trade-offs among model-based versus design-base estimators would be similar to those discussed in this article. Such a design might seem irrational. Experience suggests, however, that large-scale trials are intended to serve multiple purposes, including describing the distribution of effects or over persons. Therefore, optimizing design for a single use may not be desirable.

Trial Design. It seems to follow from these findings that experimental design should consider the risk of endogeneity of precision. Designs based on randomized lotteries seem particularly vulnerable, but other aspects of participant recruitment may also produce endogenous sample sizes. While beyond the scope of this article, the tools we have developed can readily be adapted to probe these design questions as they arise in application.

References

Angrist, J.D., Cohodes, S.R., Dynarski, S.M., Pathak, P.A., & Walters, C.R. (2016). Stand and deliver: Effects of Boston's charter high schools on college preparation, entry, and choice. *Journal of Labor Economics*, 34(2):275-318.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1):289-300.

Bloom, H.S., Hill, C.J., & Riccio J.A. (2003). Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments. *Journal of Policy Analysis and Management*, 22(4):551-575.

Bloom H.S., Raudenbush S.W., Weiss MJ, Porter K. (2017). Using multisite experiments to study cross-site variation in treatment effects: a hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*. 10(4):817-42

Bloom H.S., Unterman R. 2014. Can small high schools of choice improve educational prospects for disadvantaged students? *Journal of Policy Analysis and Management*. 33(2):290-319

Card, D., & Payne, A.A. (2002). School finance reform, the distribution of school spending, and the distribution of student test scores. *Journal of Public Economics*, 83(1):49-82.

Clark MA, Gleason P.M., Tuttle C.C., Silverberg M.K. (2015). Do charter schools improve student achievement? Educational. *Evaluation and Policy Analysis*. 37(4):419-36

Cochran, W.G. & Cox, G. M. (1992). *Experimental Designs*. New York: Wiley.

Efron, B. (1987) Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, 82:397, 171-185

Feller A, Grindal T, Miratrix L, Page L. (2015). Compared to what? Variation in the impact of early childhood education by alternative care-type settings. *Annals of Applied Statistics*. 10(3):1245-1285.

Finn JD, Achilles C.M. (1990). Answers and questions about class size: a statewide experiment. *American Educational Research Journal*, 27(3):557-77

Heckman, J., & Vytlacil, E. (1998). Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources*, 33(4):974-987.

Hong G, Raudenbush S.W. (2006). Evaluating kindergarten retention policy: a case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101(475):901-10

Imbens GW, Rubin D.B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge, UK: Cambridge Univ. Press. 1st ed.

Kautz, T., Schochet, P.Z., Tilley, C. (2017). *Comparing Impact Findings from Design-Based and Model-Based Methods: An Empirical Investigation.* US Institute of Educational Sciences.

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. Journal of the American Statistical Association, 99(466):546-556.

Meng, X.L. and X. Xie. (2014). I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? *Econometric Reviews* 33(1-4): 218-250.

Miratrix, L.W., Weiss, M.J., and Henderson, B. (2021) An applied researcher's guide to estimating effects from multisite individually randomized trials: Estimands, estimators, and estimates. *Journal of Research on Educational Effectiveness*, 14(1):270-308.

Mosteller F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children.* 5(2):113

Nieuwlaat, R., Wilczynski, N., Navarro, T., Hobson, N., Jeffery, R., Keepanasseril, A., Agoritsas, T., Mistry, N., Iorio A., Jack, S., Sivaramalingam, B., Iserman, E., Mustafa, R.A., Jeraszewski, D, Cotoi, C., and Haynes, R.B. (2014). Interventions for enhancing medication adherence. *Cochrane Database of Systematic Reviews*, 11.

Nye, B., Konstantopoulos, S. and Hedges, L.V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26(3):237-257.

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). *Head Start Impact Study Final Report.* Washington, DC: Prepared for the Office of Planning, Research and Evaluation of the Administration for Children and Families of the U.S. Department of Health and Human Services.

Raudenbush, S.W. & Bloom, H.S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, 36(4):475-499

Raudenbush, S.W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational statistics*, 10(2):75-98.

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edition. Thousand Oaks, CA: Sage Publications.

Raudenbush, S.W. & Schwartz, D. (2020). Randomized Experiments in Education, with Implications for Multilevel Causal Inference. *Annual Review of Statistics and Its Applications*, 7(1):117-208.

Reardon, S.F., Yun, J. T., & Eitle, T. M. (2000). The changing structure of school segregation: Measurement and evidence of multiracial metropolitan-area school segregation, 1989–1995. *Demography*, 37(3):351-364.

Rubin, D.B. (1986). Comment: Which ifs have causal answers? *Journal of the American Statistical Association*, 81(396):961-962.

Schochet, P.Z. (2016). *Statistical theory for the RCT-YES Software: Design-based Causal Inference for RCTs*, Second Edition (NCEE 2015–4011). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development.

Spybrook, J. (2013). Detecting Intervention Effects Across Context: An Examination of the Precision of Cluster Randomized Trials. *The Journal of Experimental Education*, 82(3):334-357.

U.S. Department of Health and Human Services (2019). Head Start Program Facts: Fiscal Year 2018. https://eclkc.ohs.acf.hhs.gov/about-us/article/head-start-program-facts-fiscal-year-2018.

Vinovskis, M. A. (2005). *The Birth of Head Start.* University of Chicago Press

Walters C.R. (2015). Inputs in the production of early childhood human capital: evidence from head start. *American Economic Journal, Applied Economics.* 7(4):76-102

Weiss M.J., Bloom H.S., Verbitsky-Savitz N., Gupta H., Vigil A.E., Cullinan D.N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness.* 10(4):843-876

Yu, C. and Hoff, P. D. (2018) Adaptive multigroup confidence intervals with constant coverage. *Biometrika*, 105(2):319-335.

APPENDICES

A Appendix to "Dynamic Borrowing From Historical Controls Via the Synthetic Prior with Covariates in Randomized Clinical Trials"

A.1 Data generation details for the simulation study

A.1.1 Factors defining the different simulation settings

In the simulation study data are generated according in a wide array of schemes, which consists of a (nearly) full factorial design with the following factors.

Direct relevance of historical rates and usefulness of covariates. The most important conditions we vary across simulations are (a) whether the historical control rates are directly relevant (similar) to the new rate or not and (b) whether the covariates predict response rates or not. These conditions define what we call Scenarios 1-4:

- Scenario 1: Historical rates similar to new rate *and* covariates predictive
- Scenario 2: Historical rates not similar but covariates predictive
- Scenario 3: Historical rates similar and covariates not predictive
- Scenario 4: Historical rates not similar and covariates not predictive

For further details see the beginning of Section A.1.2 and the attached code/simulated historical data files.

Maximum target control group size. The maximum target control group size may be either 150 or 80 patients, reflecting some of the range of sample sizes typical in non-oncology Phase 2 efficacy trials (see, e.g., Table 2.5.1). In the designs we consider 150 or 80 are also the *fixed* sample sizes of the treatment group, with equal randomization between treatment and control in Stage 1 and unbalanced randomization in Stage 2 (depending on the # of Stage 2 control patients). This leads to total trial sizes on the order of 300 or 160, more or less depending on the concordance between the historical and new control data at interim.

Whether all covariates are used. The data-generating schemes for the main simulation results in Tables 1 and 3 are given by a full factorial design defined on the *previous* factors, in a setting where the analyst only observes and uses 2 of the 6 covariates actually associated with response rates. Finally, just for Scenarios 1 and 2 (where covariates are useful) and the $n^c_{max} = 150$ maximum trial size case, we also simulate cases in which the analysts observe and may use all 6 covariates, yielding Tables 2 and 4.

A.1.2 Exact sampling procedure

We directly generate trial-level data without sampling any patient-level data. To generate random trial-level covariates and true response rates that are correlated, we first sample from a marginal distribution for the covariates and then from a conditional distribution for response rates given covariates. In Scenarios 1 and 3, the historical trials come from groups $i(h) \in \{1, 2, 3, 4, 5, 6\}$ (with 8 from each, for a total of 48 historical trials) and the new trial comes from group $i(H + 1) = 6$. In Scenarios 2 and 4, the historical trials come from groups $i(h) \in \{1, 2, 3, 4, 6\}$ (missing group 5; with 10 from each, for a total of 50 historical trials) while the new trial comes from group $i(H + 1) = 5$, leading to a lower response rate than any of the historical trials.

To sample a single trial's control data:

1. Sample covariates.

The trial-level covariates include both patient characteristics, which are randomly sampled, and trial eligibility or treatment regimen features, which are fixed under repeated sampling for a given trial (but vary deterministically across trials). To reflect a setting where different trials (indexed by h) are conducted in different settings, we pose 6 different covariate distributions and each trial's covariates are sampled from just one of these 6 (indexed by the function $i(h)$).

- Proportion male:

$$\sim Binom(n, \mu_{i(h)})/n$$

  where $\mu_{i(h)}$ is the $i(h)$-th component of the vector $(0.3, 0.2, 0.28, 0.34, 0.42, 0.35)$.

- Average weight at trial start (kg):

$$\sim N(\mu_{i(h)}, 9)$$

  where $\mu_{i(h)}$ is the $i(h)$-th component of the vector $(75, 55, 60, 64, 82, 66)$.

- Average disease duration at trial start (years):

$$\sim \mu_{i(h)} + \begin{cases} -2 & w.p.\ 0.15 \\ -1 & w.p.\ 0.2 \\ 0 & w.p.\ 0.3 \\ 1 & w.p.\ 0.2 \\ 2 & w.p.\ 0.15 \end{cases}$$

  where $\mu_{i(h)}$ is the $i(h)$-th component of the vector $(18, 16, 8, 18, 22, 12)$.

- Dosing scheme:

$$= d_{i(h)}$$

where $d_{i(h)}$ is the $i(h)$-th component of the vector

(biweekly, twice daily, weekly, weekly, biweekly, weekly).

In regressions this is represented as two binary covariates (biweekly or not, and twice daily or not).

- Previous/background treatment at trial start:

$$= b_{i(h)}$$

where $b_{i(h)}$ is the $i(h)$-th component of the vector

(none, none, MTX, MTX, none, none).

2. Sample response rate given covariates.

$$y_h|\psi_h \overset{ind.}{\sim} Binom(n_h, \psi_h)$$
$$\psi_h, x_h \overset{ind.}{\sim} Beta(\alpha_h, \beta_h)$$

α_h and β_h are such that the Beta distribution has a mean given by the regression on covariates

$$\begin{aligned} logit\left(\frac{\alpha_h}{\alpha_h+\beta_h}\right) = {} & 10\log(0.85)\cdot \text{male}_h + 0.1\log(0.95)\cdot \text{weight}_h + \\ & 0.2\log(0.95)\cdot \text{duration}_h + \log(0.9)\cdot I(d_{i(h)} = \text{biweekly}) + \\ & \log(1.1) I(d_{i(h)} = \text{twice daily}) \end{aligned}$$

and scaled so that all 6 covariates explain 80% of the variation in true response rates (i.e. considering the law of total of total variance, we have conditional variance given covariates = 0.2 * marginal variance). This scaling is determined by Monte Carlo integration to find the marginal variance of the response rates.

Sampling both ψ_h and y_h means there is randomness in the true response rates beyond just randomness in the observed response rate given the truth. This can be compared to a Frequentist repeated sampling regime in which when a trial is replicated not only would the same exact patients have different observed responses, but also the sample of patients would change (leading to a different "true" response rate).

These covariate distributions and regression equation were calibrated to mimic the range of covariate values and response rates we observed in an expanded version of Table 2.5.1.

To sample the new trial's treatment group data we take

$$y_{trt}|\psi_{trt} \sim Bin(n_{trt}, \psi_{trt})$$

where n_{trt} is the control group's maximum target sample size (half for interim analysis) and ψ_{trt} is 0 or 0.3 greater than the new trial's control rate of ψ_{H+1} (reflecting no treatment effect or an additive treatment effect of 30 percentage points).

For the Frequentist repeated sampling in the simulation study, we sampled the historical trials' data *just once* for each of Scenarios 1-4 and fixed it under repeated sampling, only resampling the new trial's data on each replicate. This reflects a setting where the historical trials have already been run, and the Frequentist replication being considered is just replication of the yet-to-be conducted new trial. The fixed historical data set that was generated for each Scenario is shared along with the code.

A.2 Simulation results for variations on the SPx prior

In Table A.1 we give simulation results from Table 2.4.1, with the addition of those for two minor variations on the original SPx prior.

"SPx C." stands for "SPx Careful", and mimics the SPx prior stated in Section 2.2 but with

$$(p_{hist}, p_{reg}, p_{ind}) = \left(\frac{3}{40}, \frac{3}{40}, \frac{34}{40}\right)$$

instead of $(p_{hist}, p_{reg}, p_{ind}) = \left(\frac{1}{8}, \frac{1}{8}, \frac{3}{4}\right)$, giving the no-borrowing submodel 85% prior weight instead of 75%.

"SPx D." stands for "SPx Diffuse", and it puts equal prior weights on the submodels while making the priors within the borrowing submodels more diffuse (i.e. pushing them to borrow *less* strongly). In particular, it changes the SPx prior stated in Section 2.2 in three ways:

1. for the historical borrowing submodel, the prior variance σ gets the less concentrated hyperior

   $$\sigma \sim TCauchy(0, 0.1, (0, \infty)),$$

   whereas in the regular SPx method the scale parameter is 0.02,

2. for the regression submodel, the prior variance is not scaled down, so we just have

   $$\theta_{reg} | \boldsymbol{\beta}, \tau^2, \mathbf{x}_{H+1} \sim N(\boldsymbol{\beta}^T \mathbf{x}_{H+1}, \tau^2),$$

   whereas in the regular SPx the prior variance here is $c\tau^2$ with $c = \frac{1}{25}$, and

3. the prior submodels weights are

   $$(p_{hist}, p_{reg}, p_{ind}) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$$

instead of $(p_{hist}, p_{reg}, p_{ind}) = \left(\frac{1}{8}, \frac{1}{8}, \frac{3}{4}\right)$ in the regular SPx.

From the SPx C. columns in Table A.1 we can see that putting slightly more prior weight on the no-borrowing submodel may be an attractive alternative, at least in this formulation where the borrowing submodels have concentrated priors and are well separated from the no-borrowing submodel. SPx C. achieves more conservative coverage (and slightly better RMSE) than SPx in Scenarios 2 and 4, where SPx can suffer slightly. At the same time its RMSE in Scenarios 1 and 3, where borrowing is "easy," is only slightly worse than that of SPx, though its crfedible intervals are somewhat wider.

In contrast, the SPx D. results are poorer and vindicate the regular SPx strategy of making the borrowing submodels have tight priors while putting most prior mass on the no-borrowing submodel. SPx D. ends up borrowing more aggressively and confidently, leading to poor coverage (and in some cases pitiful RMSE) in Scenarios 2 and 4.

A.3 Rationale for the SPx hyperprior

Near the end of Section 2.2 in the text, we highlighted two prior choices in SPx that drive its adaptive historical borrowing:

(a) make the priors in the *hist* and *reg* submodels relatively strongly concentrated (i.e. high prior probability of small σ^2 and τ^2), and

(b) give relatively high prior weight to the *ind* (no-borrowing) submodel (i.e. $p_{ind} > p_{hist}, p_{reg}$).

A more detailed explanation of these choices follows.

By posing quite informative priors for the *hist* and *reg* borrowing submodels, (a), we both allow strong borrowing when deemed appropriate (and $p^*_{hist} + p^*_{reg}$ is large) and make the marginal likelihoods of these submodels more sensitive to conflict between the new and historical data. This means that their marginal likelihoods will more quickly dwarf the no-borrowing submodel's marginal likelihood as borrowing grows safer and shrink compared to it

	$\mathbf{n^c_{max} = 150}$							$\mathbf{n^c_{max} = 80}$						
	SPx		SPx C.		SPx D.		Ind.	SPx		SPx C.		SPx D.		Ind.
	Fixed	Adapt.	Fixed	Adapt.	Fixed	Adapt.	Fixed	Fixed	Adapt.	Fixed	Adapt.	Fixed	Adapt.	Fixed
Scenario 1														
Size	150	123.4	150	127.0	150	116.7	150	80	64.9	80	66.7	80	60.0	80
RMSE	0.025	0.026	0.026	0.027	0.025	0.026	0.032	0.024	0.026	0.024	0.027	0.024	0.026	0.032
Coverage	94.2	96.0	96.0	96.5	93.0	94.2	96.7	99.0	98.7	99.5	99.2	98.1	97.7	99.6
Width	0.104	0.113	0.115	0.124	0.090	0.098	0.140	0.135	0.153	0.153	0.170	0.107	0.119	0.190
Scenario 2														
Size	150	127.7	150	133.6	150	119.2	150	80	65.0	80	68.14	80	61.3	80
RMSE	0.031	0.034	0.031	0.033	0.032	0.034	0.030	0.033	0.038	0.032	0.036	0.034	0.038	0.030
Coverage	91.4	92.0	94.1	93.6	87.7	87.1	97.1	97.2	95.6	98.0	97.8	90.5	90.3	99.6
Width	0.114	0.123	0.118	0.125	0.100	0.112	0.125	0.153	0.165	0.160	0.171	0.131	0.144	0.171
Scenario 3														
Size	150	122.7	150	126.0	150	116.6	150	80	64.1	80	65.7	80	60.9	80
RMSE	0.022	0.023	0.024	0.025	0.022	0.023	0.033	0.021	0.023	0.022	0.025	0.019	0.023	0.032
Coverage	96.6	96.8	97.2	97.4	98.5	98.8	97.3	99.8	99.6	99.9	99.9	100.0	100.0	99.7
Width	0.106	0.116	0.116	0.127	0.111	0.123	0.140	0.138	0.157	0.155	0.173	0.139	0.155	0.190
Scenario 4														
Size	150	156.4	150	156.9	150	146.7	150	80	78.6	80	79.9	80	71.1	80
RMSE	0.040	0.043	0.037	0.039	0.045	0.049	0.030	0.046	0.052	0.040	0.046	0.056	0.065	0.029
Coverage	93.5	92.1	95.3	93.7	87.3	86.7	96.4	98.1	96.0	99.6	97.4	91.4	87.5	99.7
Width	0.140	0.135	0.136	0.132	0.140	0.140	0.125	0.189	0.187	0.186	0.185	0.183	0.184	0.170

Table A.1: Control group size and Frequentist estimation accuracy for the new trial's control response rate, averaged over 1,000 simulated trials. Metrics are defined as in Table 2.4.1.

when borrowing grows more dangerous. The rationale for *a priori* favoring the no-borrowing submodel, (b), is to compensate for the behavior of the three submodels' marginal likelihoods in the challenging setting of moderate but not necessarily obvious conflict between the new and historical data. In this case, especially when the new trial size is small (e.g. at interim), the marginal likelihood of the no-borrowing submodel may not be much larger than that of the borrowing submodels because its prior is more diffuse and thus still gives nontrivial prior mass to response rates that have low likelihoods, so to get large p^*_{ind} in this case we must have large p_{ind}. This also reflects the idea that we should be conservative about historical borrowing, especially when sample sizes are small.

B Appendix to "Treatment Effect Estimation in Multi-site Trials with Endogenous Design: Old Estimators, New Results"

B.1 Computational Derivations and Formulae

B.1.1 FIRC Estimator

The maximum likelihood estimators $\hat{\sigma}^2_{bFIRC}$and $\hat{\beta}_{FIRC}$ are the solutions to the pair of estimating equations

$$\hat{\sigma}^2_{bFIRC} = \min\left\{\left(\frac{\sum \hat{\Delta}^{-2}_{FIRCj}[(\hat{\beta}_j - \hat{\beta}_{FIRC})^2 - V_j]}{\sum \hat{\Delta}^{-2}_{FIRCj}}\right), 0\right\} \tag{B.1}$$

$$\hat{\beta}_{FIRC} = \frac{\sum \hat{\Delta}^{-1}_{FIRCj}\hat{\beta}_j}{\sum \hat{\Delta}^{-1}_{FIRCj}} \tag{B.2}$$

where $\hat{\Delta}_{FIRCj} = \hat{\sigma}^2_{bFIRC} + V_j$. For $\hat{\sigma}^2_{bFIRC} = 0$, $\hat{\beta}_{FIRC} = \hat{\beta}_{FE}$. For $\hat{\sigma}^2_{bFIRC} > 0$, we find it convenient to replace the weight $\hat{\Delta}^{-1}_{FIRCj}$ with the equivalent weight $\hat{\lambda}_{FIRCj} = \hat{\sigma}^{FIRC^2}_b \hat{\Delta}^{-1}_{FIRCj}$, a number lying in the interval $(0, 1)$ that helps us evaluate the behavior of the FIRC without reference to the scale of the outcome.

We can solve this system of equations iteratively based on the method of Fisher scoring. Given values $(\hat{\sigma}^{(m)^2}_{bFIRC}, \hat{\beta}^{(m)}_{FIRC})$ at iteration m, compute weight $\hat{\Delta}^{(m)}_{FIRCj}$ and squared residuals$(\hat{\beta}_j - \hat{\beta}^{(m)}_{FIRC})^2$; substitute into the RHS of (B.1) and (B.2) and solve to obtain $(\hat{\sigma}^{(m+1)^2}_{bFIRC}, \hat{\beta}^{(m+1)}_{FIRC})$. Repeat until convergence.

Proof. The density $f(\hat{\beta}_j|V_j) = (2\pi\Delta_j)^{-\frac{1}{2}} e^{-\frac{1}{2}\Delta_j^{-1}(\hat{\beta}_j - \beta)^2}$; the log-likelihood for the sample

is

$$L(\beta, \sigma_b^2 | data) = constant - \frac{1}{2}\sum \ln(\Delta_j) - \frac{1}{2}\sum \Delta_j^{-1}(\hat{\beta}_j - \beta)^2$$

The score functions are

$$\begin{aligned} S(\sigma_b^2) &= -\tfrac{1}{2}\sum \Delta_j^{-1} + \tfrac{1}{2}\sum \Delta_j^{-2}(\hat{\beta}_j - \beta)^2 \\ &= \tfrac{1}{2}\sum \Delta_j^{-2}\left[(\hat{\beta}_j - \beta)^2 - \Delta_j\right] \\ &= \tfrac{1}{2}\sum \Delta_j^{-2}\left[(\hat{\beta}_j - \beta)^2 - V_j\right] - \tfrac{1}{2}\sum \Delta_j^{-2}\sigma_b^2; \\ S(\beta) &= \sum \Delta_j^{-1}(\hat{\beta}_j - \beta). \end{aligned}$$

Setting $S(\sigma_b^2)$=$S(\beta) = 0$ and solving for $\sigma_b^2$and β yields (B.1) and (B.2). □

B.1.2 FIRC+ Estimator

FIRC+ uses iteratively re-weighted least squares applied to the regression model

$$\hat{\beta}_j = \beta + \alpha\eta_j + \varepsilon_j + e_j. \tag{B.3}$$

Here $\varepsilon_j = b_j - \alpha\eta_j \sim N(0, \sigma_{b|\eta}^2)$, $e_j = \hat{\beta}_j - \beta_j \sim N(0, V_j)$, $\varepsilon_j \perp \eta_j$. To maximize the likelihood with respect to$\sigma_{b|\eta}^2, \alpha$ and β, we can solve the estimating equations iteratively:

$$\hat{\sigma}_{b|\eta}^{(m+1)2} = \max\left\{\left(\frac{\sum \Delta_{+j}^{(m)-2}[(\hat{\beta}_j - \hat{\beta}_{FIRC+}^{(m)} - \hat{\alpha}^{(m)}\eta_j)^2 - V_j]}{\sum \Delta_{+j}^{(m)-2}}\right), 0\right\} \tag{B.4}$$

$$\hat{\alpha}^{(m+1)} = \frac{\sum \Delta_{+j}^{(m)-1}\eta_j(\hat{\beta}_j - \bar{\beta})]}{\sum \Delta_{+j}^{(m)-1}\eta_j^2} \tag{B.5}$$

$$\hat{\beta}_{FIRC+}^{(m+1)} = \bar{\beta}^{(m)} - \hat{\alpha}^{(m)}\bar{\eta}. \tag{B.6}$$

Here $\Delta_{+j} = \sigma_{b|\eta}^2 + V_j$, $\bar{\beta} = \sum \Delta_{+j}^{-1}\beta_j / \sum \Delta_{+j}^{-1}$ and $\bar{\eta} = \Delta_{+j}^{-1}\eta_j / \sum \Delta_{+j}^{-1}$. For $\hat{\sigma}_{b|\eta}^2 > 0$,

we find it convenient to substitute $\lambda_{+j} = \sigma^2_{b|\eta}\Delta^{-1}_{+j}$ for Δ^{-1}_{+j} in (B.4)-(B.6). Given values $\hat{\sigma}^{(m)2}_{b|\eta}$ and $\hat{\alpha}^{(m)}$at iteration m, compute weight $\Delta^{(m)-1}_{+j}$ and solve the LHS for estimates at iteration m+1.

Proof. Model (B.3) can be written

$$\begin{aligned}\hat{\beta}_j &= \beta_0 + \alpha(\eta_j - \bar{\eta}) + \varepsilon_j + e_j \\ &= \begin{pmatrix} 1 & \eta_j - \bar{\eta} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \alpha \end{pmatrix} + \varepsilon_j + e_j =: x_j^T\theta + \varepsilon_j + e_j\end{aligned}$$

where $\beta_0 = \beta + \alpha\bar{\eta}$, $x_j^T = \begin{pmatrix} 1 & \eta_j - \bar{\eta} \end{pmatrix}$ and $\theta = \begin{pmatrix} \beta_0 & \alpha \end{pmatrix}^T$. The density function for the data is thus

$$h(\hat{\beta}_j|\eta_j) = (2\pi\Delta_{+j})^{-\frac{1}{2}} e^{-\frac{1}{2}\Delta^{-1}_{+j}(\hat{\beta}_j - x_j^T\theta)^2},$$

and the score functions are

$$\begin{aligned}S(\sigma^2_{b|\eta}) &= \frac{-1}{2}\sum_{j=1}^{J}\Delta_j^{-1} + \frac{1}{2}\sum_{j=1}^{J}\Delta_j^{-2}(\hat{\beta}_j - x_j^T\theta)^2 \\ &= \frac{1}{2}\sum_{j=1}^{J}\Delta_j^{-2}\left[(\hat{\beta}_j - x_j^T\theta)^2 - \delta_j\right] \\ &= \frac{1}{2}\sum_{j=1}^{J}\Delta_j^{-2}\left[(\hat{\beta}_j - x_j^T\theta)^2 - V_j\right] - \frac{1}{2}\sum_{j=1}^{J}\Delta_j^{-2}\sigma^2_{b|\eta}\end{aligned}$$

and

$$\begin{aligned}S(\theta) &= \sum \Delta_j^{-1}x_j(\hat{\beta}_j - x_j^T\theta \\ &= \sum \Delta_j^{-1}x_j\hat{\beta}_j - \sum \Delta_j^{-1}x_jx_j^T\theta.\end{aligned}$$

Setting $S(\sigma^2_{b|\eta}) = 0$ and solving for $\sigma^2_{b|\eta}$ yields (B.4). Setting $S(\theta) = 0$ and solving for θ yields

$$\begin{aligned}\hat{\theta} &= \left(\sum \Delta_j^{-1} x_j x_j^T\right)^{-1} \sum \Delta_j^{-1} x_j \hat{\beta}_j \\ &= \begin{pmatrix} \hat{\beta}_0 \\ \hat{\alpha} \end{pmatrix} = \begin{pmatrix} \sum \Delta_j^{-1} & 0 \\ 0 & \sum \Delta_j^{-1}(\eta - \bar{\eta})^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum \Delta_j^{-1} \hat{\beta}_j \\ \sum \Delta_j^{-1}(\eta_j - \bar{\eta})\hat{\beta}_j \end{pmatrix}.\end{aligned}$$

Note that $\sum \Delta_j^{-1}(\eta_j - \bar{\eta})\hat{\beta}_j = \sum \Delta_j^{-1}\eta_j(\hat{\beta}_j - \bar{\beta})$. This then gives (B.5) and (B.6). □

B.1.3 Finding V_j

Define Y_{ij} as the outcome for person i in site j. The sample means for treated and control units are $\bar{Y}_{1j} = \sum_{i=1}^{n_j} T_{ij}Y_{ij} / \sum_{i=1}^{n_j} T_{ij}$and $\bar{Y}_{0j} = \sum_{i=1}^{n_j}(1 - T_{ij})Y_{ij} / \sum_{i=1}^{n_j}(1 - T_{ij})$. We regard the site-specific treatment effect $\hat{\beta}_j = \bar{Y}_{1j} - \bar{Y}_{0j}$ as an unbiased estimator of $\beta_j = \mu_{1j} - \mu_{0j}$having sampling variance

$$V_j := Var(\hat{\beta}_j | \beta_j) = \sum_{j=1}^{n_j} T_{ij}\sigma^2_{ij} / \sum_{j=1}^{n_j} T_{ij} + \sum_{j=1}^{n_j}(1 - T_{ij})\sigma^2_{ij} / \sum_{j=1}^{n_j}(1 - T_{ij}).$$

Here μ_{1j} is the mean outcome in site j if all persons in that site were treated and μ_{0j} is the mean outcome if all such members were untreated, and $\sigma^2_{ij} = Var(Y_{ij} | \mu_{1j}, \mu_{0j})$, which we shall call the "within-site variance" for person i in site j.

Clearly some structure must be imposed upon the within-site variance if we are to estimate V_j. Many analysts have assumed a constant variance $\sigma^2_{ij} = \sigma^2$ for all i and j, in which case $V_j = \sigma^2/[n_j\bar{T}_j(1 - \bar{T}_j)]$. This is the assumption underlying $\hat{\beta}_{FE}$. However, Bloom et al. (2017) show why this assumption is unreasonable when treatment effects vary across sites and show that a reasonably robust alternative is to specify different constants for treated and control units, that is $\sigma^2_{ij} \equiv T_{ij}\sigma^2_1 + (1 - T_{ij})\sigma^2_0$, in which case $V_j = \sigma^2_1/(n_j\bar{T}_j) +$

$\sigma_1^2[n_j(1-\bar{T}_j)]$. In either case the constants σ^2 or σ_1^2, σ_0^2 can be estimated consistently by pooling sums of squares within sites or by estimating a two-level hierarchical model. For the purpose of analyzing the Head Start data in this paper, we estimated the model $Y_{ij} = \beta_0 + b_{0j} + (\beta_1 + b_{1j})T_{ij} + e_{ij}$via maximum likelihood where $b_{0j}$and $b_{1j}$are zero-mean random effects having variances $\sigma^2_{b_0}$and $\sigma^2_{b_1}$respectively and covariance $\sigma_{b_0 b_1}$. We found that the specification of a single level-1 variance $Var(e_{ij}) = \sigma^2$ was suitable in this case.

In the trials that we have seen, estimates of σ^2 or σ_1^2, σ_0^2 are quite precise. We caution against specifying unique variances for each site (e.g., $V_j = \sigma^2_{1j}/(n_j\bar{T}_j) + \sigma^2_{1j}[n_j(1-\bar{T}_j)]$) unless site sizes are very large, although one could obtain precise estimates of $\sigma^2_{1j}, \sigma^2_{1j}$ by modeling them as a function of covariates or by specifying an exchangeable prior (Kasim & Raudenbush, 1998).

References

Bloom, H.S., Hill, C.J., & Riccio J.A. (2003). Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments. *Journal of Policy Analysis and Management*, 22(4), 551-575.

Kasim, R. M., and Raudenbush, S. W. (1998). Application of Gibbs Sampling to Nested Variance Components Models with Heterogeneous Within-Group Variance. *Journal of Educational and Behavioral Statistics*, 23(2), 93-116.

B.2 Some Finite Sample Bias Results

This section proves a general characterization of the finite sample bias of a weighted estimator of the site-ATE, as well as the result showing that in finite samples the FIRC estimator is never more biased than the FE estimator for the site-ATE.

B.2.1 Bias of weighted estimators

Lemma B.1. *If an estimator* $\hat{\beta}$ *of* β *can be written* $\hat{\beta} = \sum_{j=1}^{J} w_j \hat{\beta}_j / \sum_{j=1}^{J} w_j$ *where* $\sum_{j=1}^{J} w_j \neq 0$ *and* $Cov\left(\frac{w_j}{\bar{w}}, \hat{\beta}_j - \beta_j\right) = 0$*, then under model* (3.2.1)*,* $Bias(\hat{\beta}, \beta) = Cov(w_j/\bar{w}, \beta_j)$ *, with* $\bar{w} := J^{-1} \sum_{j=1}^{J} w_j$*.*

Proof. We can immediately see that the bias is

$$J^{-1} \sum_{j=1}^{J} E\left(\frac{w_j}{\bar{w}} \hat{\beta}_j\right) - \beta = J^{-1} \sum_{j=1}^{J} \left[Cov\left(\frac{w_j}{\bar{w}}, \hat{\beta}_j\right) + \beta\right] - \beta = Cov\left(\frac{w_j}{\bar{w}}, \beta_j\right)$$

The first equality follows because $E\left(\frac{w_j}{\bar{w}}\right) = 1$ (by exchangeability) and the second follows by exchangeability and the assumption that the normalized weights have zero covariance with the sampling errors. □

In English, this lemma says that the bias for the site-ATE is simply the covariance between the estimator's (normalized) weights and the site-specific ATEs.

B.2.2 FIRC is never more biased than FE

Theorem 3.2. *For any number of sites* J*,* $\left|Bias(\hat{\beta}_{FIRC}, \beta)\right| \leq \left|Bias(\hat{\beta}_{FE}, \beta)\right|$*.*

Proof. The basic argument here is that the FIRC estimator is bounded between the FE estimator and the UW estimator, and FE is biased while UW is not.

To see this, note that the normalized FIRC weight for site j, as a function of the

estimated variance $\hat{\sigma}_b^2$, can be written

$$h_j(\hat{\sigma}_b^2) := \frac{\frac{1}{\hat{\sigma}_b^2+V_j}}{J^{-1}\sum_{j=1}^{J}\frac{1}{\hat{\sigma}_b^2+V_j}}$$

It is easy to see that $h_j(0) = P_j/\bar{P}$, the normalized FE weight, and that $\lim_{\hat{\sigma}_b^2\to\infty} h(\hat{\sigma}_b^2) = 1/J$, the UW weight. Furthermore, note that h_j is monotone in $\hat{\sigma}_b^2$ (decreasing if the j-th weight is above average, and increasing if it is below average), which can be checked for instance through the derivative (or by noting that it is continuous and open). The monotonicity means that in fact the (normalized) FIRC weights are each bounded between the respective (normalized) FE precision weight and the UW constant weight. And because the estimators are weighted sums, this implies that the FIRC estimator is bounded between the FE and UW estimators, so its expectation (and bias) is also bounded between theirs. Since UW is always unbiased whereas FE may not be, FIRC is never more biased than FE. □

B.3 Approximations of the Bias and Variance of Estimators

This section derives the bias and variance results shown in Table 3.3.1 of the text, reproduced here for convenience.

First we describe the basic derivation strategy that we apply to each of the estimators, which is fairly straightforward. We want to find the asymptotic distributions of estimators that each have the form $\hat{\beta} = \sum_{j=1}^{J} w_j \hat{\beta}_j / \sum_{j=1}^{J} w_j$. We consider the estimator $\hat{\beta}$ in a sequence of estimators as $J \to \infty$, so more formally we could write $\hat{\beta}^{(J)}$ to emphasize the dependence on J, but we avoid this notation except where necessary for clarification. Our derivations are based on the representation

$$J^{1/2}(\hat{\beta} - \beta^*) = \frac{J^{1/2}\left(\sum_{j=1}^{J} w_j(\hat{\beta}_j - \beta^*)\right)}{\sum_{j=1}^{J} w_j} = \frac{J^{1/2} \cdot \frac{1}{J}\sum_{j=1}^{J} Z_j}{\frac{1}{J}\sum_{j=1}^{J} w_j} = \frac{J^{1/2}\bar{Z}}{\bar{w}}, \tag{B.7}$$

where $\beta^* := \text{plim}(\hat{\beta})$ (which exists by direct application of the WLLN to $\hat{\beta}$) and $Z_j := w_j(\hat{\beta}_j - \beta^*)$. Our strategy for analyzing (B.7) is to deal with the numerator and denominator separately by applying the Central Limit Theorem in the numerator, the Weak Law of Large Numbers in the denominator, and then Slutsky's Theorem to combine the results. This strategy assumes that $w_j\hat{\beta}_j$ and w_j are both *iid* sequences asymptotically as $J \to \infty$, so it can be applied immediately to UW, FE, and FIRC+ but only after slight modification to FIRC, which has an estimated weight (as a function of $\hat{\sigma}_b^2$) with a mean that changes as J changes.

In particular, these *iid* assumptions and the construction of Z_j implies that Z_j is itself *iid* and that it has mean 0 because

$$\beta^* = \text{plim}\left(\frac{J^{-1}\sum_{j=1}^{J} w_j\hat{\beta}_j}{J^{-1}\sum_{j=1}^{J} w_j}\right) = \frac{E(w_j\hat{\beta}_j)}{E(w_j)} \Rightarrow E(Z_j) = E(w_j\hat{\beta}_j) - E(w_j)\beta^* = 0,$$

where the second equality comes from Slutsky and the WLLN assuming that the summands

	Estimator	Asymptotic Bias	Asymptotic Variance
UW	$\frac{1}{J}\sum_{j=1}^{J}\hat{\beta}_j$	0	$\sigma_b^2 + \mu_{\tilde{V}} e^{\frac{1}{2}\sigma_\eta^2}$
FE	$\frac{\sum_{j=1}^{J} e^{\eta_j}\hat{\beta}_j}{\sum_{j=1}^{J} e^{\eta_j}}$	$\sigma_{\eta b}$	$\sigma_b^2(1+\rho_{\eta b}^2\sigma_\eta^2)e^{\sigma_\eta^2} + \mu_{\tilde{V}} e^{-\frac{1}{2}\sigma_\eta^2}$
FIRC	$\frac{\sum_{j=1}^{J}\hat{\lambda}_j\hat{\beta}_j}{\sum_{j=1}^{J}\hat{\lambda}_j}$	$\frac{\sigma_{\eta b}(1-\tilde{\lambda}^*)}{1+\sigma_\eta^2\tilde{\lambda}(1-\tilde{\lambda}^*)}$	$(i):\quad \left[\sigma_b^2\left(1+\rho_{\eta b}^2 A\right)e^{v_2^*(1-\tilde{\lambda}^*)^2} + \mu_{\tilde{V}} e^{-\frac{1}{2}v_2^*(1-2\tilde{\lambda}^{*2})}\right]\left(1+\sigma_\eta^2 d_1^{*2} v_2^*\right)^{1/2}$ $(ii):\quad \frac{(\sigma_b^{*2}+\mu_{\tilde{V}})[1+\sigma_\eta^2\tilde{\lambda}^*(1-\tilde{\lambda}^*)]^{\frac{1}{2}}}{e^{\frac{1}{2}v_1^{*2}(1-\tilde{\lambda}^*)^2}}$
FIRC+	$\frac{\sum_{j=1}^{J}\hat{\lambda}_j^+(\hat{\beta}_j-\hat{\alpha}\eta_j)}{\sum_{j=1}^{J}\hat{\lambda}_j^+}$	0	$\frac{(\sigma_{b\|\eta}^2+\mu_{\tilde{V}})[1+\sigma_\eta^2\tilde{\lambda}^+(1-\tilde{\lambda}^+)]^{\frac{1}{2}}[1+v_1^+(1-\tilde{\lambda})^2]}{e^{\frac{1}{2}v_1^+(1-\tilde{\lambda}^+)^2}}$

Table 3.3.1: Asymptotic Bias and Variance of the estimators. For UW the results are finite-sample under (3.2.1), with the variance (normalized by J^{-1}) also requiring marginal normality of η_j. The FE, FIRC, and FIRC+ results are asymptotic under (3.2.1) with bivariate normality, and FIRC and FIRC+ also use the Laplace approximations. As an exception, the FIRC+ bias result is finite-sample and only requires that the conditional expectation of β_j is linear in η_j (as happens under bivariate normality).
The FIRC results use the following new definitions:

$$\tilde{\lambda}^* := \sigma_b^{*2}/(\sigma_b^{*2}+\mu_{\tilde{V}}), \quad \text{where } \sigma_b^{*2} := \text{plim}(\hat{\sigma}_b^2) \approx \sigma_b^2[1-2\rho_{\eta b}^2 v_2^*\tilde{\lambda}^*(1-\tilde{\lambda}^*)]$$
$$v_1^* := \frac{\sigma_\eta^2}{1+\sigma_\eta^2\tilde{\lambda}^*(1-\tilde{\lambda}^*)}$$
$$v_2^* := \frac{\sigma_\eta^2}{1+2\sigma_\eta^2\tilde{\lambda}^*(1-\tilde{\lambda}^*)}$$
$$A := \frac{1+(1-\tilde{\lambda}^*)^2 v_2^*}{1+2\sigma_\eta^2\tilde{\lambda}^*(1-\tilde{\lambda}^*).}$$

The FIRC+ results use the following new definitions:

$$v_1^+ := \frac{\sigma_\eta^2}{1+\sigma_\eta^2\tilde{\lambda}^+(1-\tilde{\lambda}^+)}$$
$$\tilde{\lambda}^+ := \sigma_{b|\eta}^2/(\sigma_{b|\eta}^2+\mu_{\tilde{V}}e^{-\eta_j}), \quad \text{where } \sigma_{b|\eta}^2 = \sigma_\eta^2(1-\rho_{\eta b}^2).$$

are *iid* as $J \to \infty$; also Z_j has finite variance which we denote $\sigma_Z^2 := Var(Z_j)$ and this variance is constant in j. So we can apply the Central Limit Theorem to the numerator on the far RHS of (B.7), and we find that $J^{1/2}\bar{Z} \xrightarrow{d} N(0, \sigma_Z^2)$. Applying the Weak Law of Large Numbers to the denominator we get $\text{plim}(\bar{w}) = E(w_j) =: \mu_w$. These results can be combined with Slutsky's Theorem, so we have the main result

$$J^{1/2}(\hat{\beta} - \beta^*) \xrightarrow{d} N\left(0, \frac{\sigma_Z^2}{\mu_w^2}\right). \tag{B.8}$$

To summarize the estimator's probability limit, we have

$$\hat{\beta} \xrightarrow{p} \beta^* = \beta + \frac{\mu_{w(\hat{\beta}-\beta)}}{\mu_w}. \tag{B.9}$$

We refer to the second term as the asymptotic bias of $\hat{\beta}$ (as an estimator of β), and we define $\mu_{w(\hat{\beta}-\beta)} := E[w_j(\hat{\beta}_j - \beta)]$. The asymptotic variance of $\hat{\beta}$ is given by

$$AVar(\hat{\beta}) = \frac{\sigma_Z^2}{\mu_w^2}. \tag{B.10}$$

Then the derivations mostly amount to finding $\mu_{w\hat{\beta}}$, σ_Z^2, and μ_w for each estimator. To compute these expectations we use (a) varying features of and reliance on the bivariate normality assumption for β_j and η_j, and (b) Laplace-type approximations for expectations involving the FIRC and FIRC+ weights, which are otherwise opaque. In the subsections that follow we show the details for each estimator.

B.3.1 Unweighted (UW)

Recall that $\hat{\beta}_{UW} = \sum_{j=1}^{J} \hat{\beta}_j / J$.

Theorem B.1. *If* $\eta_j \sim N(0, \sigma_\eta^2)$*, then* $J^{\frac{1}{2}}(\hat{\beta}_{UW} - \beta) \xrightarrow{d} N(0, \sigma_b^2 + \mu_{\tilde{V}} e^{\frac{1}{2}\sigma_\eta^2})$.

Proof. This estimator is simple enough to be analyzed with the CLT alone, but in terms

of our framework above we have $w_j = 1$ and $Z_j = \hat{\beta}_j - \beta$, so $\mu_w = 1$, $\mu_Z = 0$, and $\sigma_Z^2 = \sigma_b^2 + \mu_{\tilde{V}} E(e^{-\eta_j}) = \sigma_b^2 + \mu_{\tilde{V}} e^{\frac{1}{2}\sigma_\eta^2}$ under the assumption that $\eta_j \sim N(0, \sigma_\eta^2)$ since $e^{-\eta_j}$ is lognormal. Then applying (B.8) yields the result. □

B.3.2 Fixed Effects (FE)

Recall that $\hat{\beta}_{FE} = \sum_{j=1}^{J} e^{\eta_j} \hat{\beta}_j / \sum_{j=1}^{J} e^{\eta_j}$.

Theorem B.2. *If we assume bivariate normality of site-ATEs and centered log precisions, that is*

$$\begin{pmatrix} \beta_j \\ \eta_j \end{pmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} \beta \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_b^2 & \sigma_{\eta b} \\ \sigma_{\eta b} & \sigma_\eta^2 \end{bmatrix} \right) \tag{B.11}$$

then we have

$$\hat{\beta}_{FE} \xrightarrow{p} \beta_{FE}^* = \beta + \sigma_{\eta b}$$

and

$$J^{\frac{1}{2}}(\hat{\beta}_{FE} - \beta_{FE}^*) \xrightarrow{d} N\{0, e^{2\sigma_\eta^2}[\sigma_b^2(1 + \rho_{\eta b}^2 \sigma_\eta^2) + \tilde{V} e^{-\frac{3}{2}\sigma_\eta^2}]\}.$$

Proof. Here we have $w_j = e^{\eta_j}$, $Z_j = e^{\eta_j}(\hat{\beta}_j - \beta_{FE}^*)$, and $\beta_{FE}^* = \frac{\mu_{w\beta}}{\mu_w}$. To start, the log normality implied by (B.11) implies that

$$\mu_w = E(e^{\eta_j}) = e^{\frac{1}{2}\sigma_\eta^2}. \tag{B.12}$$

Then we just need to calculate $\mu_{w(\hat{\beta}-\beta)} = E[e^{\eta_j}(\hat{\beta}_j - \beta)]$ and $\sigma_Z^2 = Var(e^{\eta_j}[\hat{\beta}_j - \beta_{FE}^*])$.

To more easily evaluate these expectations we note that under model (3.2.1) and (B.11) we can decompose the site-ATE estimate $\hat{\beta}_j$ into sampling error e_j, a term predicted by the log precision η_j, and an additional error independent of the log precision, ε_j. In particular,

$$\hat{\beta}_j = \beta_j + e_j = \beta + b_j + e_j = \beta + \alpha\eta_j + \varepsilon_j + e_j \tag{B.13}$$

where the errors satisfy $E(e_j|\eta_j) = 0$, $Var(e_j|\eta_j) =: V_j = \mu_{\tilde{V}} e^{-\eta_j}$, $sgn(e_j) \perp \beta_j$ (thanks to symmetry of the sampling distribution f in (3.2.1)), and $\varepsilon_j \sim N[0, \sigma_b^2(1-\rho_{\eta b}^2)]$ with $\eta_j \perp \varepsilon_j$. We may also write $\alpha = \sigma_{\eta b}/\sigma_\eta^2$ and $\rho_{\eta b} = \sigma_{\eta b}/(\sigma_\eta \sigma_b)$.

We will also use the facts that, under (B.11),

$$\begin{aligned} E(e^{\eta_j}) &= e^{\frac{1}{2}\sigma_\eta^2} \qquad & E(\eta_j e^{\eta_j}) &= \sigma_\eta^2 e^{\frac{1}{2}\sigma_\eta^2} \\ E(e^{2\eta_j}) &= e^{2\sigma_\eta^2} \qquad & E(\eta_j e^{2\eta_j}) &= 2\sigma_\eta^2 e^{2\sigma_\eta^2} \\ & & E(\eta_j^2 e^{2\eta_j}) &= (4\sigma_\eta^4 + \sigma_\eta^2) e^{2\sigma_\eta^2}. \end{aligned} \tag{B.14}$$

These equalities can be confirmed for example by evaluating, for any real numbers p and q, the integral

$$\begin{aligned} E(\eta_j^p e^{q\eta_j}) &= (2\pi\sigma_\eta^2)^{-\frac{1}{2}} \int_{\mathbb{R}} \eta_j^p e^{q\eta_j - \frac{\eta_j^2}{2\sigma_\eta^2}} d\eta_j \\ &= e^{\frac{1}{2}q^2\sigma_\eta^2} E_{q\sigma_\eta^2, \sigma_\eta^2}(\eta_j^p) \end{aligned} \tag{B.15}$$

where $E_{q\sigma_\eta^2, \sigma_\eta^2}(.)$ is an expectation taken over a normal density with mean $q\sigma_\eta^2$ and variance σ_η^2.

Therefore using (B.13) we can write

$$\mu_{w(\hat{\beta}-\beta)} = E(e^{\eta_j} E[\hat{\beta}_j - \beta|\eta_j]) = \frac{\sigma_{\eta b}}{\sigma_\eta^2} E(\eta_j e^{\eta_j}) = \sigma_{\eta b} e^{\frac{1}{2}\sigma_\eta^2} \tag{B.16}$$

where we use the tower law, (B.13), and (B.14) (to evaluate the final expectation). So applying (B.9) the asymptotic bias is simply

$$\beta_{FE}^* - \beta_{FE} = \sigma_{\eta b} e^{\frac{1}{2}\sigma_\eta^2 - \frac{1}{2}\sigma_\eta^2} = \sigma_{\eta b}. \tag{B.17}$$

To evaluate σ_Z^2 (and the asymptotic variance), we first recall that Z_j has mean 0 and

use (B.17) to write

$$\sigma_Z^2 = E(Z_j^2) = E\left[e^{2\eta_j}(\alpha\eta_j + \varepsilon_j + e_j - \sigma_{\eta b})^2\right]. \tag{B.18}$$

And (B.13) and (B.11) imply $E(\varepsilon_j + e_j|\eta_j) = 0$ and $Var(\varepsilon_j + e_j|\eta_j) = \sigma_b^2(1-\rho_{\eta b}^2) + \mu_{\tilde{V}}e^{-\eta_j}$, so further we have

$$\begin{aligned}
\sigma_Z^2 &= E\left[e^{2\eta_j}E\left([\alpha\eta_j + \varepsilon_j + e_j - \sigma_{\eta b}]^2|\eta_j\right)\right] \\
&= E\left(e^{2\eta_j}[\alpha^2\eta_j^2 + \sigma_b^2(1-\rho_{\eta b}^2) + \mu_{\tilde{V}}e^{-\eta_j} + \sigma_{\eta b}^2 - 2\alpha\sigma_{\eta b}\eta_j]\right) \\
&= \alpha^2 E\left(\eta_j^2 e^{2\eta_j}\right) + \sigma_b^2\left(1-\rho_{\eta b}^2\right)E\left(e^{2\eta_j}\right) + \mu_{\tilde{V}}E\left(e^{\eta_j}\right) + \sigma_{\eta b}^2 E\left(e^{2\eta_j}\right) - 2\alpha\sigma_{\eta b}E\left(\eta_j e^{2\eta_j}\right)
\end{aligned} \tag{B.19}$$

where we use the tower law on the first line. Now plugging (B.14) in to (B.19) yields

$$\begin{aligned}
\sigma_Z^2 &= \left(\sigma_{\eta b}^2/\sigma_\eta^2\right)e^{2\sigma_\eta^2} + \sigma_b^2\left(1-\rho_{\eta b}^2\right)e^{2\sigma_\eta^2} + \mu_{\tilde{V}}e^{\frac{1}{2}\sigma_\eta^2} + \sigma_{\eta b}^2 e^{2\sigma_\eta^2} \\
&= e^{2\sigma_\eta^2}\left[\sigma_b^2\left(1+\rho_{\eta b}^2\sigma_\eta^2\right) + \mu_{\tilde{V}}e^{-\frac{3}{2}\sigma_\eta^2}\right].
\end{aligned} \tag{B.20}$$

Finally, from (B.12) we have $\mu_w^2 = e^{\sigma_\eta^2}$, so combined with (B.20) we can see that the asymptotic variance is

$$AVar(\hat{\beta}_{FE}) = \sigma_b^2(1+\rho_{\eta b}^2\sigma_\eta^2)e^{\sigma_\eta^2} + \mu_{\tilde{V}}e^{-\frac{1}{2}\sigma_\eta^2}. \tag{B.21}$$

At last, using (B.8) concludes the proof. □

B.3.3 Fixed Intercept Random Coefficients (FIRC)

Recall that for the FIRC estimator, we have $\hat{\beta}_{FIRC} = \sum_{j=1}^J \hat{\lambda}_j\hat{\beta}_j / \sum_{j=1}^J \hat{\lambda}_j$ where

$$\hat{\lambda}_j := \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + V_j} \quad \text{and} \quad \hat{\sigma}_b^2 := \left(\frac{\sum_{j=1}^J \hat{\lambda}_j^2\left[(\hat{\beta}_{FIRC} - \hat{\beta}_j)^2 - V_j\right]}{\sum_{j=1}^J \hat{\lambda}_j^2}\right)_+$$

and the $()_+$ indicates that $\hat{\sigma}_b^2$ is taken to be 0 when the inside is negative.

Theorem B.3. *Under bivariate normality assumption (C.5),*

$$\hat{\beta}_{FIRC} \xrightarrow{p} \beta^*_{FIRC} \approx \beta + \frac{\sigma_{\eta b}(1-\tilde{\lambda}^*)}{1+\sigma_\eta^2 d_1^*}$$

and

$$J^{\frac{1}{2}}(\hat{\beta}_{FIRC} - \beta^*_{FIRC}) \xrightarrow{d} N(0, \sigma^{*2}_{FIRC}).$$

The asymptotic variance is

$$\sigma^{*2}_{FIRC} \approx \frac{(\sigma_b^{*2} + \mu_{\tilde{V}})(1+\sigma_\eta^2 d_1^*)^{\frac{1}{2}}}{e^{\frac{1}{2}v_1^*(1-\tilde{\lambda}^*)^2}},$$

where $\tilde{\lambda}^* = \sigma_b^{*2}/(\sigma_b^{*2} + \mu_{\tilde{V}})$, $d_1^* = \tilde{\lambda}^*(1-\tilde{\lambda}^*)$, $v_1^* = \sigma_\eta^2/(1+\sigma_\eta^2 d_1^*)$, *and*

$$\hat{\sigma}_b^2 \xrightarrow{p} \sigma_b^{*2} \approx \sigma_b^2 + \sigma_b^2 \rho_{\eta b}^2 (A-1)$$

where

$$A = \frac{1+(1-\tilde{\lambda}^*)^2 v_2^*}{1+2\sigma_\eta^2 d_1^*}.$$

Furthermore, we have the alternative asymptotic variance approximation

$$\sigma^{*2}_{FIRC} \approx \left(\sigma_b^2\left[1+\rho_{\eta b}^2 A\right] e^{v_2^*(1-\tilde{\lambda}^*)^2} + \mu_{\tilde{V}} e^{-\frac{1}{2}v_2^*(1-2\tilde{\lambda}^{*2})}\right)[1+\sigma_\eta^2 d_1^{*2} v_2^*]^{1/2}.$$

Remark.

These approximations, which are shown through simulation (see Section B.6) to be very accurate under our assumptions, are based on Laplace approximations as shown below (Section B.3.3.2). To clarify the approach we used to derive these, we'll first consider the simple case in which site-specific treatment effect,β_j, and the natural logarithm of sampling precision,η_j, are uncorrelated. We then turn to the more challenging task of deriving

Theorem B.3 when β_j and η_j are correlated.

Before jumping into the derivations we recall that our basic CLT decomposition (B.8) cannot be directly applied to $\hat{\beta}_{FIRC}$ because its actual weights do not form an *iid* sequence, an assumption needed to use the WLLN and the CLT. To clarify this point, we emphasize the FIRC estimator's dependence on the number of sites J by using the notation $\hat{\beta}_{FIRC} = \hat{\beta}^{(J)}_{FIRC} = \sum_{j=1}^{J} \hat{\lambda}_j^{(J)} \hat{\beta}_j / \sum_{j=1}^{J} \hat{\lambda}_j^{(J)}$. The notation $\hat{\lambda}_j^{(J)}$ may seem strange, but it comes from the fact that when there are a total of J sites, the FIRC estimator's weight for site j depends on both the sampling variance just in site j (V_j) and the overall effect heterogeneity estimate using data from all J sites ($(\hat{\sigma}_b^2)^{(J)}$), i.e. $\hat{\lambda}_j^{(J)} = (\hat{\sigma}_b^2)^{(J)} / [(\hat{\sigma}_b^2)^{(J)} + V_j]$. This makes clear that the FIRC weights cannot form an *iid* sequence because $(\hat{\sigma}_b^2)^{(J)}$ is, of course, not *iid* as $J \to \infty$ (since both its mean and variance will be changing as it contracts around its probability limit). (And this puts aside the slightly confusing question of how to simultaneously pick j and J when considering $\hat{\lambda}_j^{(J)}$ as an infinite sequence).

Instead, in the derivations below we will work with the asymptotic, not estimated, weights $\lambda_j^* := \frac{\sigma_b^{*2}}{\sigma_b^{*2} + V_j}$ where $\hat{\sigma}_b^2$ is replaced with its probability limit $\sigma_b^{*2} := \text{plim}(\hat{\sigma}_b^2)$. Since this weight *is iid* as $j \to \infty$, we can apply (B.7) and (B.8) to find the asymptotic distribution of $J^{1/2}(\hat{\beta}^*_{FIRC} - \beta^*_{FIRC})$, where $\hat{\beta}^*_{FIRC} := \sum_{j=1}^{J} \lambda_j^* \hat{\beta}_j / \sum_{j=1}^{J} \lambda_j^*$ is the FIRC estimator replacing $\hat{\lambda}_j$ with its probability limit λ_j^*. (Note that $\hat{\beta}^*_{FIRC}$ is not to be confused with β^*_{FIRC}, the probability limit of $\hat{\beta}_{FIRC}$). In fact, this is also the asymptotic distribution of $J^{1/2}(\hat{\beta}_{FIRC} - \beta^*_{FIRC})$ because

$$
\begin{aligned}
J^{1/2}(\hat{\beta}^*_{FIRC} - \beta^*_{FIRC}) & \\
- J^{1/2}(\hat{\beta}_{FIRC} - \beta^*_{FIRC}) & = J^{1/2}\left(\frac{\sum_{j=1}^J \lambda_j^* \hat{\beta}_j}{\sum_{j=1}^J \lambda_j^*} - \frac{\sum_{j=1}^J \hat{\lambda}_j \hat{\beta}_j}{\sum_{j=1}^J \hat{\lambda}_j}\right) \\
& = J^{1/2}\left(\frac{\left(\sum_{j=1}^J \hat{\lambda}_j\right)\left(\sum_{j=1}^J \lambda_j^* \hat{\beta}_j\right) - \left(\sum_{j=1}^J \lambda_j^*\right)\left(\sum_{j=1}^J \hat{\lambda}_j \hat{\beta}_j\right)}{\left(\sum_{j=1}^J \lambda_j^*\right)\left(\sum_{j=1}^J \hat{\lambda}_j\right)}\right) \\
& \xrightarrow{p} 0
\end{aligned}
\tag{B.22}
$$

(using the traditional argument that if $X_n \xrightarrow{d} X$ and $|X_n - Y_n| \xrightarrow{p} 0$ then $Y_n \xrightarrow{d} X$). The last line can be seen by noting that $\hat{\lambda}_j = \lambda_j^* + o_p(1)$ (as $J \to \infty$ and j stays fixed, by the Continuous Mapping Theorem applied to $\hat{\sigma}_b^2$), plugging this result in to the numerator of the second line, and using the appropriate o_p notation arithmetic rules.

The upshot of this detailed argument is that to find the asymptotic distribution of $\hat{\beta}_{FIRC}$ we can replace the estimated weight $\hat{\lambda}_j$ with the asymptotic weight $\lambda_j^* := \sigma_b^{*2}/(\sigma_b^{*2} + V_j)$ in our usual calculations in (B.7)-(B.10).

B.3.3.1 Asymptotic distribution when $\sigma_{\eta b} = 0$

First we derive the asymptotic distribution of $\hat{\beta}^*_{FIRC}$ in this easy special case when in fact $\sigma_{\eta b} = 0$ (as the FIRC model assumes) to illustrate the basic process without getting bogged down in long computations.

Proof. If the data are normally distributed as in (B.11) and $\sigma_{\eta b} = 0$, then FIRC is MLE under the true model so it is consistent for both β and σ_b^2, meaning that $\hat{\beta}^*_{FIRC}$ is asymptotically unbiased (without even needing to calculate (B.9)) and $\hat{\beta}^*_{FIRC} = \sum_{j=1}^J \lambda_j \hat{\beta}_j / \sum_{j=1}^J \lambda_j$, where $\lambda_j = \sigma_b^2/(\sigma_b^2 + V_j)$ is the "true" weight using the true effect heterogeneity $\sigma_b^2 = \text{plim}(\hat{\sigma}_b^2)$.

Then in (B.7), we have $w_j = \lambda_j$ and $Z_j = \lambda_j(\hat{\beta}_j - \beta)$. So $\mu_w = E(\lambda_j)$, and

$$\begin{aligned}
\sigma_Z^2 &= E[\lambda_j^2(b_j + e_j)^2] \\
&= E[\lambda_j^2 E((b_j + e_j)^2|\eta_j)] \\
&= E[\lambda_j^2(\sigma_b^2 + V_j)] \\
&= \sigma_b^2 E(\lambda_j)
\end{aligned}$$

where the first line uses the fact $E(Z_j) = 0$ and (B.13), the second line uses the tower law, the third line uses (B.13), and the last line follows from the definition of λ_j. Using the Laplace approximations listed in Section B.3.3.2 below (and derived in Section B.4) to evaluate $E(\lambda_j)$, we find that the asymptotic variance is

$$AVar(\hat{\beta}^*_{FIRC}) = \frac{\sigma_b^2 E[\lambda_j]}{E[\lambda_j]^2} = \frac{\sigma_b^2}{E[\lambda_j]} \approx \frac{(\sigma_b^2 + \mu_{\tilde{V}})(1 + \sigma_\eta^2 d_1)^{1/2}}{e^{\frac{1}{2}v_1(1-\tilde{\lambda})^2}}$$

where $d_1 = \tilde{\lambda}(1 - \tilde{\lambda})$, $\tilde{\lambda} = \sigma_b^2/(\sigma_b^2 + \mu_{\tilde{V}})$, and $v_1 = \sigma_\eta^2/(1 + \sigma_\eta^2 d_1)$. Then applying (B.8) to $\hat{\beta}^*_{FIRC}$ and recalling that $\hat{\beta}_{FIRC}$ has the same asymptotic distribution as $\hat{\beta}^*_{FIRC}$ gives the special case of Theorem B.3 and concludes the proof. □

Remark. Note that d_1 is bounded by $0 \leq d_1 \leq 0.25$, reaching its minimum when $\tilde{\lambda}$ is 0 or 1 and reaching its maximum when $\tilde{\lambda}$ is 0.5. Also, v_1 is bounded by $0 \leq v_1 \leq \sigma_\eta^2$.

B.3.3.2 Laplace Approximations Useful for FIRC

To derive (B.15) and results below, we apply a Laplace transform to obtain the following

approximations:

$$
\begin{aligned}
E(\lambda_j) &\approx \frac{\tilde{\lambda} e^{\frac{1}{2} v_1 (1-\tilde{\lambda})^2}}{(1+\sigma_\eta^2 d_1)^{\frac{1}{2}}}, & E(\lambda_j \eta_j) &\approx \frac{v_1 (1-\tilde{\lambda}) \tilde{\lambda} e^{\frac{1}{2} v_1 (1-\tilde{\lambda})^2}}{(1+\sigma_\eta^2 d_1)^{\frac{1}{2}}} \\
E(\lambda_j^2) &\approx \frac{\tilde{\lambda}^2 e^{2 v_2 (1-\tilde{\lambda})^2}}{(1+2\sigma_\eta^2 d_1)^{\frac{1}{2}}}, & E(\eta_j \lambda_j^2) &\approx \frac{2 v_2 (1-\tilde{\lambda}) \tilde{\lambda}^2 e^{2 v_2 (1-\tilde{\lambda})^2}}{(1+2\sigma_\eta^2 d_1)^{\frac{1}{2}}}, \\
E(\eta_j^2 \lambda_j^2) &\approx \frac{[4 v_2^2 (1-\tilde{\lambda})^2 + v_2] \tilde{\lambda}^2 e^{2 v_2 (1-\tilde{\lambda})^2}}{(1+2\sigma_\eta^2 d_1)^{\frac{1}{2}}}, & E(e^{-\eta_j} \lambda_j^2) &\approx \frac{\tilde{\lambda}^2 e^{\frac{1}{2} v_2 (1-2\tilde{\lambda})^2}}{(1+2\sigma_\eta^2 d_1)^{\frac{1}{2}}}
\end{aligned}
\tag{B.23}
$$

where $v_1 := \sigma_\eta^2/(1+\sigma_\eta^2 d_1)$ and $v_2 := \sigma_\eta^2/(1+2\sigma_\eta^2 d_1)$, the subscripts 1 and 2 denoting which power of λ_j these terms appear in (i.e. v_1 appearing in expectations with λ_j^1 and v_2 appearing in expectations with λ_j^2). A general derivation underlying these results is given in Section B.4.

B.3.3.3 Asymptotic distribution when $\sigma_{\eta b} \neq 0$

In the general case, FIRC is the MLE under the wrong model and, as we will show, inconsistent for both β and σ_b^2. We prove the fully general Theorem B.3 in pieces.

Setup

As before, we apply (B.7)-(B.10) to $\hat{\beta}^*_{FIRC} = \sum_{j=1}^J \lambda_j^* \hat{\beta}_j / \sum_{j=1}^J \lambda_j^*$ where $\lambda_j^* = \sigma_b^{*2}/(\sigma_b^{*2} + V_j)$ and now $\sigma_b^{*2} := \text{plim}(\hat{\sigma}_b^2) \neq \sigma_b^2$. Now in (B.7), we have $w_j = \lambda_j^*$ and $Z_j = \lambda_j^*(\hat{\beta}_j - \beta^*_{FIRC})$. The Laplace approximations we will use here are the same as those in (B.19) except for σ_b^{*2} replacing σ_b^2 and the same notation affecting the λ terms.

Asymptotic bias

Proof. To find the asymptotic bias $\beta^*_{FIRC} - \beta$ using (B.9), we have $\mu_w = E(\lambda_j^*)$ and

$$
\begin{aligned}
\mu_{w(\hat{\beta}-\beta)} &= E[\lambda_j^*(\hat{\beta}_j - \beta)] \\
&= E[\lambda_j^* E(\hat{\beta}_j - \beta | \eta_j)] \\
&= \alpha E[\lambda_j^* \eta_j].
\end{aligned}
$$

The second line uses the tower law and the third line uses (B.13). So from (B.9) the asymp-

totic bias of $\hat{\beta}^*_{FIRC}$, which we also write with the notation B_{FIRC}, is

$$B_{FIRC} := \beta^*_{FIRC} - \beta = \frac{\mu_{w(\hat{\beta}-\beta)}}{\mu_w} = \alpha \frac{E(\lambda^*_j \eta_j)}{E(\lambda^*_j)} \approx \alpha(1-\tilde{\lambda}^*)v_1 = \frac{\sigma_{\eta b}(1-\tilde{\lambda}^*)}{1+\sigma^2_\eta d^*_1}, \quad \text{(B.24)}$$

where the third line plugs in Laplace approximations for the expectations (recall that $d^*_1 := \tilde{\lambda}^*(1-\tilde{\lambda}^*)$). The last expression shows how our asymptotics reflect the finite-sample fact (Theorem 3.2) that FIRC is never more biased than FE. □

Asymptotic variance (similar to FIRC+ expression)

Proof. Next, for the asymptotic variance we have

$$\begin{aligned}
\sigma^2_Z &= E[\lambda^{*2}_j(\alpha\eta_j + \varepsilon_j + e_j - B_{FIRC})^2] \\
&= E[\lambda^{*2}_j E[(\alpha\eta_j + \varepsilon_j + e_j - B_{FIRC})^2 | \eta_j]] \\
&= E[(\lambda^{*2}_j E[(\varepsilon_j + e_j)^2 + (\alpha\eta_j - B_{FIRC})^2 | \eta_j]] \\
&= E[\lambda^{*2}_j(\sigma^2_{b|\eta} + V_j)] + E[\lambda^{*2}_j(\alpha\eta_j - B_{FIRC})^2].
\end{aligned} \quad \text{(B.25)}$$

The first line uses (B.13), the second line uses the tower law, the third line again uses (B.13) (especially the fact that the errors have conditional mean 0), and the fourth line uses (B.13). Considering the first term in the last line of (B.25), we have

$$\begin{aligned}
E[\lambda^{*2}_j(\sigma^2_{b|\eta} + V_j)] &= E\left\{\lambda^{*2}_j[\sigma^{*2}_b + V_j + (\sigma^2_{b|\eta} - \sigma^{*2}_b)]\right\} \\
&= E\left\{\lambda^{*2}_j(\sigma^{*2}_b + V_j) + \lambda^{*2}_j(\sigma^2_{b|\eta} - \sigma^{*2}_b)\right\} \\
&= \sigma^{*2}_b E(\lambda^*_j) + (\sigma^2_{b|\eta} - \sigma^{*2}_b)E[\lambda^{*2}_j].
\end{aligned}$$

On the first line we add and subtract σ^{*2}_b, and after the second line we simplify the first term by using the definition of λ^*_j. And from our derivation of σ^{*2}_b, coming later, and in particular (C.24), we see that in fact the second term in (B.25) satisfies

$$E[\lambda^{*2}_j(\alpha\eta_j - B_{FIRC})^2] = -(\sigma^2_{b|\eta} - \sigma^{*2}_b)E(\lambda^{*2}_j) \quad \text{(B.26)}$$

so we get a cancellation in (B.25) and the simple, familiar result that

$$\sigma_Z^2 = \sigma_b^{*2} E(\lambda_j^*).$$

Then applying (B.10) we get the asymptotic variance

$$AVar(\hat{\beta}_{FIRC}^*) = \frac{\sigma_b^{*2} E(\lambda_j^*)}{E^2(\lambda_j^*)} = \frac{\sigma_b^{*2}}{E(\lambda_j^*)} \approx \frac{(\sigma_b^* + \mu_{\tilde{V}})(1 + \sigma_\eta^2 {d_1}^*)^{1/2}}{e^{\frac{1}{2} v_1 (1-\tilde{\lambda}^*)^2}},$$

where the last expression uses the appropriate Laplace approximation from (B.23), replacing σ_b^2 everywhere with σ_b^{*2} (although (B.23) approximates moments of λ_j, not λ_j^*, a quick look at the general Laplace approximation derivation confirms that this is the only difference).

Finally, as before, we apply (B.8) to $\hat{\beta}_{FIRC}^*$ and recall that $\hat{\beta}_{FIRC}$ has the same asymptotic distribution as $\hat{\beta}_{FIRC}^*$ to conclude the proof. □

Alternate approximation for the asymptotic variance (similar to FE expression)

Proof. In the previous asymptotic variance derivation we could have continued to simplify (B.25) differently, instead proceeding as

$$\begin{aligned} \sigma_Z^2 &= E[\lambda_j^{*2}(\sigma_{b|\eta}^2 + V_j)] + E[\lambda_j^{*2}(\alpha\eta_j - B_{FIRC})^2] \\ &= \left[\sigma_{b|\eta}^2 E(\lambda_j^{*2}) + \mu_{\tilde{V}} E(\lambda_j^{*2} e^{-\eta_j})\right] + \left[\alpha^2 E(\lambda_j^{*2}\eta_j^2) - 2\alpha B_{FIRC} E(\lambda_j^{*2}\eta_j) + B_{FIRC}^2 E(\lambda_j^{*2})\right] \end{aligned} \tag{B.27}$$

where for the first term we have not added and subtracted σ_b^{*2} in an attempt to lower the power on λ_j^*, and in the second term we simply expand the square. Before plugging in any approximations to (B.26), we consider the key ratio

$$\frac{\sigma_Z^2}{\mu_w^2} = \sigma_{b|\eta}^2 \frac{E(\lambda_j^{*2})}{E^2(\lambda_j^*)} + \mu_{\tilde{V}} \frac{E(\lambda_j^{*2} e^{-\eta_j})}{E^2(\lambda_j^*)} + \alpha^2 \frac{E(\lambda_j^{*2}\eta_j^2)}{E^2(\lambda_j^*)} - 2\alpha B_{FIRC} \frac{E(\lambda_j^{*2}\eta_j)}{E^2(\lambda_j^*)} + B_{FIRC}^2 \frac{E(\lambda_j^{*2})}{E^2(\lambda_j^*)} \tag{B.28}$$

Plugging in our Laplace approximations to (B.27), with the notation

$$\begin{aligned}\frac{E(\lambda_j^{*2})}{E^2(\lambda_j^*)} &\approx r_1 := e^{2v_2^*(1-\tilde{\lambda}^*)^2 - v_1^*(1-\tilde{\lambda}^*)^2}\frac{1+\sigma_\eta^2 d_1^*}{(1+2\sigma_\eta^2 d_1^*)^{1/2}} \\ \frac{E(\lambda_j^{*2}e^{-\eta_j})}{E^2(\lambda_j^*)} &\approx r_2 := e^{-v_2^*(1/2-\tilde{\lambda}^{*2})^2}\frac{1+\sigma_\eta^2 d_1^*}{(1+2\sigma_\eta^2 d_1^*)^{1/2}},\end{aligned}$$

we have

$$\begin{aligned}\frac{\sigma_Z^2}{\mu_w^2} &\approx \sigma_{b|\eta}^2 \cdot r_1 + \mu_{\tilde{V}} \cdot r_2 + \alpha^2 \cdot \left[4v_2^{*2}(1-\tilde{\lambda}^*)^2 + v_2^*\right] r_1 - 2\alpha \cdot \alpha v_1^*(1-\tilde{\lambda}^*) \cdot 2v_2^*(1-\tilde{\lambda}^*)r_1 \\ &\qquad +\alpha^2 v_2^{*2}(1-\tilde{\lambda}^*)^2 \cdot r_1 \\ &\approx \sigma_{b|\eta}^2 r_1 + \mu_{\tilde{V}} r_2 + \alpha^2 \left[v_2^* + v_2^{*2}(1-\tilde{\lambda}^*)^2\right] r_1 \\ &\approx \sigma_b^2(1-\rho_{\eta b}^2)r_1 + \mu_{\tilde{V}} r_2 + \sigma_b^2 \rho_{\eta b}^2 \sigma_\eta^{-2}\left[v_2^* + v_2^{*2}(1-\tilde{\lambda}^*)^2\right] r_1 \\ &\approx \sigma_b^2\left[1 - \rho_{\eta b}^2 + \rho_{\eta b}^2\sigma_\eta^{-2}v_2^*\left(1 + v_2^*(1-\tilde{\lambda}^*)^2\right)\right] r_1 + \mu_{\tilde{V}} r_2 \\ &\approx \sigma_b^2\left(1 + \rho_{\eta b}^2 A\right) r_1 + \mu_{\tilde{V}} r_2\end{aligned}$$

where on the second line we combine the α^2 terms (using the fact that $v_1^* - v_2^* = O(v_1^2)$ so they are interchangeable up to our order of approximation), on the third line we plug in characterizations of $\sigma_{b|\eta}^2$ and α^2 in terms of σ_b^2, on the fourth line we combine the σ_b^2 terms, and on the fifth line we use the notation $A := \frac{1+v_2^*(1-\tilde{\lambda}^*)^2}{1+2\sigma_\eta^2 d_1^*}$. If we simplify the exponential term in r_1 by using the fact that $v_1^* - v_2^* = O(v_1^2)$, and notice that in the definitions of r_1 and r_2 we can also slightly simplify by writing

$$\frac{1+\sigma_\eta^2 d_1^*}{(1+2\sigma_\eta^2 d_1^*)^{1/2}} = \left[\frac{(1+\sigma_\eta^2 d_1^*)^2}{1+2\sigma_\eta^2 d_1^*}\right]^{1/2} = \left[\frac{1+2\sigma_\eta^2 d_1^* + \sigma_\eta^4 d_1^{*2}}{1+2\sigma_\eta^2 d_1^*}\right]^{1/2} = [1+\sigma_\eta^2 d_1^{*2} v_2^*]^{1/2},$$

then we get that

$$\frac{\sigma_Z^2}{\mu_w^2} \approx \left(\sigma_b^2\left[1+\rho_{\eta b}^2 A\right] e^{v_2^*(1-\tilde{\lambda}^*)^2} + \mu_{\tilde{V}} e^{-\frac{1}{2}v_2^*(1-2\tilde{\lambda}^{*2})}\right)[1+\sigma_\eta^2 d_1^{*2} v_2^*]^{1/2}. \tag{B.29}$$

□

Derivation of σ_b^{*2} and (B.26)

Proof. To understand the asymptotic behavior of the FIRC heterogeneity estimator, $\hat{\sigma}_b^2$, we consider the probability limit of its estimating equation:

$$\begin{aligned}\sigma_b^{*2} &= \text{plim}\left(\frac{J^{-1}\sum_{j=1}^{J}\hat{\lambda}_j^2[(\hat{\beta}_j-\hat{\beta}_{FIRC})^2-V_j]}{J^{-1}\sum_{j=1}^{J}\hat{\lambda}_j^2}\right)\\ &= \text{plim}\left(\frac{J^{-1}\sum_{j=1}^{J}\lambda_j^{*2}[(\hat{\beta}_j-\beta_{FIRC}^*)^2-V_j]}{J^{-1}\sum_{j=1}^{J}\lambda_j^{*2}}\right)\end{aligned}$$

where on the second line we replace $\hat{\lambda}_j$ and $\hat{\beta}_{FIRC}^*$ for the same reasons and by the same argument discussed in the remark on Theorem B.3. Continuing on, we use the WLLN and the tower law to get

$$\begin{aligned}\sigma_b^* &= \frac{E\left[\lambda_j^{*2}E[(\hat{\beta}_j-\beta_{FIRC}^*)^2-V_j|\eta_j]\right]}{E(\lambda_j^{*2})}\\ &= \frac{E\left[\lambda_j^{*2}E[(\alpha\eta_j-B_{FIRC}+\varepsilon_j+e_j)^2-V_j|\eta_j]\right]}{E(\lambda_j^{*2})}\\ &= \frac{E\left[\lambda_j^{*2}[(\alpha\eta_j-B_{FIRC})^2+\sigma_{b|\eta}^2+V_j-V_j]\right]}{E(\lambda_j^{*2})}\\ &= \sigma_{b|\eta}^2+\frac{E\left[\lambda_j^{*2}(\alpha\eta_j-B_{FIRC})^2\right]}{E(\lambda_j^{*2})},\end{aligned} \tag{B.30}$$

where the second and third lines use (B.13) just like in (B.25) above, and the last line simplifies. The last line immediately implies (B.26). Continuing further by expanding the

square, we have

$$\sigma_b^{*2} = \sigma_{b|\eta}^2 + \alpha^2 \frac{E(\lambda_j^{*2}\eta_j^2)}{E(\lambda_j^{*2})} - 2\alpha B_{FIRC} \frac{E(\lambda_j^{*2}\eta_j)}{E(\lambda_j^{*2})} + B_{FIRC}^2 \tag{B.31}$$

and from the Laplace approximations in (B.23) we have

$$\begin{aligned}
\frac{E\left(\lambda_j^{*2}\eta_j^2\right)}{E\left(\lambda_j^{*2}\right)} &\approx \left[4v_2^{*2}\left(1-\tilde{\lambda}^*\right)^2 + v_2^*\right] \\
\frac{E\left(\lambda_j^{*2}\eta_j\right)}{E\left(\lambda_j^{*2}\right)} &\approx 2v_2^*\left(1-\tilde{\lambda}^*\right).
\end{aligned}$$

So plugging these approximations and expression (B.24) for B_{FIRC} into (B.31) we get

$$\begin{aligned}
\sigma_b^{*2} &\approx \sigma_{b|\eta}^2 + \alpha^2\left[4v_2^{*2}(1-\tilde{\lambda}^*)^2 + v_2^*\right] - 2\alpha\cdot\left[\alpha(1-\tilde{\lambda}^*)v_1^*\right]\cdot\left[2v_2^*(1-\tilde{\lambda}^*)\right] + \left[\alpha(1-\tilde{\lambda}^*)v_1^*\right]^2 \\
&\approx \sigma_{b|\eta}^2 + \alpha^2\left[4v_2^{*2}(1-\tilde{\lambda}^*)^2 + v_2^* - 4(1-\tilde{\lambda}^*)^2 v_1^* v_2^* + (1-\tilde{\lambda}^*)^2 v_1^{*2}\right] \\
&\approx \sigma_{b|\eta}^2 + \alpha^2 v_2^*\left[1 + (1-\tilde{\lambda}^*)^2 v_2^*\right] \\
&\approx \sigma_b^2 + \sigma_b^2\rho_{\eta b}^2\left[\frac{1+(1-\tilde{\lambda}^*)^2 v_2^*}{1+2\sigma_\eta^2 d_1^*} - 1\right] = \sigma_b^2 + \sigma_b^2\rho_{\eta b}^2(A-1).
\end{aligned} \tag{B.32}$$

where the third line uses the convenient fact that $v_1^* - v_2^* = v_1^* v_2^* d_1^* = O(v_1^2)$ (which is the order of our Laplace approximations, so v_1 can be replaced with v_2 or vice versa), and the last line uses the definitions of $\sigma_{b|\eta}^2$, α, and v_2^*. This concludes the proof. □

B.3.4 FIRC+

Recall for FIRC+ that we have the estimator:

$$\hat{\beta}_{FIRC+} := \frac{\sum_{j=1}^{J} \hat{\lambda}_j^+ (\hat{\beta}_j - \hat{\alpha}\eta_j)}{\sum_{j=1}^{J} \hat{\lambda}_j^+} \tag{B.33}$$

where

$$\hat{\lambda}_j^+ := \frac{\hat{\sigma}^2_{b|\eta}}{\hat{\sigma}^2_{b|\eta} + V_j}, \quad \hat{\sigma}^2_{b|\eta} := \left(\frac{\sum_{j=1}^{J} \hat{\lambda}_j^{+2} \left[(\hat{\beta}_{FIRC+} - \hat{\beta}_j - \hat{\alpha}\eta_j)^2 - V_j \right]}{\sum_{j=1}^{J} \hat{\lambda}_j^{+2}} \right)_+$$

(and the $()_+$ indicates that $\hat{\sigma}^2_{b|\eta}$ is taken to be 0 when the inside is negative), and

$$\hat{\alpha} := \frac{\sum\limits_{j=1}^{J} \hat{\lambda}_j^+ \eta_j (\hat{\beta}_j - \bar{\beta})}{\sum\limits_{j=1}^{J} \hat{\lambda}_j^+ \eta_j (\eta_j - \bar{\eta})}, \quad \text{with} \quad \bar{\beta} := \frac{\sum\limits_{j=1}^{J} \hat{\lambda}_j^+ \hat{\beta}_j}{\sum\limits_{j=1}^{J} \hat{\lambda}_j^+} \quad \text{and} \quad \bar{\eta} := \frac{\sum\limits_{j=1}^{J} \hat{\lambda}_j^+ \eta_j}{\sum\limits_{j=1}^{J} \hat{\lambda}_j^+}.$$

Theorem B.4. *Under bivariate normality assumption* (B.11)*,*

$$J^{\frac{1}{2}} (\hat{\beta}_{FIRC+} - \beta) \xrightarrow{d} N(0, \sigma^2_{FIRC+})$$

where

$$\sigma^2_{FIRC+} \approx \frac{(\sigma^2_{b|\eta} + \mu_{\tilde{V}})[1 + v_1^+ (1 - \tilde{\lambda}^+)^2](1 + \sigma^2_\eta d_1^+)^{\frac{1}{2}}}{e^{\frac{1}{2} v_1^+ (1 - \tilde{\lambda}^+)^2}},$$

with $\tilde{\lambda}^+ = \sigma^2_{b|\eta}/(\sigma^2_{b|\eta} + \mu_{\tilde{V}})$*,* $v_1^+ = \sigma^2_\eta/(1 + \sigma^2_\eta d_1^+)$*, and* $d_1^+ = \tilde{\lambda}^+(1 - \tilde{\lambda}^+)$*.*

Setup and Asymptotic Bias

Proof. We make two changes to the general derivation strategy we have used earlier. First, we consider the FIRC+ estimator where the estimated weight $\hat{\lambda}_j^+$ (a function of the estimate $\hat{\sigma}^2_{b|\eta}$) is replaced by its probability limit λ_j^+ (a function of the true conditional variance $\sigma^2_{b|\eta}$). This is done for the same reason and same justification as in the FIRC derivations. (And since FIRC+ is MLE under the bivariate normality assumption of the theorem, $\hat{\sigma}^2_{b|\eta}$ is consistent

for $\sigma^2_{b|\eta}$ and the continuous mapping theorem then implies that $\hat{\lambda}^+_j \xrightarrow{p} \lambda^+_j := \sigma^2_{b|\eta}/(\sigma^2_{b|\eta} + V_j))$.

And second, we jointly analyze the FIRC+ estimators of β and α, that is we consider the multivariate estimator $\hat{\theta}^*_{FIRC+} := (\hat{\beta}^*_{FIRC+}, \hat{\alpha}^*_{FIRC+})^T$, where the * indicates that we have already substituted λ^+_j for $\hat{\lambda}^+_j$ in these estimators. This notation comes from writing the FIRC+ regression model as

$$\begin{aligned}\hat{\beta}_j &= \beta + \alpha\eta_j + \varepsilon_j + e_j \\ &= \begin{pmatrix} 1 & \eta_j \end{pmatrix} \begin{pmatrix} \beta \\ \alpha \end{pmatrix} + \varepsilon_j + e_j =: x_j^T\theta + \varepsilon_j + e_j\end{aligned} \tag{B.34}$$

where $x_j := \begin{pmatrix} 1 & \eta_j \end{pmatrix}^T$, $\theta := \begin{pmatrix} \beta & \alpha \end{pmatrix}^T$, and $\hat{\beta}_j|\eta_j \sim N(x_j^T\theta, \sigma^2_{b|\eta} + V_j)$. Thus the FIRC+ estimator (with $\sigma^2_{b|\eta}$ known) may be written as

$$\begin{aligned}\hat{\theta}^*_{FIRC+} &= \left(\sum_{j=1}^J \lambda^+_j x_j x_j^T\right)^{-1} \sum_{j=1}^J \lambda^+_j x_j \hat{\beta}_j \\ &= \theta + \left(\sum \lambda^+_j x_j x_j^T\right)^{-1} \sum \lambda^+_j x_j (\varepsilon_j + e_j).\end{aligned} \tag{B.35}$$

This will turn out to be easier to analyze than $\hat{\beta}^*_{FIRC+}$ alone, which depends on $\hat{\alpha}^*_{FIRC+}$.

For the asymptotic bias, we simply note that under the bivariate normality assumption (B.11) FIRC+ is the MLE under the correct model (whether or not $\sigma^2_{b|\eta}$ is known), so $\hat{\theta}^*_{FIRC+} \xrightarrow{p} \theta$, which means the estimator is asymptotically unbiased. In fact, FIRC+ is also finite-sample unbiased for θ; although we do not give a proof here, it is very easy to see by applying the tower law (conditioning on all η_j) to the estimator. □

For intuition as to why we cannot *also* send $\hat{\alpha}_{FIRC+}$ to its probability limit (α) when analyzing $\hat{\beta}_{FIRC+}$, like we did with $\hat{\sigma}^2_{b|\eta}$, note that in fact an argument like (B.22) that justifies doing this with $\hat{\sigma}^2_{b|\eta}$ would fail for $\hat{\alpha}_{FIRC+}$. If we write $\hat{\beta}^{(\alpha)}_{FIRC+}$ for the FIRC+

estimator $\hat{\beta}_{FIRC+}$ where we have replaced $\hat{\alpha}_{FIRC+}$ with α, it is fairly quick to see that

$$\begin{aligned} J^{1/2}(\hat{\beta}^{(\alpha)}_{FIRC+} - \beta) - J^{1/2}(\hat{\beta}_{FIRC+} - \beta) &= J^{1/2}\bar{\eta}(\alpha - \hat{\alpha}_{FIRC+}) \\ &\xrightarrow{p} 0. \end{aligned} \tag{B.36}$$

The first line is basic algebra, and to understand the lack of convergence in probability, note that

$$\begin{aligned} Var\left[J^{1/2}\bar{\eta}\left(\alpha - \hat{\alpha}_{FIRC+}\right)\right] &= JE\left[\bar{\eta}^2(\alpha - \hat{\alpha}_{FIRC+})^2\right] \\ &= JE\left[\bar{\eta}^2 Var\left(\hat{\alpha}_{FIRC+}|\eta_j\,\forall j\right)\right] \end{aligned}$$

and that by standard weighted least squares results $Var(\hat{\alpha}_{FIRC+}|\eta_j\,\forall j)$ is only of order $J^{-1/2}$, so although (B.36) has mean 0 (since $\hat{\alpha}_{FIRC+}$ is conditionally unbiased) it does not have decreasing variance as $J \to \infty$, meaning that it cannot converge in distribution to 0 and thus cannot converge in probability to 0.

Asymptotic Variance

Proof. Then to find the asymptotic distribution of the estimator, we will consider

$$\begin{aligned} \sqrt{J}(\hat{\theta}^* - \theta) &= \sqrt{J}\cdot\left(\sum \lambda_j^+ x_j x_j^T\right)^{-1}\sum \lambda_j^+ x_j(\varepsilon_j + e_j) \\ &= \sqrt{J}\cdot\left(\tfrac{1}{J}\sum \lambda_j^+ x_j x_j^T\right)^{-1}\tfrac{1}{J}\sum \lambda_j^+ x_j(\varepsilon_j + e_j) \\ &:= \sqrt{J}\left(\tfrac{1}{J}\sum c_j\right)^{-1}\tfrac{1}{J}\sum Z_j \\ &\xrightarrow{d} \mu_c^{-1} N_2\left(\mathbf{0}, \sigma_Z^2\right) \end{aligned}$$

where we use the notation $c_j := \lambda_j^+ x_j x_j^T$, $Z_j := \lambda_j^+ x_j(\varepsilon_j + e_j)$, $\mu_c := E(c_j)$, and $\sigma_Z^2 := Var(Z_j) = E(Z_j Z_j^T)$. The last line follows from multivariate forms of Slutsky's theorem, the WLLN, and the CLT, analogously to our basic approach (B.7)-(B.8). For the CLT, note that Z_j indeed has expectation 0 because in the tower law (conditioning on η_j) $E(\varepsilon_j + e_j|\eta_j) = 0$ as we stated in (B.13). Then we just need to calculate μ_c and σ_Z^2 to find the asymptotic

variance

$$\sigma^2_{FIRC+} = \mu_c^{-1}\sigma^2_Z\mu_c^{-1}.$$

Note that

$$Z_j {Z_j}^T = {\lambda_j^+}^2(\varepsilon_j + e_j)^2 x_j x_j^T = \lambda_j^+(\varepsilon_j + e_j)^2 c_j \tag{B.37}$$

and

$$x_j x_j^T = \begin{bmatrix} 1 & \eta_j \\ \eta_j & \eta_j^2 \end{bmatrix}$$

so for the asymptotic variance we have

$$\begin{aligned}
\sigma^2_{FIRC+} &= \mu_c^{-1} E\left(Z_j Z_j^T\right)\mu_c^{-1} \\
&= \mu_c^{-1} E\left({\lambda_j^+}^2 E\left[(\varepsilon_j + e_j)^2|\eta_j\right] x_j x_j^T\right)\mu_c^{-1} \\
&= \mu_c^{-1} E\left({\lambda_j^+}^2(\sigma^2_{b|\eta} + V_j) x_j x_j^T\right)\mu_c^{-1} \\
&= \mu_c^{-1}\sigma^2_{b|\eta} E\left(\lambda_j^+ x_j x_j^T\right)\mu_c^{-1} \\
&= \sigma^2_{b|\eta}\mu_c^{-1} \\
&= \sigma^2_{b|\eta}\begin{bmatrix} E(\lambda_j^+) & E(\lambda_j^+\eta_j) \\ E(\lambda_j^+\eta_j) & E(\lambda_j^+\eta_j^2) \end{bmatrix}^{-1} \\
\sigma^2_{FIRC+} &= \frac{\sigma^2_{b|\eta}}{E(\lambda_j^+)E(\lambda_j^+\eta_j^2) - E^2(\lambda_j^+\eta_j)}\begin{bmatrix} E(\lambda_j^+\eta_j^2) & -E(\lambda_j^+\eta_j) \\ -E(\lambda_j^+\eta_j) & E(\lambda_j^+) \end{bmatrix}
\end{aligned}$$

where the second line uses the tower law, the third line uses (B.13), the fourth line uses the definition of λ_j^+, and the fifth line uses the definitions of c_j and μ_c. Substituting the Laplace approximations

$$\begin{aligned}
E(\lambda_j^+) &\approx \frac{\tilde{\lambda}^+ e^{\frac{1}{2}v_1^+(1-\tilde{\lambda}^+)^2}}{(1+\sigma_\eta^2 d_1^+)^{\frac{1}{2}}} \\
E(\lambda_j^+\eta_j) &\approx v_1^+(1-\tilde{\lambda}^+)E(\lambda_j^+) \\
E(\lambda_j^+\eta_j^2) &\approx v_1^+[1 + v_1^+(1-\tilde{\lambda}^+)^2]E(\lambda_j^+)
\end{aligned}$$

into the (1, 1) element of this final matrix, we get that the asymptotic variance of $\hat{\beta}_{FIRC+}$ is approximately

$$\sigma^2_{FIRC+} \approx (\sigma^2_{b|\eta} + \mu_{\tilde{V}})[1 + v_1^+(1-\tilde{\lambda}^+)^2][1 + \sigma^2_\eta \tilde{\lambda}^+(1-\tilde{\lambda}^+)]^{\frac{1}{2}} e^{-\frac{1}{2}v_1^+(1-\tilde{\lambda}^+)^2},$$

concluding the proof. □

B.4 Laplace Approximations

B.4.1 General Result

At various points in the asymptotic derivations we want to approximate $E\left(f(\eta_j)\eta_j^k\right)$ for some *fixed* positive function f where the expectation is with respect to $\eta_j \sim N(0, \sigma_\eta^2)$. In particular we need the cases when $k = 0, 1, 2$ and $f(\eta_j) = \lambda_j^*, \lambda_j^{*^2}, \lambda_j^{*^2} e^{-\eta_j}$ (where the $*$ indicates that λ_j^* depends on a fixed probability limit $\sigma_b^{*^2}$ instead of a random estimate $\hat{\sigma}_b^2$, whether for FIRC or FIRC+), but the Laplace-type approximations we derive below apply more generally.

Theorem B.5. *Suppose $h(\eta_j) := \frac{-1}{2\sigma_\eta^2}\eta_j^2 + \log f(\eta_j)$ has a bounded third derivative. Then*

$$E_{0,\sigma_\eta^2}\left(\eta_j^k f(\eta_j)\right) = \frac{v^{1/2}}{\sigma_\eta} f(0) e^{Q(\hat{\eta})} E_{\hat{\eta},v}\left(\eta_j^k\right)\left[1 + O\left(\frac{E_{\hat{\eta},v}\left(\eta_j^{k+3}\right)}{E_{\hat{\eta},v}\left(\eta_j^k\right)}\right)\right] \tag{B.38}$$

where $Q(\eta_j) := h'(0)\cdot\eta_j + \frac{1}{2}h''(0)\cdot\eta_j^2$, $\hat{\eta} := \frac{-h'(0)}{h''(0)}$ (the maximizer of Q), and $v := \frac{-1}{Q''(\hat{\eta})}$, and on the LHS E_{0,σ_η^2} denotes expectation with respect to $\eta_j \sim N(0,\sigma_\eta^2)$ while on the RHS $E_{\hat{\eta},v}$ denotes expectation with respect to $\eta_j \sim N(\hat{\eta}, v)$.

Proof. First note that

$$\begin{aligned} E\left(f(\eta_j)\eta_j^k\right) &= (2\pi\sigma_\eta^2)^{\frac{-1}{2}} \int_{-\infty}^{\infty} f(\eta_j)\eta_j^k e^{\frac{-1}{2\sigma_\eta^2}\eta_j^2} d\eta_j \\ &= (2\pi\sigma_\eta^2)^{\frac{-1}{2}} \int_{-\infty}^{\infty} \eta_j^k e^{h(\eta_j)} d\eta_j \end{aligned} \tag{B.39}$$

where the second line follows by definition of h. Now Taylor expand h around $\eta_j = 0$ using the Mean Value Theorem remainder, so

$$h(\eta_j) = h(0) + \underbrace{h'(0)\cdot\eta_j + \frac{1}{2}h''(0)\cdot\eta_j^2}_{:=\mathrm{Q}(\eta_j)} + \underbrace{\frac{1}{6}h^{(3)}(c_j)\eta_j^3}_{:=\mathrm{S}(\eta_j)}$$

for some c_j bounded between 0 and η_j. It is important to note that c_j is a random variable because one of its bounds is random. Finally Taylor expand Q around its maximizer $\hat{\eta}$ (or complete the square) to see the factorization

$$Q(\eta_j) = Q(\hat{\eta}) + \frac{1}{2}Q''(\hat{\eta}) \cdot (\eta_j - \hat{\eta})^2,$$

so we have

$$h(\eta_j) = h(0) + Q(\hat{\eta}) + \frac{1}{2}Q''(\hat{\eta}) \cdot (\eta_j - \hat{\eta})^2 + S(\eta_j). \tag{B.40}$$

Plugging (B.40) into (B.39), we get

$$\begin{aligned} E_{0,\sigma_\eta^2}\left(f(\eta_j)\eta_j^k\right) &= (2\pi\sigma_\eta^2)^{\frac{-1}{2}} f(0) e^{Q(\hat{\eta})} \int_{-\infty}^{\infty} \eta_j^k e^{S(\eta_j)} e^{\frac{1}{2}Q''(\hat{\eta})\cdot(\eta_j-\hat{\eta})^2} d\eta_j \\ &= (2\pi\sigma_\eta^2)^{\frac{-1}{2}} f(0) e^{Q(\hat{\eta})} \left(2\pi\frac{-1}{Q''(\hat{\eta})}\right)^{\frac{1}{2}} E_{\hat{\eta},v}\left(\eta_j^k e^{S(\eta_j)}\right) \\ &= \frac{v^{\frac{1}{2}}}{\sigma_\eta} f(0) e^{Q(\hat{\eta})} E_{\hat{\eta},v}\left(\eta_j^k e^{S(\eta_j)}\right) \end{aligned} \tag{B.41}$$

where $v := \frac{-1}{Q''(\hat{\eta})}$ (note that $Q''(\hat{\eta})$ is negative because $\hat{\eta}$ is the maximizer of Q), giving us nearly the full result. To evaluate $E_{\hat{\eta},v}\left(\eta_j^k e^{S(\eta_j)}\right)$, note that

$$
\begin{aligned}
E_{\hat{\eta},v}\left(\eta_j^k e^{S(\eta_j)}\right) &= E_{\hat{\eta},v}\left(\eta_j^k\left[1+S(\eta_j)+\frac{1}{2}S^2(\eta_j)+\cdots\right]\right) && \text{, by expanding } e^{S(\eta_j)} \\
&= E_{\hat{\eta},v}\left(\eta_j^k\left[1+O(\eta_j^3)\right]\right) && \text{, as } \eta_j^2 \to 0 \text{ by def. of } S \text{ and } (*) \\
&= E_{\hat{\eta},v}\left(\eta_j^k\right) + E_{\hat{\eta},v}\left(O(\eta_j^{k+3})\right) && \text{, by def. of big } O \\
&= E_{\hat{\eta},v}\left(\eta_j^k\right)\cdot\left[1+\frac{E_{\hat{\eta},v}\left(O(\eta_j^{k+3})\right)}{E_{\hat{\eta},v}\left(\eta_j^k\right)}\right] \\
&= E_{\hat{\eta},v}\left(\eta_j^k\right)\cdot\left[1+\frac{O(E_{\hat{\eta},v}\left(\eta_j^{k+3}\right))}{E_{\hat{\eta},v}\left(\eta_j^k\right)}\right] && \text{, by def. of big } O \text{ as } \sigma_\eta^2 \to 0 \\
&= E_{\hat{\eta},v}\left(\eta_j^k\right)\cdot\left[1+O\left(\frac{E_{\hat{\eta},v}\left(\eta_j^{k+3}\right)}{E_{\hat{\eta},v}\left(\eta_j^k\right)}\right)\right] && \text{, by def. of big } O
\end{aligned}
$$

so approximately we can just evaluate a familiar normal moment for this expectation. By (*) we note also that $h^{(3)}$ is bounded so the value of c_j does not affect the order relation. Plugging this into (B.41) concludes the proof. □

For the cases we consider for f, we have $\hat{\eta} \propto v$ which causes the order term in (B.38) to be $O(v)$ when k is odd and $O(v^2)$ when k is even. This can be seen by evaluating the normal expectations, which will be polynomials in v, and then taking the difference between the order of the numerator and the order of the denominator (which just depends on whether k is even or odd; consider the cases $k = 1$ and $k = 2$) – recall that big O notation allows comparisons of this kind.

B.4.2 Specific Cases

When $f(\eta_j) = {\lambda_j}^{*^p} := \left[\frac{{\sigma_b^*}^2}{{\sigma_b^*}^2+\mu_{\tilde{V}}e^{-\eta_j}}\right]^p$ for $p = 1, 2$, then $f(0) = {\tilde{\lambda}}^{*^p} := \left[\frac{{\sigma_b^*}^2}{{\sigma_b^*}^2+\mu_{\tilde{V}}}\right]^p$ and we find that $v := v_p^* = \frac{\sigma_\eta^2}{1+p\sigma_\eta^2\tilde{\lambda}^*(1-\tilde{\lambda}^*)}$ (corresponding to the v_1^* and v_2^* elsewhere), $\hat{\eta} = v_p^* p(1-\tilde{\lambda}^*)$, and $h^{'}(0) = p(1-\tilde{\lambda}^*)$.

Therefore, a first-order Laplace approximation of interest is

$$E(\eta^q \tilde{\lambda}^{*p}) \approx \frac{\tilde{\lambda}^{*p} e^{\frac{1}{2}p^2 v_p^*(1-\tilde{\lambda}^*)^2}}{[1+p\sigma_\eta^2 \tilde{\lambda}^*(1-\tilde{\lambda}^*)]^{\frac{1}{2}}} E_{N(\hat{\eta},v_p^*)}(\eta^q) \tag{B.42}$$

Specific cases we need evaluate to

$$E(\lambda_j^*) \approx \frac{\tilde{\lambda}^* e^{\frac{1}{2}v_1^*(1-\tilde{\lambda}^*)^2}}{[1+\sigma_\eta^2\tilde{\lambda}^*(1-\tilde{\lambda}^*)]^{\frac{1}{2}}}, \quad \text{with } v_1^* = \frac{\sigma_\eta^2}{1+\sigma_\eta^2\tilde{\lambda}^*(1-\tilde{\lambda}^*)}$$
$$E({\lambda_j^*}^2) \approx \frac{\tilde{\lambda}^* e^{2v_2^*(1-\tilde{\lambda}^*)^2}}{[1+2\sigma_\eta^2\tilde{\lambda}^*(1-\tilde{\lambda}^*)]^{\frac{1}{2}}}, \quad \text{with } v_2^* = \frac{\sigma_\eta^2}{1+2\sigma_\eta^2\tilde{\lambda}^*(1-\tilde{\lambda}^*)}$$

and

$$E(\lambda_j^*\eta_j) \approx E(\lambda_j^*)v_1^*(1-\tilde{\lambda}^*)$$
$$E(\lambda_j^*\eta_j^2) \approx E(\lambda_j^*)v_1^*[1+v_1^*(1-\tilde{\lambda}^*)^2]$$
$$E({\lambda_j^*}^2\eta_j) \approx 2E({\lambda_j^*}^2)v_2^*(1-\tilde{\lambda}^*)$$
$$E({\lambda_j^*}^2\eta_j^2) \approx E({\lambda_j^*}^2)[4{v_2^*}^2(1-\tilde{\lambda}^*)^2+v_2^*].$$

We are also interested in evaluating$E(\lambda_j^{*2}e^{-\eta_j})$, the case of (D.1) where $f(\eta_j) = \lambda_j^{*2}e^{-\eta_j}$ with $v = v_2^*, \hat{\eta} = v_2^*(1-2\tilde{\lambda}^*),\ h^{(1)}(0) = (1-2\tilde{\lambda}^*)$. Therefore we have

$$E({\lambda_j^*}^2 e^{-\eta_j}) \approx \frac{{\tilde{\lambda}^*}^2 e^{\frac{1}{2}v_2^*(1-2\tilde{\lambda}^*)^2}}{[1+2\tilde{\lambda}^*(1-\tilde{\lambda}^*)]^{\frac{1}{2}}}.$$

The results used for the FIRC+ derivations are identical, except for replacing $\tilde{\lambda}^*$ with $\tilde{\lambda}^+$ and ${\sigma_b^*}^2$ with $\sigma_{b|\eta}^2$.

B.5 Simulation Scenario Details

B.5.1 Normal Case

In the simple normal case with linear conditional expectation, which is the FIRC+ model, we let

$$\beta_j|\eta_j \overset{\perp}{\sim} N(\beta + \alpha\eta_j, \sigma^2_{b|\eta}) \tag{B.43}$$

where recall that $\alpha = \sigma_{\eta b}/\sigma^2_\eta$ and $\sigma^2_{b|\eta} = \sigma^2_b(1-\rho^2_{\eta b})$.

B.5.2 Heavy-tailed Case

In the heavy-tailed case we let

$$\beta_j = \beta + \alpha\eta_j + \kappa_j \tag{B.44}$$

where $\kappa_j \overset{iid}{\sim} t_3$ (t-distributed with 3 degrees of freedom) independently of η_j (we take $\kappa = \left(\frac{1}{3}\sigma^2_b(1-\rho^2_{\eta b})\right)^{1/2}$ so β_j still has marginal variance σ^2_b). Thus marginally β_j has heavy tails (mirroring the t_3 component more than the normal η_j), meaning that more sites are especially effective or ineffective (compared to the average site) than in the normal model. To illustrate, see Figure B.1.

B.5.3 Nonlinear Case

In this case we make β_j conditionally normal with a quadratic mean function:

$$\beta_j|\eta_j \overset{\perp}{\sim} N(\beta + \alpha\eta_j + \gamma(\eta_j^2 - \sigma^2_\eta),\ \tau^2) \tag{B.45}$$

where we take $\gamma = 0.7\sigma_b/\sigma^2_\eta * \sqrt{\frac{1}{2}(1-\rho^2_{\eta b})}$ and $\tau^2 = \sigma^2_b(1-\rho^2_{\eta b}) - 2\gamma^2\sigma^4_\eta$ so again β_j has marginal variance σ^2_b. So as σ_b increases and σ^2_η and ${\rho_{\eta b}}^2$ decrease the conditional mean will be more nonlinear and the conditional variance will become accordingly smaller. This

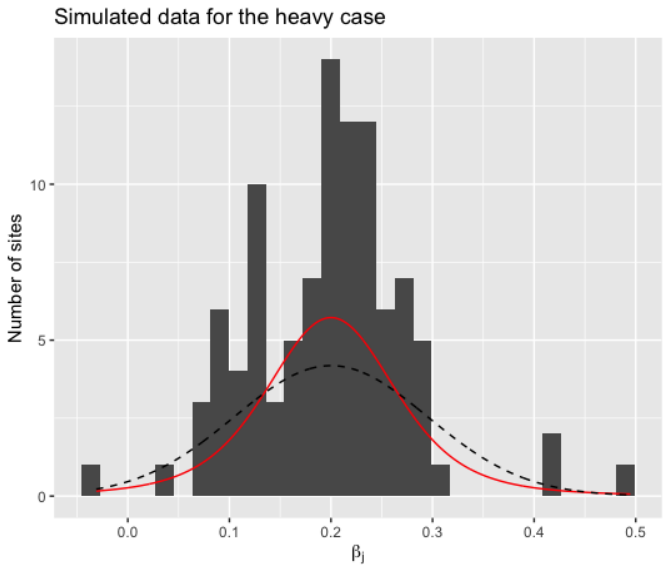

Figure B.1: Illustration of the heavy-tailed case. The vertical bars show a histogram of the true β_j's in a randomly sampled trial of 100 sites. The plotted curves are densities: the red one is the true marginal distribution of the β_j in this scenario, and the dotted black one is normal marginal distribution for the corresponding normal scenario. Although it is hard to see, the red curve does indeed have heavier tails for extreme values for β_j (but since these distributions have identical variance it is not apparent as quickly). In this case, we had $\beta = 0.2$, $\sigma_b = 0.1$, $\rho_{\eta b} = 0.3$, and $\sigma_\eta^2 = 0.6$.

model says that sites with extreme log precisions (low and high) are more effective than sites with average log precisions. This exceptional circumstance could happen in practice, for example, if both small and large sites have different relative advantages over medium-sized sites (suppose the smallest sites have an easier time tailoring treatment to a relatively homogeneous clientele and the largest sites tend to be better funded or more professionally run). However, in many studies is probably not an appropriate default assumption. This type of nonlinearity is uniquely challenging for FIRC+ since compared to parabolas most other nonlinear functions are better approximated by a line. To illustrate, see Figure B.2.

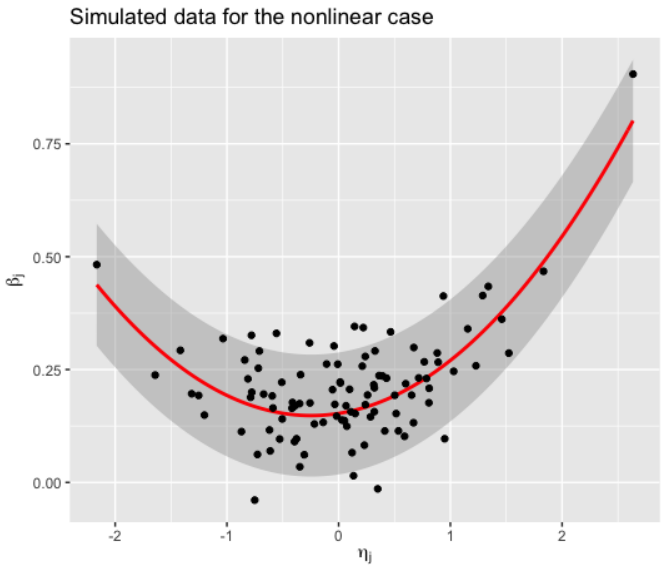

Figure B.2: Illustration of the nonlinear case. Shown is a scatterplot of the true η_j and β_j's in a randomly sampled trial of 100 sites. The red curve shows the true conditional mean function of β_j given η_j, and the grey bands show the 90% central interval of this conditional distribution for each value of η_j. In this case, we had $\beta = 0.2$, $\sigma_b = 0.1$, $\rho_{\eta b} = 0.3$, and $\sigma_\eta^2 = 0.6$.

B.5.4 Heteroscedastic Case

Finally, in the heteroscedastic case

$$\beta_j | \eta_j \overset{\perp}{\sim} N(\beta + \alpha\eta_j,\ \tau^2 e^{-\frac{1}{2}\eta_j}) \tag{B.46}$$

where $\tau^2 = \sigma_b^2(1 - \rho_{\eta b}^2)e^{-\frac{1}{2}\sigma_\eta^2}$ to make the marginal variance of β_j be σ_b^2. This model says that sites with higher precisions have (exponentially) *less* variable effects. Again, this is a fairly particular situation but it could happen if larger sites have more standardized interventions or are less different from one another in their subject populations (compared to the smaller sites). To illustrate, see Figure B.3

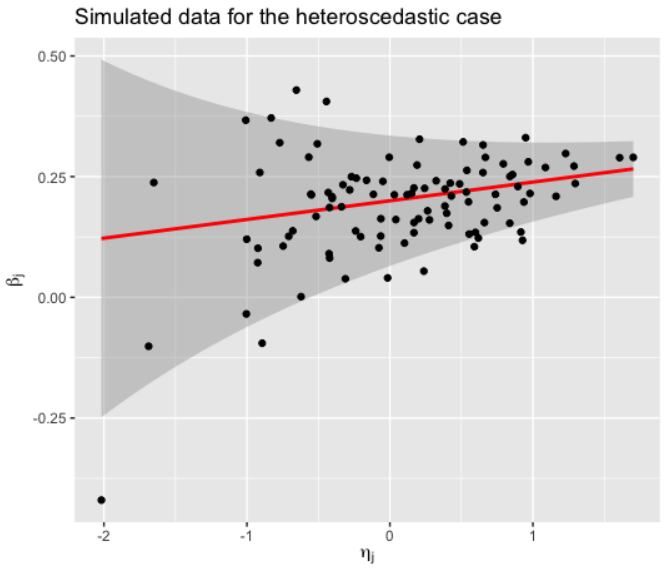

Figure B.3: Illustration of the heteroscedastic case. Shown is a scatterplot of the true η_j and β_j's in a randomly sampled trial of 100 sites. The red curve shows the true conditional mean function of β_j given η_j, and the grey bands show the 90% central interval of this conditional distribution for each value of η_j. In this case, we had $\beta = 0.2$, $\sigma_b = 0.1$, $\rho_{\eta b} = 0.3$, and $\sigma_\eta^2 = 0.6$.

B.6 Additional Figures

This section includes figures with further details of the asymptotic and simulation results.

B.6.1 Asymptotic RMSE, Variance, and Bias

The first three figures here plot the relative RMSE approximations of the estimators for different values of J, and the next two plot the variance ratios (which do not depend on J) and biases (which do not depend on J) respectively. All are based on the expressions given in Table 3.3.1.

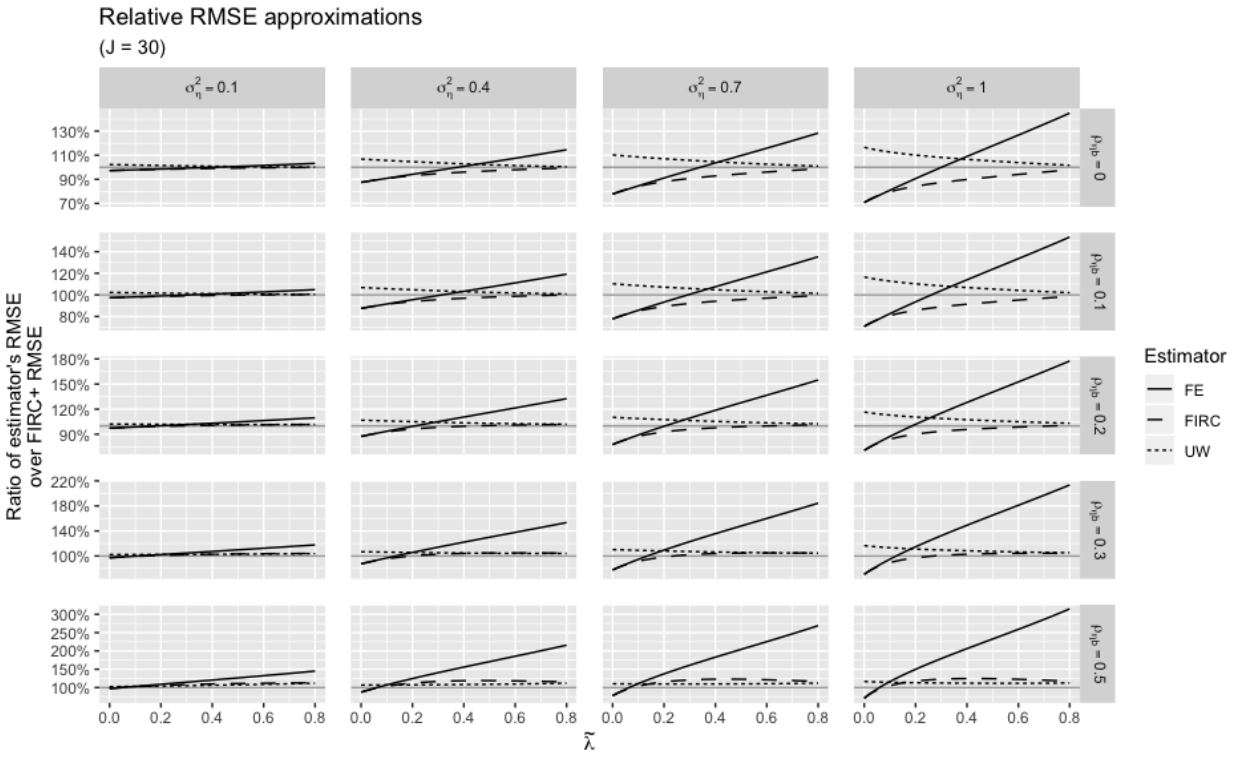

Figure B.4: Relative RMSE approximations when $J = 30$. The ratio of each estimator's RMSE to that of FIRC+ is plotted on the y-axis, as $\tilde{\lambda}$ varies on the x-axis. In the grid of plots the columns let σ_η^2 vary while the rows let $\rho_{\eta b}$ vary. The type of the line indicates the estimator being compared to FIRC+, which itself is noted by the flat, solid grey line at 100% in each plot.

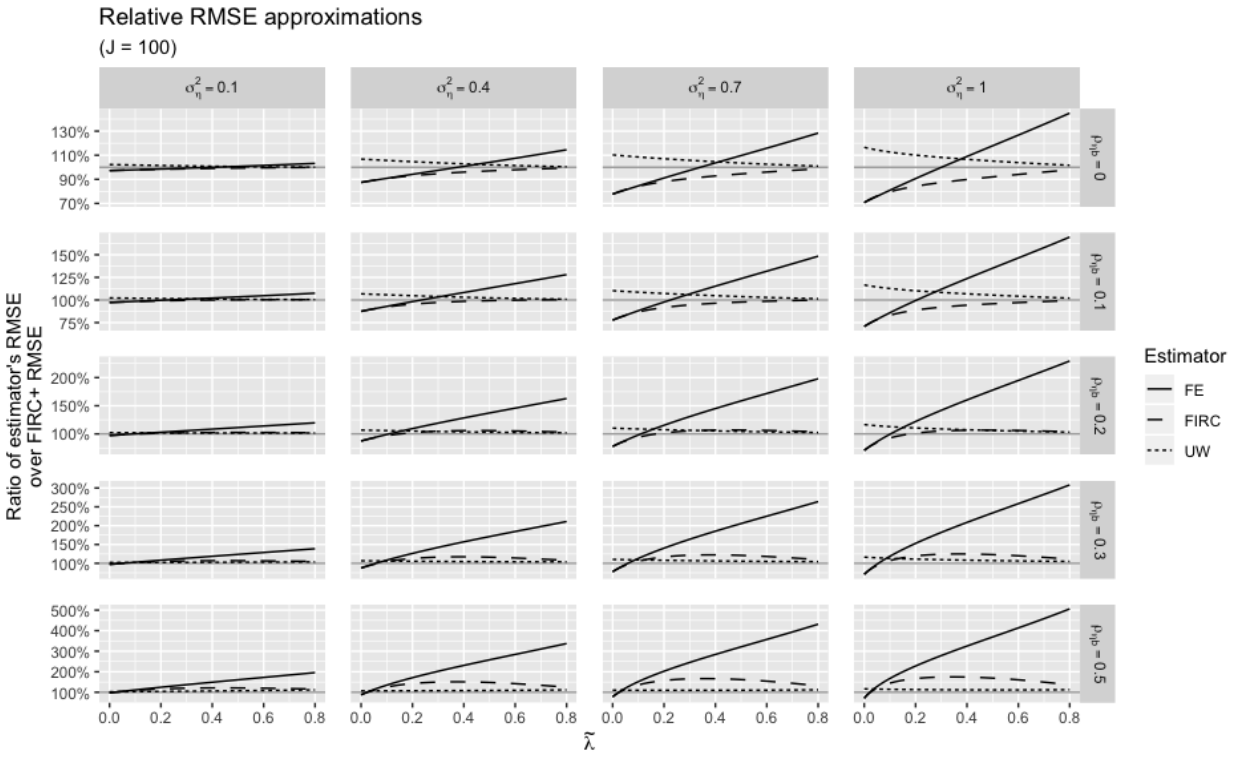

Figure B.5: Relative RMSE approximations when $J = 100$. The ratio of each estimator's RMSE to that of FIRC+ is plotted on the y-axis, as $\tilde{\lambda}$ varies on the x-axis. In the grid of plots the columns let σ_η^2 vary while the rows let $\rho_{\eta b}$ vary. The type of the line indicates the estimator being compared to FIRC+, which itself is noted by the flat, solid grey line at 100% in each plot.

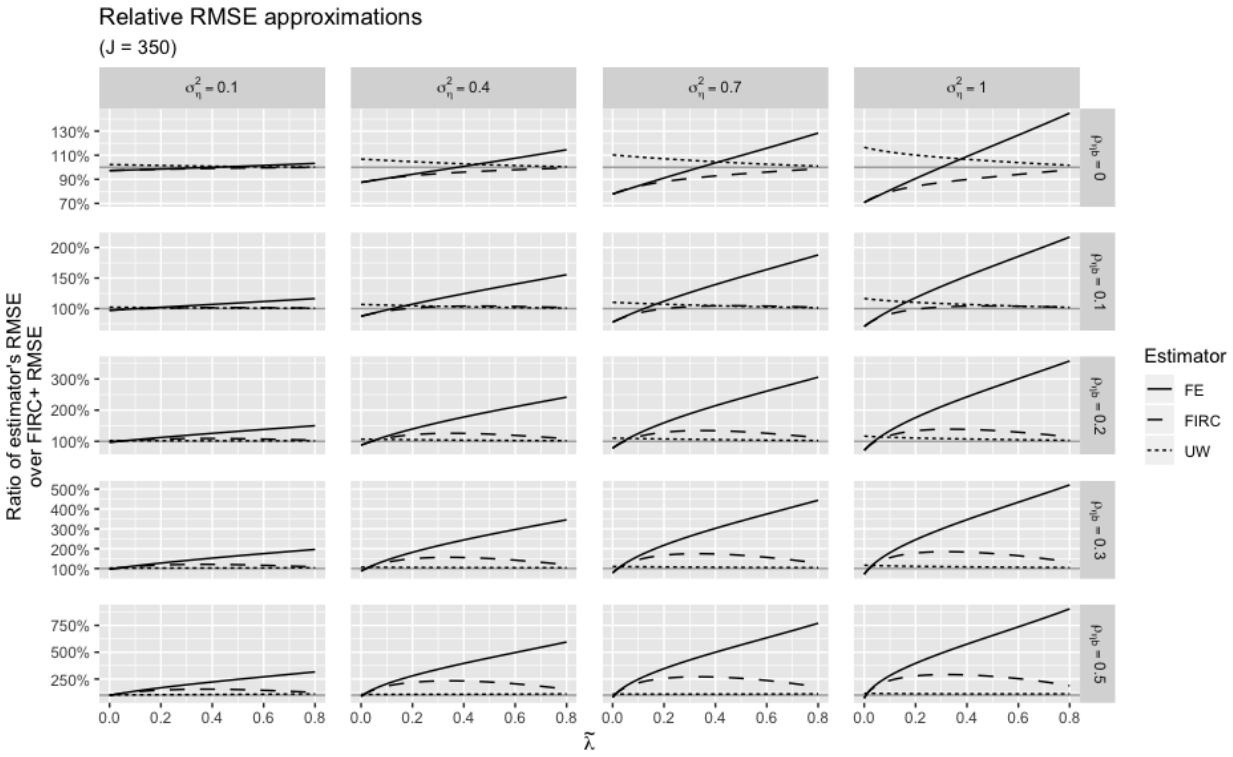

Figure B.6: Relative RMSE approximations when $J = 350$. The ratio of each estimator's RMSE to that of FIRC+ is plotted on the y-axis, as $\tilde{\lambda}$ varies on the x-axis. In the grid of plots the columns let σ_η^2 vary while the rows let $\rho_{\eta b}$ vary. The type of the line indicates the estimator being compared to FIRC+, which itself is noted by the flat, solid grey line at 100% in each plot.

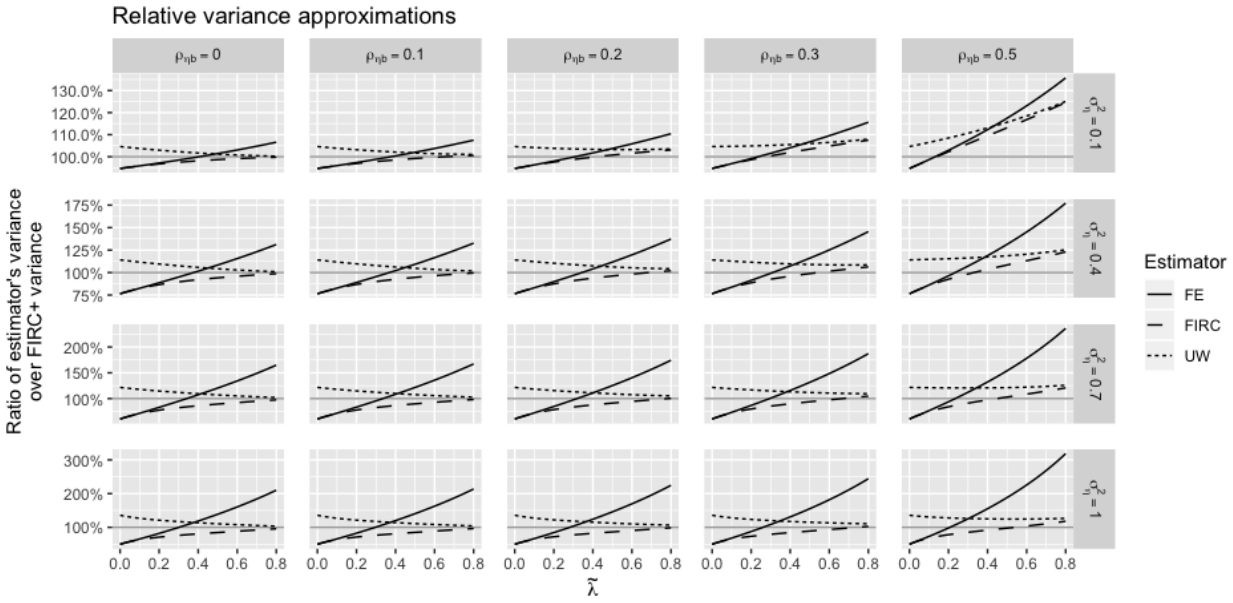

Figure B.7: Relative variance approximations. The ratio of each estimator's asymptotic variance to that of FIRC+ is plotted on the y-axis, as as $\tilde{\lambda}$ varies on the x-axis. In the grid of plots the columns let $\rho_{\eta b}$ vary while the rows let σ^2_η vary (compared to the RMSE and bias figures, this was reversed to accommodate the wide range of ratios as a function of σ^2_η). The type of the line indicates the estimator being compared to FIRC+, which itself is noted by the flat, solid grey line at 100% in each plot.

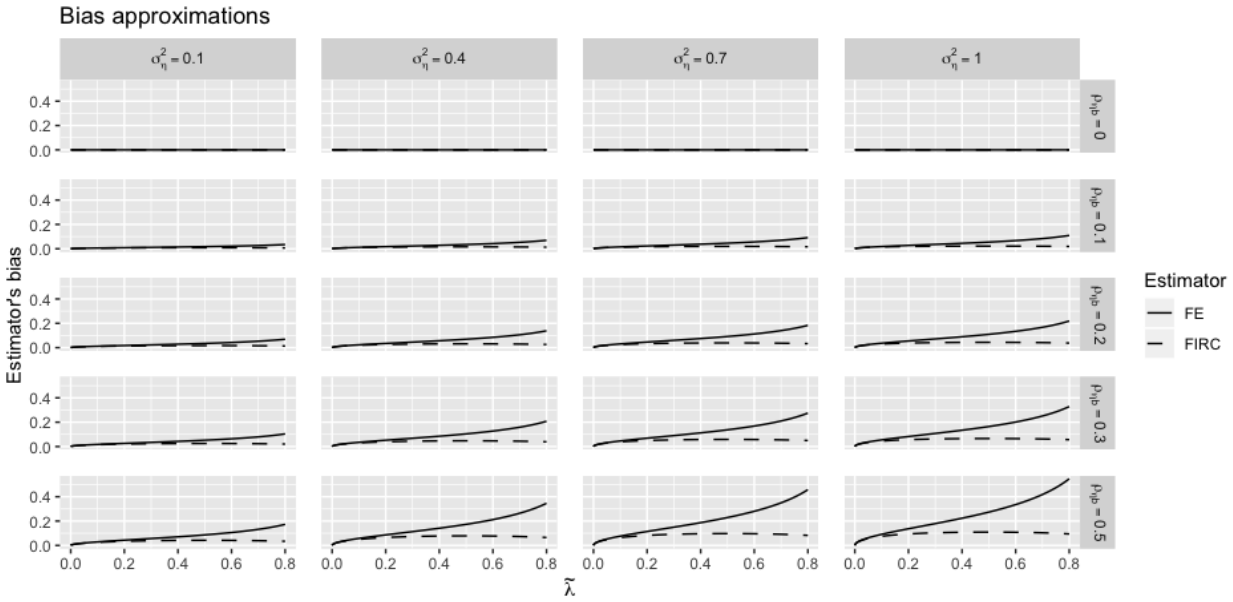

Figure B.8: Bias approximations for FIRC and FE. The asymptotic bias of FIRC and FE is plotted on the y-axis, as $\tilde{\lambda}$ varies on the x-axis. In the grid of plots the columns let σ_η^2 vary while the rows let $\rho_{\eta b}$ vary. FE is denoted by the solid line and FIRC is denoted by the dashed line.

B.6.2 Accuracy of the Asymptotic Approximations

These figures show the accuracy of the analytic RMSE approximations given in Table 3.3.1, which is generally quite good with few exceptions.

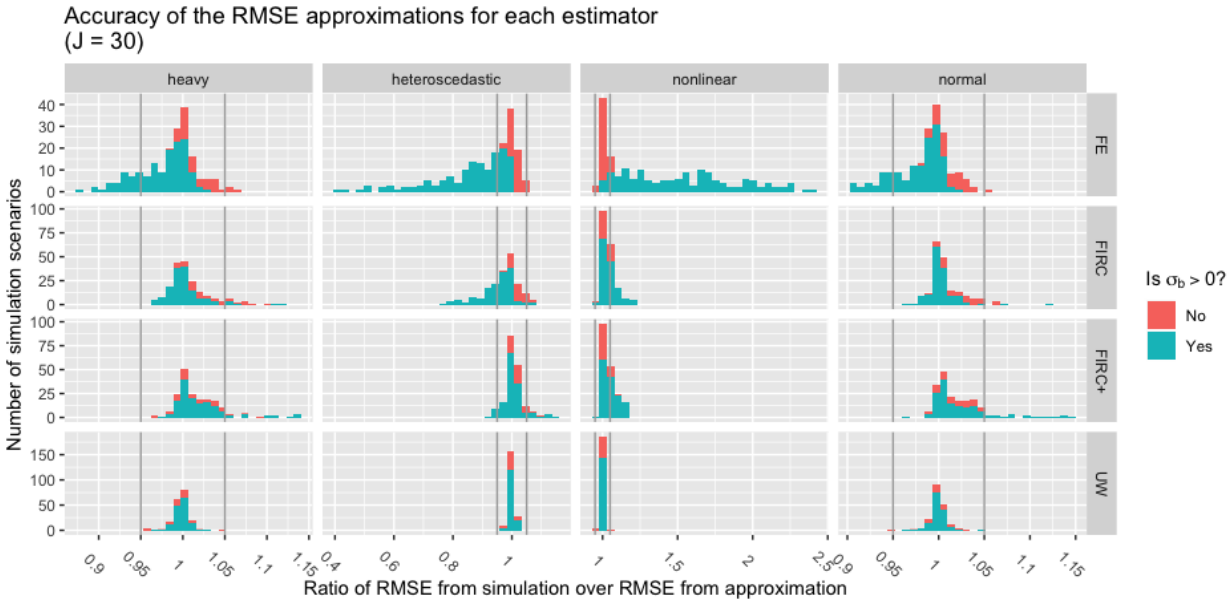

Figure B.9: Accuracy of the RMSE approximations when $J = 30$. Each observation in the histograms is the ratio of simulation RMSE over approximation RMSE for one of the 2,304 scenarios in the simulation study, grouped by distribution type (columns) and estimator (rows). The color of the bars indicates whether $\sigma_b > 0$, for example in the heavy, FE histogram roughly 20 scenarios where $\sigma_b = 0$ have a RMSE ratio near 1 (the teal portion of the tallest bar) while about 15 scenarios where $\sigma_b > 0$ have a RMSE in this range (the red portion of the tallest bar). Note that the x-axis has a different range for each simulation distribution, reflecting the widely varying accuracies in some cases. The gray vertical bars at 0.95 and 1.05 roughly indicate the interval in which the true RMSE would fall (due to approximately unbiased Monte Carlo error in the simulation RMSE), based on the fact that the UW RMSE "approximation" is an exact finite-sample result giving the true RMSE and the simulation RMSE for UW always falls in this range.

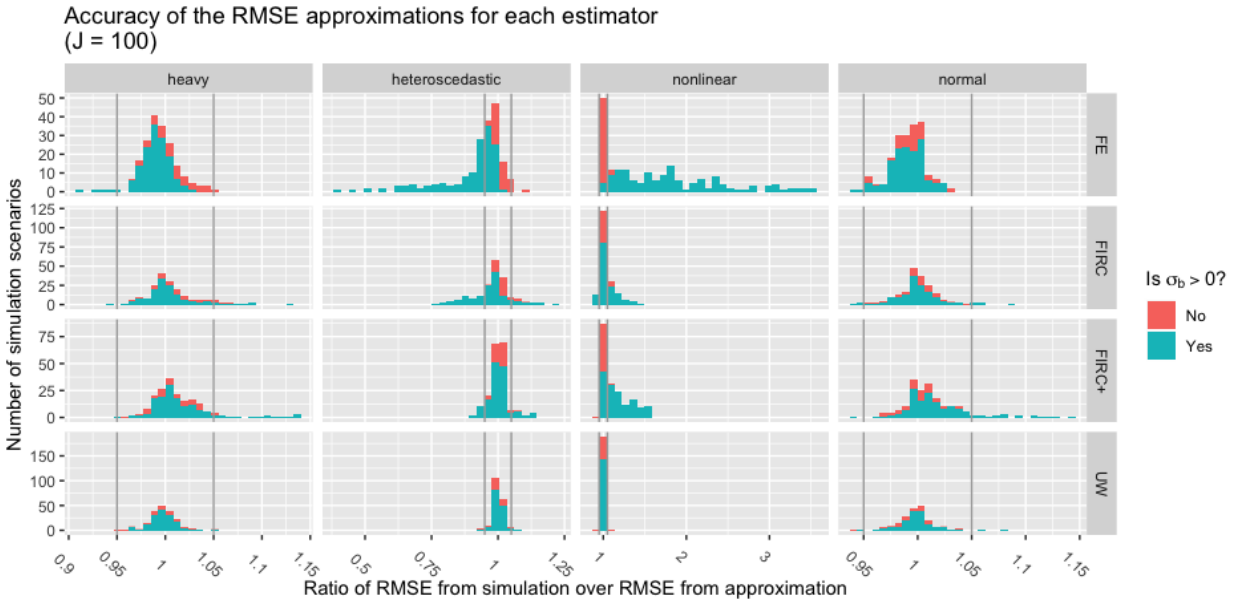

Figure B.10: Accuracy of the RMSE approximations from Table 3.3.1 when $J = 100$. See Figure B.9 caption for full explanation.

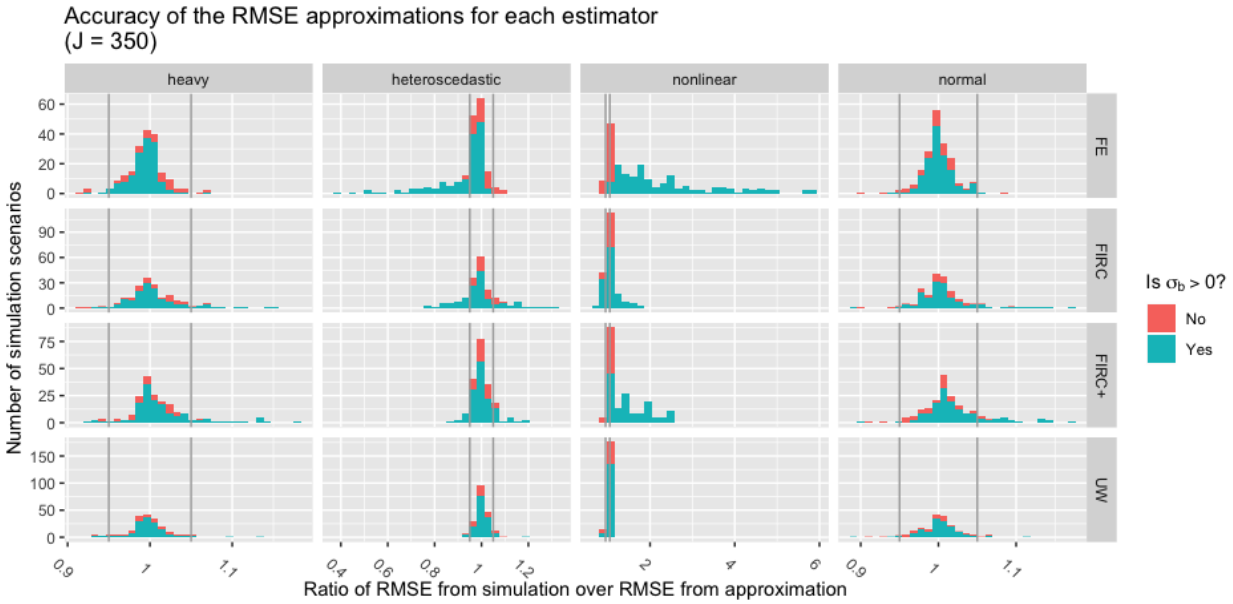

Figure B.11: Accuracy of the RMSE approximations from Table 3.3.1 when $J = 350$. See Figure B.9 caption for full explanation.

B.6.3 Simulation-based RMSE

These figures show the RMSE of the estimators in simulation when $J = 30, 100, 350$ (separate figures) and the conditional distribution of $\beta_j|\eta_j$ is normal, heavy, heteroscedastic, and nonlinear (each large row in the figure grids). They are analogous to Figure 3.4.1 in the text, which only gave the $J = 100$ case for the heteroscedastic and nonlinear scenarios. The normal and heavy cases largely follow the analytic results, as suggested by the approximation accuracy figures in Section B.6.2, while the heteroscedastic and nonlinear cases for $J = 30, 350$ are qualitatively similar to Figure 3.4.1.

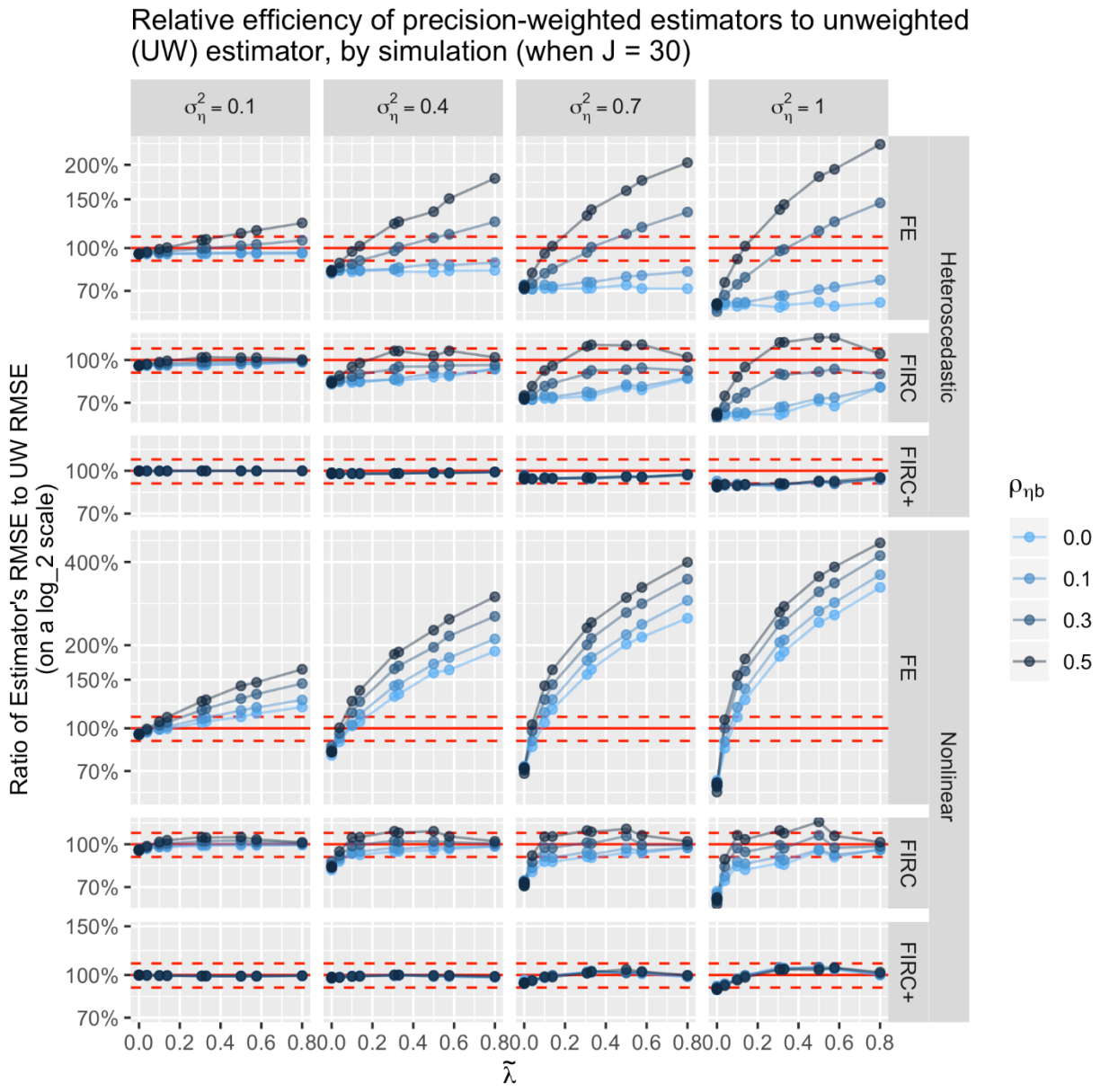

Figure B.12: Scatterplots of the ratio of each estimator's RMSE relative to the RMSE of UW, as $\tilde{\lambda}$ varies on the x-axis, for $J = 30$ in the heteroscedastic and nonlinear cases. Separate plots are given for each σ^2_η(columns) and estimator/distribution (rows). A RMSE ratio of 100% is shown by the horizontal solid red lines, while the dashed red lines are at 110% and 90% to demarcate cases where estimators perform comparably to UW. The color of each point corresponds to the correlation $\rho_{\eta b}$. These ratios only depend on σ^2_b and $\mu_{\tilde{V}}$ (not shown) through $\tilde{\lambda}$.

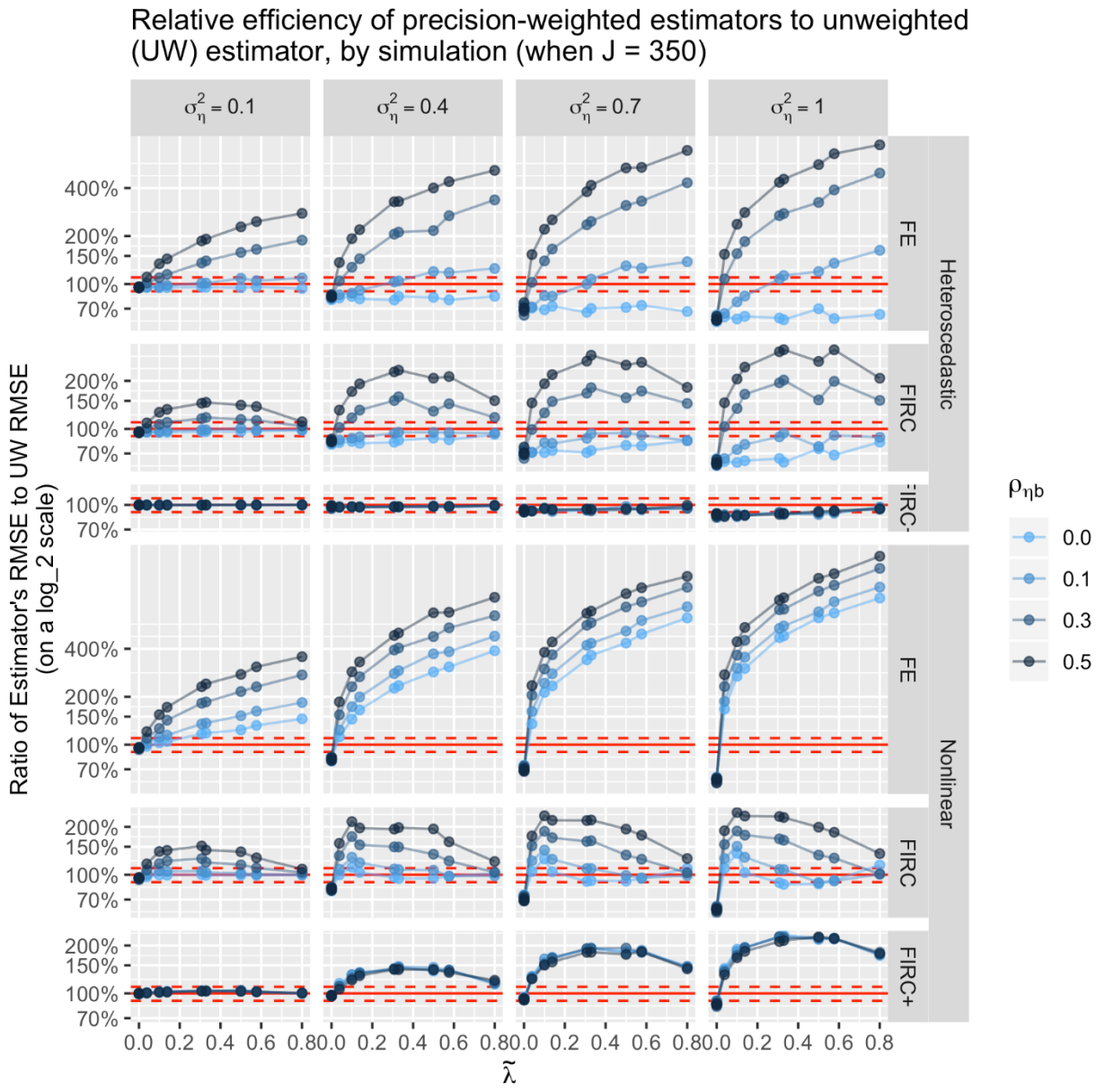

Figure B.13: Scatterplots of the ratio of each estimator's RMSE relative to the RMSE of UW, as $\tilde{\lambda}$ varies on the x-axis, for $J = 350$ in the heteroscedastic and nonlinear cases. Separate plots are given for each σ^2_η(columns) and estimator/distribution (rows). A RMSE ratio of 100% is shown by the horizontal solid red lines, while the dashed red lines are at 110% and 90% to demarcate cases where estimators perform comparably to UW. The color of each point corresponds to the correlation $\rho_{\eta b}$. These ratios only depend on σ^2_b and $\mu_{\tilde{V}}$ (not shown) through $\tilde{\lambda}$.

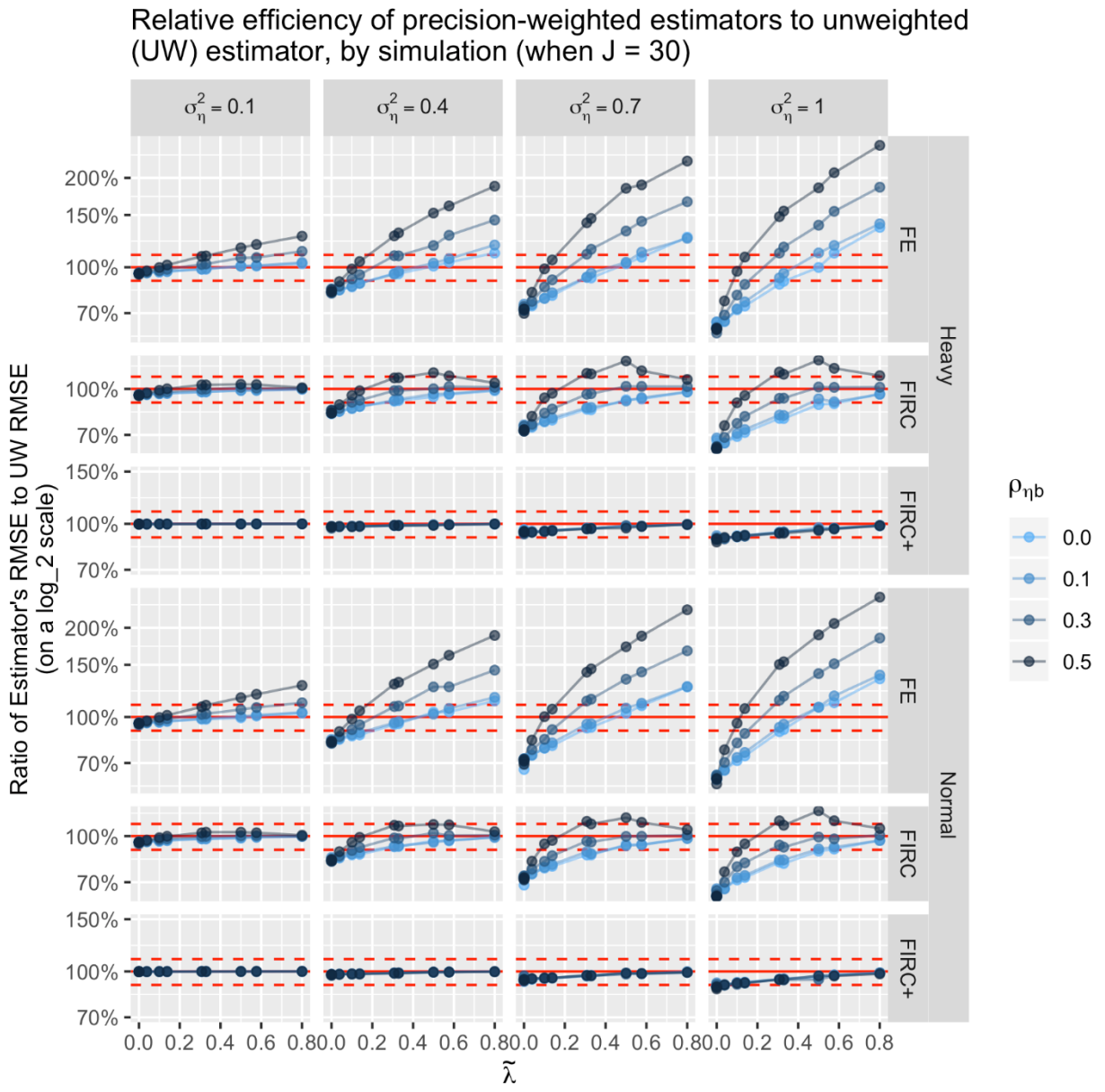

Figure B.14: Scatterplots of the ratio of each estimator's RMSE relative to the RMSE of UW, as $\tilde{\lambda}$ varies on the x-axis, for $J = 30$ in the heavy and normal cases. Separate plots are given for each σ^2_η(columns) and estimator/distribution (rows). A RMSE ratio of 100% is shown by the horizontal solid red lines, while the dashed red lines are at 110% and 90% to demarcate cases where estimators perform comparably to UW. The color of each point corresponds to the correlation $\rho_{\eta b}$. These ratios only depend on σ^2_b and $\mu_{\tilde{V}}$ (not shown) through $\tilde{\lambda}$.

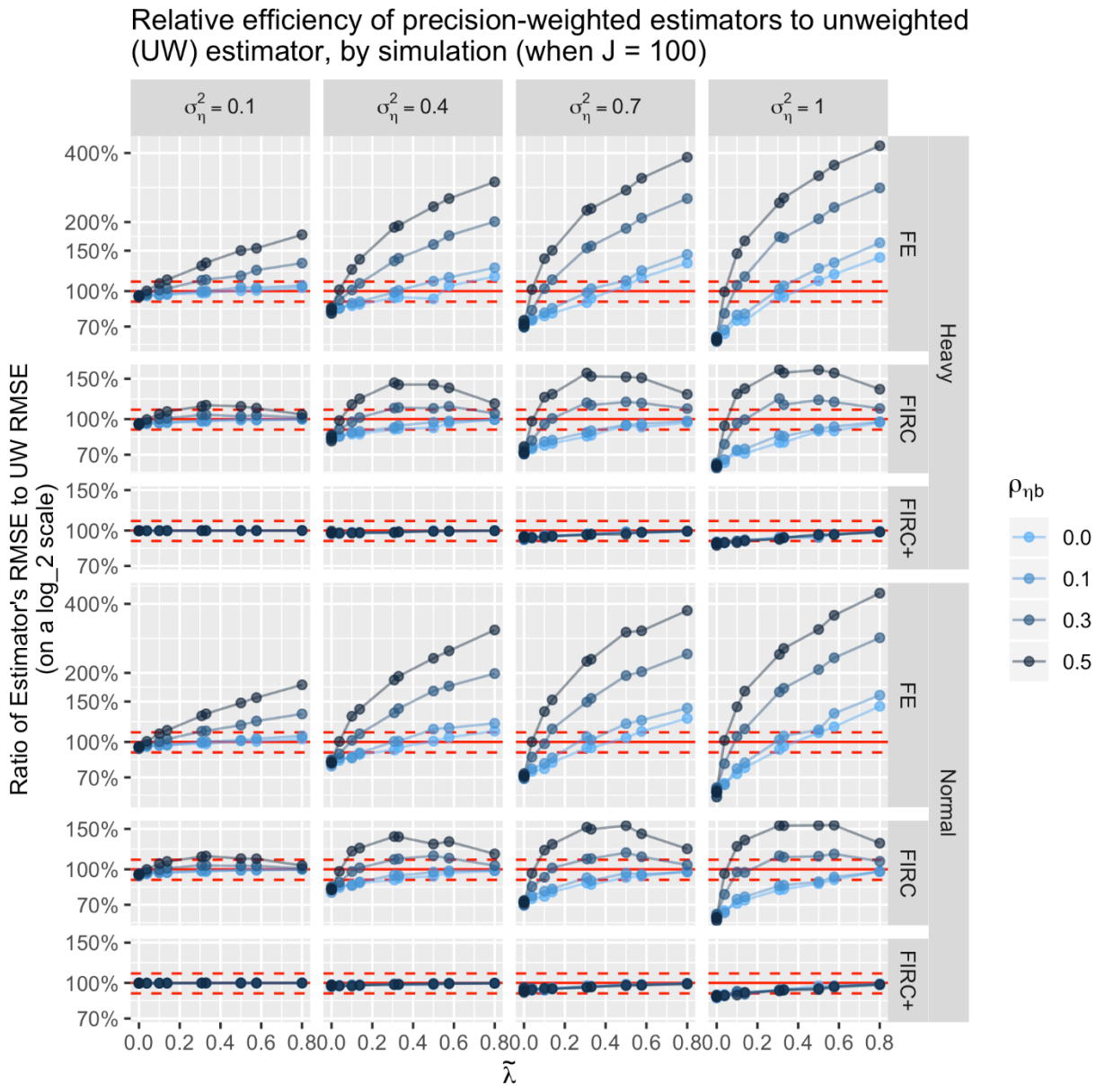

Figure B.15: Scatterplots of the ratio of each estimator's RMSE relative to the RMSE of UW, as $\tilde{\lambda}$ varies on the x-axis, for $J = 100$ in the heavy and normal cases. Separate plots are given for each σ^2_η(columns) and estimator/distribution (rows). A RMSE ratio of 100% is shown by the horizontal solid red lines, while the dashed red lines are at 110% and 90% to demarcate cases where estimators perform comparably to UW. The color of each point corresponds to the correlation $\rho_{\eta b}$. These ratios only depend on σ^2_b and $\mu_{\tilde{V}}$ (not shown) through $\tilde{\lambda}$.

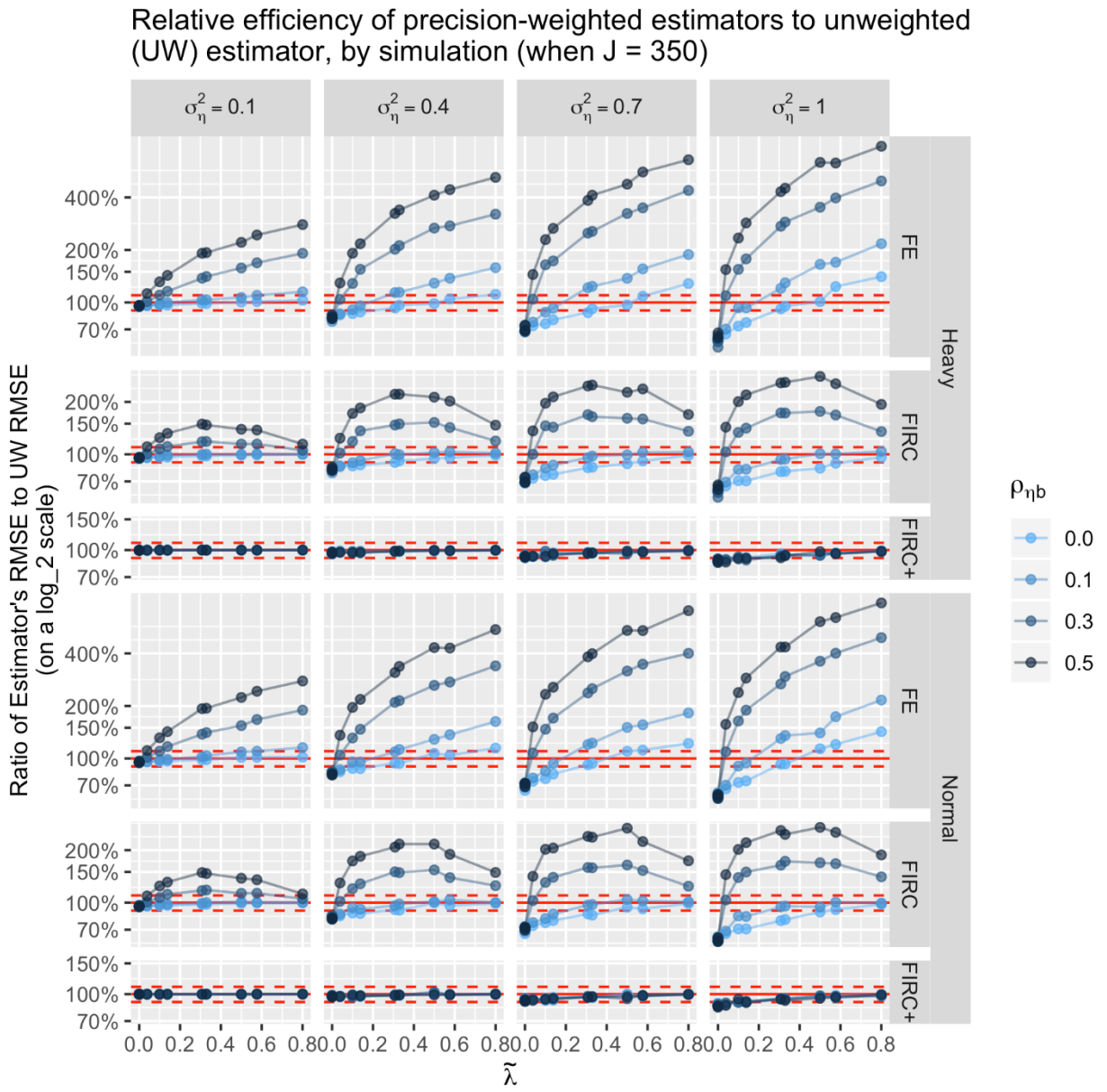

Figure B.16: Scatterplots of the ratio of each estimator's RMSE relative to the RMSE of UW, as $\tilde{\lambda}$ varies on the x-axis, for $J = 350$ in the heavy and normal cases. Separate plots are given for each σ^2_η(columns) and estimator/distribution (rows). A RMSE ratio of 100% is shown by the horizontal solid red lines, while the dashed red lines are at 110% and 90% to demarcate cases where estimators perform comparably to UW. The color of each point corresponds to the correlation $\rho_{\eta b}$. These ratios only depend on σ^2_b and $\mu_{\tilde{V}}$ (not shown) through $\tilde{\lambda}$.

B.6.4 Simulation-based Coverage

These figures simulate coverage of nominal 95% Wald-type confidence intervals for FE and FIRC, analogously to Figure 3.4.2 except for the heavy, heteroscedastic, and nonlinear cases. The basic conclusion is the same: in situations where the estimator has nontrivial bias coverage can quickly become unacceptable.

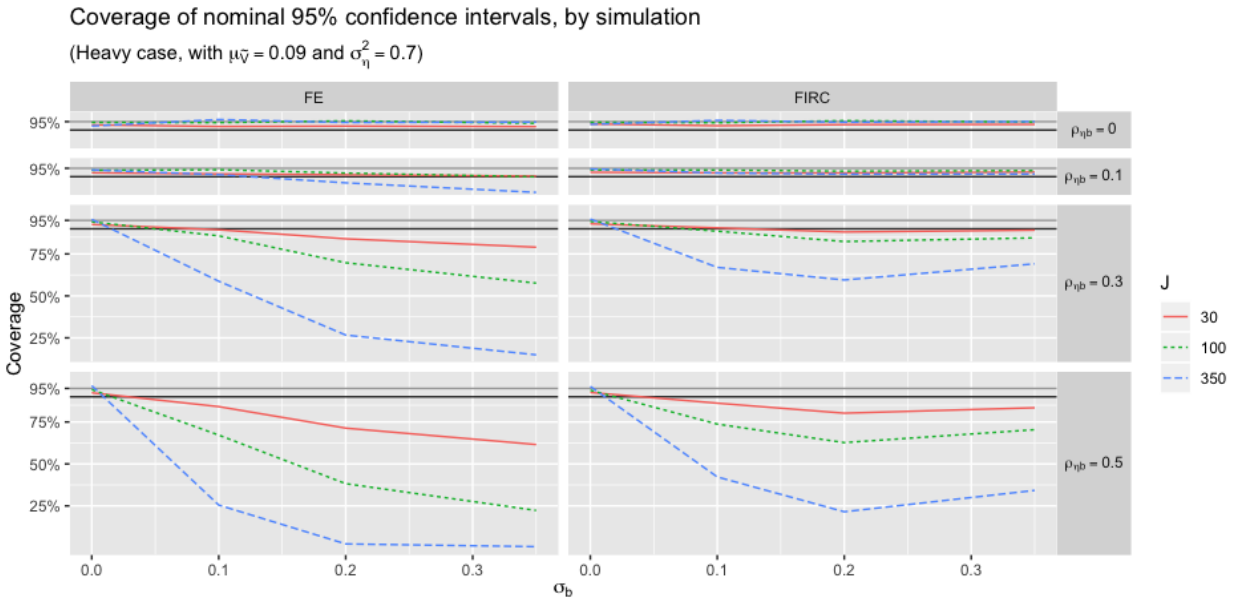

Figure B.17: Actual coverage rates (by simulation) for the standard nominal 95% confidence intervals centered on the different point estimators, as a function of σ_b (x-axis) in the heavy case. Plots are given for FIRC and FE (columns) and values of $\rho_{\eta b}$ (rows). Each colored line shows scenarios with a different sample size$J$. The horizontal light grey lines mark the nominal 95% rate, and the dark grey lines mark 90%. The data where simulated from the normal case (as in the simulation study in Sections 4), with $\mu_{\tilde{V}} = 0.09$ and σ^2_η, though changing these parameters does not change the results qualitatively (i.e. the coverage can still be bad).

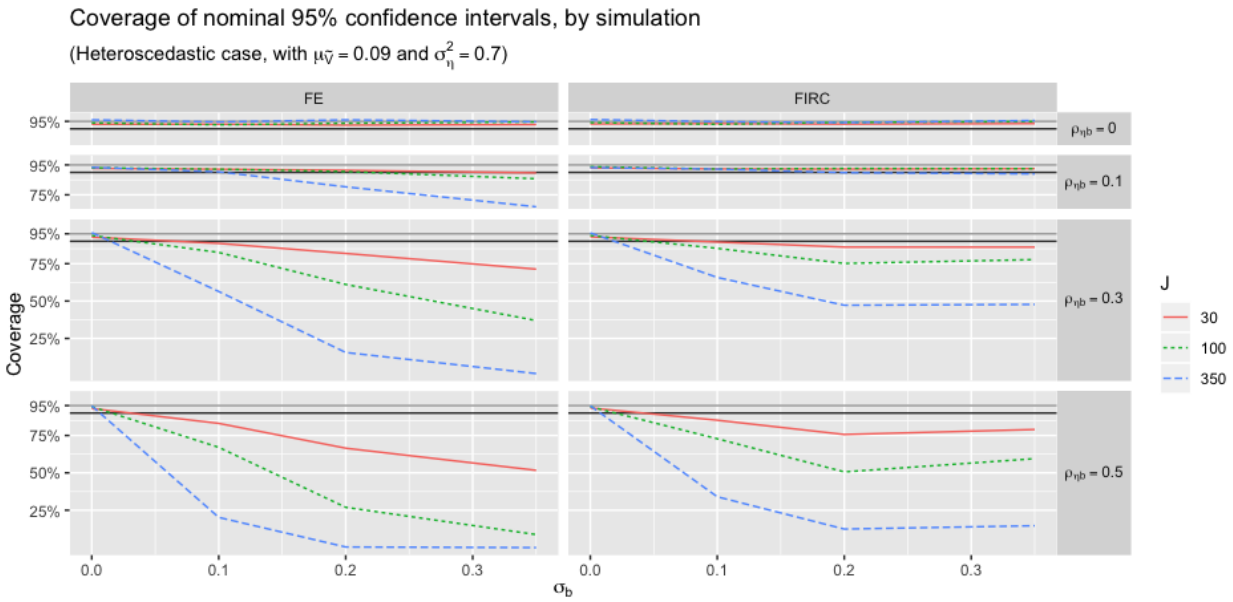

Figure B.18: Actual coverage rates (by simulation) for the standard nominal 95% confidence intervals centered on the different point estimators, as a function of σ_b (x-axis) in the heteroscedastic case. Plots are given for FIRC and FE (columns) and values of $\rho_{\eta b}$ (rows). Each colored line shows scenarios with a different sample size$J$. The horizontal light grey lines mark the nominal 95% rate, and the dark grey lines mark 90%. The data where simulated from the normal case (as in the simulation study in Sections 4), with $\mu_{\tilde{V}} = 0.09$ and σ^2_η, though changing these parameters does not change the results qualitatively (i.e. the coverage can still be bad).

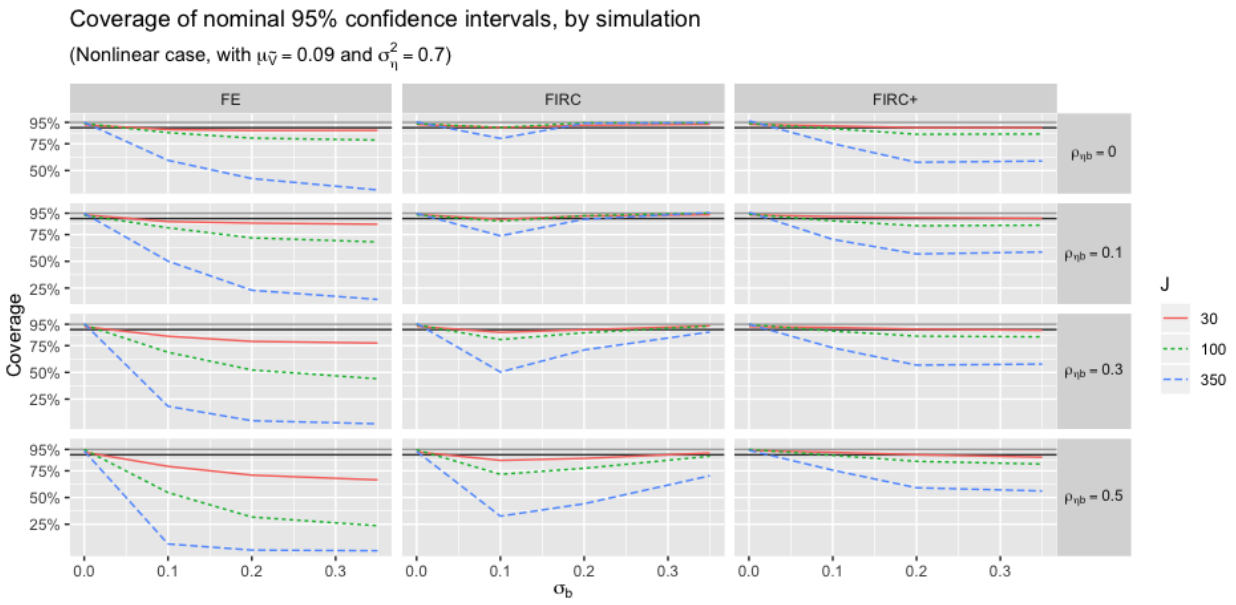

Figure B.19: Actual coverage rates (by simulation) for the standard nominal 95% confidence intervals centered on the different point estimators, as a function of σ_b (x-axis) in the nonlinear case. Plots are given for FIRC and FE (columns) and values of $\rho_{\eta b}$ (rows). Each colored line shows scenarios with a different sample size$J$. The horizontal light grey lines mark the nominal 95% rate, and the dark grey lines mark 90%. The data where simulated from the normal case (as in the simulation study in Sections 4), with $\mu_{\tilde{V}} = 0.09$ and σ^2_η, though changing these parameters does not change the results qualitatively (i.e. the coverage can still be bad).

B.7 Self-efficiency

In the results below, we consider a version of self-efficiency that is more narrow than Meng & Xie's notion in two ways. First, we restrict attention to the case when $\lambda = 0$ (or whatever notation we're using) so self-efficiency reduces to comparing the risks (MSEs) of the complete and subset estimators. This covers the most intuitive case of self-efficiency and simplifies the derivations. And second, we just consider subsets where the most imprecise sites (probably the smallest) are discarded so $S = \{j : \eta_j > \eta_{(J-s)}\}$ (where $\eta_{(J-s)}$ is the $(J-s)$-th order statistic) and the size of the subset is fixed at s and not a random variable. Again, this is the most intuitive way to discard data (since we're trying to find a more efficient estimator) and simplifies the math in some regards. (This does, however, mean that we truncate η by discarding a random region determined by the observed order statistic).

First we find the variance of the unweighted estimator, $\hat{\beta}_{UW}$, and the bias and variance of the subset unweighted estimator, $\hat{\beta}_S$, and then using these results we derive the condition for self-inefficiency of the unweighted estimator.

B.7.1 Variance of the Unweighted Estimator

This derivation is very straightforward:

$$
\begin{aligned}
Var\left(\hat{\beta}_{UW}\right) &= Var\left(\frac{1}{J}\sum_{i=1}^{J}\hat{\beta}_j\right) \\
&= \frac{1}{J^2}\sum_{j=1}^{J} Var\left(\hat{\beta}_j\right) && \text{, by independence} \\
&= \frac{1}{J^2}\sum_{j=1}^{J} Var\left(\beta_j + e_j\right) \\
&= \frac{1}{J^2}\sum_{j=1}^{J} \left[\sigma_b^2 + E(V_j)\right] && \text{, by law of total variance} \\
Var\left(\hat{\beta}_{UW}\right) &= \frac{1}{J}\left[\sigma_b^2 + E(V_j)\right] && \text{, by independence.}
\end{aligned}
$$

To help make this comparable to the subset estimator's variance which we will describe below, we relate the marginal variance of site effects σ_b^2 to the conditional variances of site effects in and out of the subset. Leting $\sigma^2_{b\in S} := Var(\beta_j | j \in S)$ and $\sigma^2_{b\notin S} := Var(\beta_j | j \notin S)$, we can note

$$
\begin{aligned}
\sigma_b^2 &= E\left[Var(\beta_j|\mathbb{1}_j)\right] + Var\left[E(\beta_j|\mathbb{1}_j)\right] \\
&= \pi_s\sigma^2_{b\in S} + (1-\pi_S)\sigma^2_{b\notin S} + Var(\mathbb{1}_j\beta_{\in S} + \left[1-\mathbb{1}_j\right]\beta_{\notin S}) \\
&= \pi_s\sigma^2_{b\in S} + (1-\pi_S)\sigma^2_{b\notin S} + \left[\pi_S\beta^2_{\in S} + (1-\pi_S)\beta^2_{\notin S}\right] \\
&\qquad - \left[\pi_S^2\beta^2_{\in S} + (1-\pi_S)^2\beta^2_{\notin S} + 2\pi_S(1-\pi_S)\beta_{\in S}\beta_{\notin S}\right] \\
&= \pi_s\sigma^2_{b\in S} + (1-\pi_S)\sigma^2_{b\notin S} + \pi_s(1-\pi_S)\left[\beta^2_{\in S} + \beta^2_{\notin S} - 2\beta_{\in S}\beta_{\notin S}\right] \\
\sigma_b^2 &= \pi_s\sigma^2_{b\in S} + (1-\pi_S)\sigma^2_{b\notin S} + \pi_s(1-\pi_S)\Delta_S^2,
\end{aligned}
$$

where on the first line we use the law of total variance, on the second line we use the tower

law, and on the third line we use the fact that $Var(X) = E(X^2) - E^2(X)$.

B.7.2 Bias and Variance of the Subset Estimator

The key issue in characterizing the bias and variance of $\hat{\beta}_S$ is that we select the random subset $S \subset \{1, \ldots, J\}$ as a function of the random vector $(\eta_1, \ldots, \eta_J)$ so the distributions of β_j in selected and unselected sites may differ if $(\beta_1, \ldots, \beta_J)$ is not independent of $(\eta_1, \ldots, \eta_J)$. In addition, the distribution of the noise $e_j = \hat{\beta}_j - \beta_j$ will differ between selected and unselected sites (since we select on η_j and $Var(e_j|\eta_j) = V_j = \tilde{\mu}_V e^{-\eta_j}$).

We will discuss subsets of fixed size s chosen using the order statistic $\eta_{(J-s)}$ to keep only the sites with the s most precise effect estimates, so we may write $S := \{j \in \{1, \ldots, J\} : \eta_j \geq \eta_{(J-s)}\}$. Our model almost surely excludes the possibility of ties, since the distribution of the transformed precisions is continuous, despite that in practice they are technically possible (in sites with the same size n_j and proportion treated $\bar{T}_j$). This is conceptually intuitive in practice, and the fixed size of S simplifies the following derivations.

To describe the subset estimator's bias, let $\beta_{\in S} := E(\beta_j | j \in S)$, $\beta_{\notin S} := E(\beta_j | j \notin S)$, and $\Delta_S := \beta_{\in S} - \beta_{\notin S}$ so by the tower law we have $\beta := E(\beta_j) = \pi_S \beta_{\in S} + (1 - \pi_S)\beta_{\notin S} = \beta_{\notin S} + \pi_S \Delta_S$. To make the equations below more concise, we adopt the notational shorthand for inclusion indicators that $\mathbb{1}_j := \mathbb{1}\{j \in S\}$ for any $j \in \{1, \ldots, J\}$. We emphasize that these indicators are random variables whose randomness is a result of the randomness of $(\eta_1, \ldots, \eta_J)$.

Then

$$
\begin{aligned}
Bias\left(\hat{\beta}_S, \beta\right) &= E\left(\hat{\beta}_S\right) - \beta \\
&= E\left(s^{-1}\sum_{j=1}^{J}\hat{\beta}_j \mathbb{1}_j\right) - \beta && \text{, by def'n of } \hat{\beta}_S \\
&= \pi_S^{-1} E\left(\hat{\beta}_j \mathbb{1}_j\right) - \beta && \text{, by independence} \\
&= \pi_S^{-1} E\left[\mathbb{1}_j E\left(\hat{\beta}_j | \mathbb{1}_j\right)\right] - \beta && \text{, by tower law} \\
&= \beta_{\in S} - \left[\beta_{\notin S} + \pi_S \Delta_S\right] && \text{, by (*) below} \\
Bias(\hat{\beta}_S, \beta) &= (1 - \pi_S)\Delta_S.
\end{aligned}
$$

Now for the variance observe that

$$
\begin{aligned}
Var\left(\hat{\beta}_S\right) &= Var\left(\frac{1}{s}\sum_{j=1}^{J}\hat{\beta}_j \mathbb{1}_j\right) \\
&= \frac{1}{s^2}\sum_{j=1}^{J} Var\left(\hat{\beta}_j \mathbb{1}_j\right) + \frac{1}{s^2}\sum_{j\neq k} Cov\left(\hat{\beta}_j \mathbb{1}_j, \hat{\beta}_k \mathbb{1}_k\right) \\
&= \frac{J}{s^2} Var\left(\hat{\beta}_j \mathbb{1}_j\right) + \frac{J(J-1)}{s^2} Cov\left(\hat{\beta}_j \mathbb{1}_j, \hat{\beta}_k \mathbb{1}_k\right) && \text{, by independence}
\end{aligned}
$$

where the covariance term appears because sites are selected for the subset without replacement (so the indicators $\mathbb{1}_j$ and $\mathbb{1}_k$ are dependent).

To simplify $Var\left(\hat{\beta}_S\right)$ further we must evaluate $Var\left(\hat{\beta}_j \mathbb{1}_j\right)$ and $Cov\left(\hat{\beta}_j \mathbb{1}_j, \hat{\beta}_k \mathbb{1}_k\right)$

above, and to do so we will rely on the useful identities

$$
\begin{aligned}
E\left(\hat{\beta}_j|\mathbb{1}_j\right) &= \mathbb{1}_j\beta_{\in S} + (1-\mathbb{1}_j)\beta_{\notin S} \\
Var\left(\hat{\beta}_j|\mathbb{1}_j\right) &= \mathbb{1}_j Var\left(\hat{\beta}_j|j\in S\right) + \left(1-\mathbb{1}_j\right) Var\left(\hat{\beta}_j|j\notin S\right) \\
Cov\left(\hat{\beta}_j,\hat{\beta}_k|\mathbb{1}_j,\mathbb{1}_k\right) &= \mathbb{1}_j\mathbb{1}_k Cov\left(\hat{\beta}_j,\hat{\beta}_k|j\in S,k\in S\right) \qquad (*) \\
&\quad + (1-\mathbb{1}_j)(1-\mathbb{1}_k)Cov\left(\hat{\beta}_j,\hat{\beta}_k|j\notin S,k\notin S\right) \\
&\quad + 2\mathbb{1}_j(1-\mathbb{1}_k)Cov\left(\hat{\beta}_j,\hat{\beta}_k|j\in S,k\notin S\right).
\end{aligned}
$$

Then we have

$$
\begin{aligned}
Var\left(\hat{\beta}_j\mathbb{1}_j\right) &= E\left[\mathbb{1}_j Var\left(\hat{\beta}_j|\mathbb{1}_j\right)\right] \\
&\quad + Var\left[\mathbb{1}_j E\left(\hat{\beta}_j|\mathbb{1}_j\right)\right] && \text{, by law of total variance} \\
&= \left[\pi_S Var\left(\hat{\beta}_j|j\in S\right)\right] \\
&\quad + \left[\pi_S E^2\left(\hat{\beta}_j|j\in S)\right) - \pi_S^2 E^2\left(\hat{\beta}_j|j\in S\right)\right] && \text{, by } (*) \\
&= \pi_S Var\left(\hat{\beta}_j|j\in S\right) + \pi_S(1-\pi_S)E^2\left(\hat{\beta}_j|j\in S\right) \\
&= \pi_S\left[\sigma^2_{b\in S} + E\left(V_j|j\in S\right)\right] + \pi_S(1-\pi_S)\beta^2_{\in S}
\end{aligned}
$$

where the final line follows from our model for $\hat{\beta}_j$ (the noise around β_j is symmetric and its sign is independent of β_j and η_j).

Similarly, we can evaluate the covariance term by using the law of total covariance and

noting

$$
\begin{aligned}
Cov\left(\hat{\beta}_j \mathbb{1}_j, \hat{\beta}_k \mathbb{1}_k\right) &= Cov\left[E\left(\hat{\beta}_j \mathbb{1}_j | \mathbb{1}_j, \mathbb{1}_k\right), E\left(\hat{\beta}_k \mathbb{1}_k | \mathbb{1}_j, \mathbb{1}_k\right)\right] \\
&\quad + E\left[Cov\left(\hat{\beta}_j \mathbb{1}_j, \hat{\beta}_k \mathbb{1}_k | \mathbb{1}_j, \mathbb{1}_k\right)\right] \\
&= Cov\left[\mathbb{1}_j E\left(\hat{\beta}_j | \mathbb{1}_j\right), \mathbb{1}_k E\left(\hat{\beta}_k | \mathbb{1}_k\right)\right] \\
&\quad + E\left[\mathbb{1}_j \mathbb{1}_k Cov\left(\hat{\beta}_j, \hat{\beta}_k | \mathbb{1}_j, \mathbb{1}_k\right)\right] \\
&= Cov\left(\mathbb{1}_j \beta_{\in S}, \mathbb{1}_k \beta_{\in S}\right) \\
&\quad + E\left[\mathbb{1}_j \mathbb{1}_k Cov\left(\hat{\beta}_j, \hat{\beta}_k | j \in S, k \in S\right)\right] \qquad \text{, by (*)} \\
&= Cov(\mathbb{1}_j, \mathbb{1}_k)\beta^2_{\in S} + E(\mathbb{1}_j \mathbb{1}_k) Cov\left(\hat{\beta}_j, \hat{\beta}_k | j \in S, k \in S\right) \\
&= \pi_S \left(\frac{s-1}{J-1} - \pi_S\right) \beta^2_{\in S} \\
&\quad + \pi_S \left(\frac{s-1}{J-1}\right) Cov\left(\beta_j, \beta_k | j \in S, k \in S\right)
\end{aligned}
$$

where the coefficients in the last line can be found directly (for the covariance, take the mean of the product minus the product of the means, and for the expectation, use the tower law) or by noting that jointly the indicators come from a multivariate hypergeometric distribution. For notational simplicity, we write $C_{\in S} := Cov\left(\beta_j, \beta_k | j \in S, k \in S\right)$ (and it was ok to replace $\hat{\beta}_j$ with β_j and the same for k because the errors e_j and e_k are assumed to be symmetrically distributed around 0). We note this covariance term is not zero in the uninteresting case that the subset contains only one site (throwing out all but one) since then in the second-to-last line $\mathbb{1}_j \mathbb{1}_k = 1$ with probability 1.

And so finally we have (from above)

$$\begin{aligned}
Var\left(\hat{\beta}_S\right) &= \frac{J}{s^2} Var\left(\hat{\beta}_j \mathbb{1}_j\right) + \frac{J(J-1)}{s^2} Cov\left(\hat{\beta}_j \mathbb{1}_j, \hat{\beta}_k \mathbb{1}_k\right) \\
&= \frac{J}{s^2}\left[\pi_S\left(\sigma^2_{b\in S} + E[V_j | j \in S]\right) + \pi_S(1-\pi_S)\beta^2_{\in S}\right] \\
&\qquad + \frac{J(J-1)}{s^2}\left[\pi_s\left(\frac{s-1}{J-1} - \pi_S\right)\beta^2_{\in S} + \pi_S\left(\frac{s-1}{J-1}\right) C_{\in S}\right] \\
Var\left(\hat{\beta}_S\right) &= \frac{1}{s}\left[\sigma^2_{b\in S} + E(V_j | j \in S) + (s-1) C_{\in S}\right].
\end{aligned}$$

B.7.3 The Covariance Term $C_{\in S}$ is Likely to be Negligible

There are a few reasons that $C_{\in S}$ will tend to be very small, likely negligible. In fact, if we make basic regression-type assumptions on the joint distribution of $(\beta_1, \ldots, \beta_J)$ and $(\eta_1, \ldots, \eta_J)$ then $C_{\in S}$ is nonzero only because the truncation region is random (since we truncate using the order statistic $\eta_{(J-s)}$). In addition, $C_{\in S} \to 0$ as either $\pi_S \to 1$ (when we discard a smaller proportion of sites) or $J \to \infty$ (asymptotically in the number of sites).

In particular, using the law of total covariance to condition on the order statistic, we can see that

$$\begin{aligned}
C_{\in S} = E_{\eta_{(J-s)}}&\left[Cov(\beta_j, \beta_k | j, k \in S, \eta_{(J-s)})\right] \\
&+ Cov_{\eta_{(J-s)}}\left[E(\beta_j | j \in S, \eta_{(J-s)}), E(\beta_k | k \in S, \eta_{(J-s)})\right]. \qquad (+)
\end{aligned}$$

The point of conditioning on the order statistic is to make the truncation region (the cutoff on η for whether a site is kept) fixed. This makes it convenient for us use the results (45.156) and (45.157) from Kotz, Balakrsihnan, & Johnson (2000) (originally due to Aitken (1936) and Lawley (1943)), which gives the joint mean vector and covariance matrix of two random vectors after conditioning on a *fixed* selection event (of arbitrary form) on the first random vector, at least under some regression-type assumptions on the two vectors' joint distribution.

In particular, their expressions are valid when (1) the conditional expectation of the second vector given the first vector is linear in the first vector and (2) the conditional covariance matrix of the second vector given the first vector does not depend on the first vector.

Lemma B.6. *Assume additionally that the conditional expectation of* $(\beta_1, \ldots, \beta_J)$ *given* $(\eta_1, \ldots, \eta_J)$ *is linear in* $(\eta_1, \ldots, \eta_J)$ *and that the corresponding conditional covariance matrix does not depend on* $(\eta_1, \ldots, \eta_J)$ *(in other words, is homoscedastic in the log precisions), while also conditioning on* $\eta_{(J-s)}$*. Then* $Cov(\beta_j, \beta_k | j, k \in S, \eta_{(J-s)}) = 0$*.*

Proof. We apply KBJ expression (45.157) to the two vectors (η_j, η_k) and (β_j, β_k) with (in their notation) $V_{11} = \sigma^2_{\eta,\sim} I_2$, $V_{12} = V_{21} = \sigma_{\eta b,\sim} I_2$, and $V_{22} = \sigma^2_{b,\sim} I_2$, where the $\sim$ notation indicates that these terms are conditional variances and covariances given $\eta_{(J-s)}$. This gives

$$Cov\left(\left[\begin{smallmatrix}\beta_j\\ \beta_k\end{smallmatrix}\right] | j \in S, k \in S, \eta_{(J-s)}\right) = \sigma^2_{b,\sim} I_2 - \sigma^2_{\eta b,\sim}/\sigma^2_{\eta,\sim}(I_2 - \sigma^{-2}_{\eta,\sim} U_{11})$$

where $U_{11} = Cov(\left[\begin{smallmatrix}\eta_j\\ \eta_k\end{smallmatrix}\right] | j \in S, k \in S, \eta_{(J-s)}) = Var(\eta_j | \eta_j \geq \eta_{(J-s)}, \eta_{(J-s)}) I_2$. Since U_{11} is diagonal there is no conditional covariance between β_j and β_k. □

Remark. The assumptions of Lemma B.6 (linear conditional expectation and homoscedasticity) are essentially the FIRC+ model without normality.

Applying Lemma B.6 to (+) implies that

$$C_{\in S} = Cov_{\eta_{(J-s)}}\left[E(\beta_j | j \in S, \eta_{(J-s)}), E(\beta_k | k \in S, \eta_{(J-s)})\right]$$

and we can see from (45.156) in KBJ that both conditional expectations are the product of the conditional mean of η_j after truncation on the same random cutoff $\eta_{(J-s)}$ and the regression coefficient of β_j on η_j (conditioning on the random order statistic). Because these terms depend weakly on the same order statistic, this covariance will be nonzero if small. However, asymptotically the dependence will fade as the order statistic becomes fixed.

B.7.4 Self-inefficiency of the Unweighted Estimator

Theorem 3.2. *Under* (3.2.1)*, suppose that for some* $s \in \{1, \ldots, J-1\}$ *the set of the* s *most precise sites,* $S := \{j \in \{1, \ldots, J\} : \eta_j > \eta_{(J-s)}\}$*, satisfies*

$$E(V_j | j \in S) \leq \pi_S E(V_j) - \left[\sigma^2_{b\in S} - \pi_S \sigma^2_b\right] - s(1-\pi_S)^2 \Delta^2_S - (s-1) C_{\in S}$$

or, equivalently,

$$\Delta^2_S \leq (1-\pi_S)^{-2} s^{-1} \left(\left[\pi_S E(V_j) - E(V_j | j \ \in S)\right] + \left[\pi_S \sigma^2_b - \sigma^2_{b\in S}\right] \right) - (1-\pi_S)^{-2} \frac{s-1}{s} C_{\in S}.$$

Then the unweighted estimator $\hat{\beta}_{UW}$ *is self-inefficient as an estimator of* β*.*

Proof. Using the results from above the proof just requires very basic algebra, which we show for completeness. For the bound on $E(V_j | j \in S)$, note that $MSE(\hat{\beta}_{UW}, \beta) \geq MSE(\hat{\beta}_S, \beta)$ if and only if

$$\begin{aligned}
Var(\hat{\beta}_{UW}) &\geq Bias^2(\hat{\beta}_S, \beta) + Var(\hat{\beta}_S) \\
\frac{1}{J}\left[\sigma^2_b + E(V_j)\right] &\geq (1-\pi_S)^2 \Delta^2_S + \frac{1}{s}\left[\sigma^2_{b\in S} + E(V_j | j \in S)\right] + \frac{s-1}{s} C_{\in S} \\
\pi_S \sigma^2_b + \pi_S E(V_j) &\geq s(1-\pi_S)^2 \Delta^2_S + \sigma^2_{b\in S} + E(V_j | j \in S) \\
E(V_j | j \in S) &\leq \pi_S E(V_j) + \left[\pi_S \sigma^2_b - \sigma^2_{b\in S}\right] - J\pi_S (1-\pi_S)^2 \Delta^2_S - (s-1) C_{\in S}
\end{aligned}$$

Similarly, we may solve for Δ^2_S:

$$\begin{aligned}
(1-\pi_S)^2 \Delta^2_S &\leq \left[\frac{1}{J}\sigma^2_b - \frac{1}{s}\sigma^2_{b\in S}\right] + \left[\frac{1}{J} E(V_j) - \frac{1}{s} E(V_j | j \in S)\right] \\
\Delta^2_S &\leq (1-\pi_S)^{-2} \frac{1}{J\pi_S} \left[\left(\pi_S E[V_j] - E[V_j | j \in S]\right) + \left(\pi_S \sigma^2_b - \sigma^2_{b\in S}\right) \right].
\end{aligned}$$

□

Remark. The condition in this theorem is sufficient for Meng & Xie's notion of self-efficiency but necessary and sufficient for the limited notion we focus on.

References

Aitken, A. C. (1936), "A further note on multivariate selection," Proceedings of the Edinburgh Mathematical Society, 5, 37–40.

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000), "Truncated Multivariate Normal Distributions," in Continuous Multivariate Distributions, Wiley series in probability and statistics, Wiley, pp. 204–205.

Lawley, D. N. (1944), "IV.—A Note on Karl Pearson's Selection Formulæ," Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Sciences, 62, 28–30.

DESCRIPTION OF SUPPLEMENTARY FILES

Supplementary files may be found attached, and include code to run the analyses and simulation studies included in each of the three chapters.