Clinical Trial Designs that Utilize Historical Controls in the setting of Cystic Fibrosis

Mark Niklas Warden

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Stephen Mooney, Chair

Amalia Magaret

Nicole Mayer-Hamblett

Program Authorized to Offer Degree:

Epidemiology

University of Washington

Abstract

Clinical Trial Designs that Utilize Historical Controls in the setting of Cystic Fibrosis

Mark Niklas Warden

Chair of the Supervisory Committee:

Stephen Mooney

Department of Epidemiology

In certain rare disease settings, traditional randomized controlled trials face ethical and feasibility challenges that can exclude their use. Adequate participant recruitment can be prohibitively difficult, and a large placebo control arm could unethically expose vulnerable participants to risk. Development of new therapeutics for these diseases still require regulatory grade evidence of a treatment effect, and researchers must seek alternative but robust trial designs to meet these demands. Historically controlled trials are one potentially statistically efficient approach that can reduce needed sample sizes, and their applicability can be well demonstrated in cystic fibrosis (CF) research. Statistical methods for historically controlled trials include, among others, inverse probability weighting (Rosenbaum 1983), propensity score-based power priors (Lin 2017), and hierarchical Bayesian modeling with commensurate priors (Hobbs 2012). Additionally, an adaptive study design method for incorporating historical controls was proposed by Psioda in 2018. We assessed these four approaches in a large simulation study to quantify bias, coverage, precision, and power. All approaches performed well in terms of type I error and power even when

there were different covariate distributions between the simulated historical and active trials. However, notable strengths and weaknesses of each were observed. We then reanalyzed two sequentially completed CF trials of treatments for pulmonary exacerbations to test whether the later trial could have been historically controlled with data from the prior trial using these methods. Specifically, we assessed whether the earlier trial data could be used to reduce the size of the concurrent control group in the later trial. This hypothetical historically controlled trial achieved similar results to the real trial results, establishing proof of concept for the validity of these methods in CF. We conclude that the later trial could have been conducted with a smaller concurrent control group without sacrificing precision and that historically controlled trials have great potential for use in future CF clinical trials.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The author sincerely appreciates the support of the following individuals and organizations without which this work would not have been possible:

Dr. Stephen Mooney, committee chair, for careful consideration and insights of casual inference assumptions in historical trials, for his teaching in my advanced epidemiology coursework, and for his networking and early career advice and support.

Dr. Nicole Mayer-Hamblett, committee member and head of the Cystic Fibrosis Therapeutics Development Network Coordinating Center (CF TDNCC), for conceiving of and helping to develop this work and for providing financial support.

Dr. Amalia Magaret, committee member, for substantive week to week methodological oversight, for excellent advice at all stages, and for emotional support dealing with the pressure of this work's completion. I am exceptionally grateful.

Dr. Joseph (Chris) Delaney, committee member, for excellent suggestions and critique during the process that expanded the work to be more applicable in practical contexts.

The Cystic Fibrosis Therapeutics Development Network Coordinating Center's Biostatistics and Clinical Data Management team, especially Dr. Noah Simon who informally consulted on the work and provided his expertise.

Additional thanks to the Cystic Fibrosis Foundation, Dr. Katie Davis for her flexibility and support, Dr. Ben Marwick as the Graduate School Representative, Lisya VanHousen for financial management and assistance in many logistical matters, and the Seattle Quake rugby team members who helped lighten the load of life during this dissertation process.

DEDICATION

To my late mother, Dr. Kendall Carnes Warden, who loved and encouraged me to strive for excellence, and to my father and brother, Dennis and Nathan Warden, for their unwavering support.

Chapter 1. BACKGROUND AND MOTIVATION

1.1 RARE DISEASES, THE FDA, AND MODERN APPROACHES TO CLINICAL TRIALS

The FDA defines a rare disease as one that affects fewer than 200,000 people in the United States (U.S.), and examples include Huntington disease, Duchenne muscular dystrophy, and cystic fibrosis. There are approximately 7,000 rare diseases that affect over 25 million Americans, and 94% still lack an approved treatment[1,2]. Genetic disorders account for 80% of rare diseases and often manifest from birth. Therefore, about half of rare diseases are also classified as pediatric rare diseases[1,3] and affect a vulnerable population, further intensifying the need for quality research. The FDA acknowledges that "developing safe and effective products to treat rare diseases can be very challenging," noting that there are often subpopulations within the rare disease of smaller sizes and different clinical manifestations[1]. These characteristics lead to challenging research environments with relatively few clinical trials and fewer investigators.

Between January 2010 and December 2012, Rees et al.[3] identified 659 rare disease trials registered at ClinicalTrials.gov that enrolled a total 70,305 patients. The authors found 199 (30.2%) were discontinued and 142 (31.5%) remaining unpublished four years after trial completion, stifling research progress. These percentages were similar between pediatric (79 trials) and adult (580 trials) studies. When a reason was provided, 62 of the 199 discontinued trials reported patient accrual as the cause. The authors conclude that innovations are needed to improve patient participation and trial quality to advance scientific research in these settings.

Cognizant of research struggles in rare diseases and lack of treatments, the FDA has increasingly approved new drugs and treatments without a traditional randomized controlled trial. Downing et al. 2014[4] report that between 2005 and 2012, the FDA approved 188 novel therapeutic agents, 31 (16.5%) in orphan diseases and 22 (11.7%) using an accelerate approval pathway (likely rare diseases). An orphan disease is defined as a rare disease for which treatments are often not considered profitable due to their cost to develop and limited patient population. A total of 448 pivotal efficacy trials were identified with 143 (31.9%) approved with *only* an active comparator and 58 (12.9%) approved *without any* comparator arm. Furthermore, in the trials of orphan

diseases, 50% had no comparator arms and 15% had only an active comparator, and these percentages were 55% and 21% respectively trials on the accelerated approval pathway[4]. Alternative trial designs have become increasingly acceptable to regulators to meet the modern needs of patients with rare diseases and limited or no treatment options.

In 2016, the United States congress passed the 21st Century Cures Act[5] to improve research pipelines, especially for rare diseases. In section C, (3021-3024), congress instructed the FDA to provide guidance to assist sponsors "in incorporating complex adaptive and other novel trial designs into proposed clinical protocols and applications for new drugs."[6] One such novel trial design in a rare disease setting is an externally controlled trial, which is the focus of the research in this dissertation. Externally controlled trials use historical control data taken from other sources such as previously conducted randomized controlled trials (RCTs). Historical controls can also be taken observational studies (notably prospective and retrospective cohort studies) and other transactional data sources such as information from registries, electronic health records, and transactional insurance claims. This kind of data is called real-world data (RWD).[7]

In response to the act, the FDA drafted and certified a new guidance on the design and conduct of externally controlled trials[6]. The central consideration is the potential treat to validity introduced by bias associated with differences between the external control group and active trial participants. These differences could include time periods, geographic regions, diagnosis criteria, treatment definitions, outcome definitions, and others.[6] These concerns are the same whether the external data is derived from a high quality completed RCT or from possibly lesser quality RWD. Therefore, the FDA guidance for externally controlled trials is intertwined with concepts regarding the utilization of RWD, meaning literature related to RWD is useful to review.

An additional mandate in the 21st Century Cures act was to create a program called the Real-World Evidence program, including a document describing its framework[8]. The mandate acknowledges that trials that utilize external RWD data can create robust, regulatory grade evidence, termed real-world evidence (RWE), for drugs and biologics but not devices[9,10]. The Real-World Evidence program seeks to provide guidelines and standards to encourage researchers "to modernize clinical trial designs, including the use of real-world evidence, and

clinical outcome assessments, which will speed the development and review of novel medical products"[5] as one of its key goals. Almost all the literature and guidance regarding RWE is applicable to externally controlled trials that use previous RCT data. Conversely, the methods assessed in this work also may be applicable to RWE settings in future research.

Primary investigators funded by the FDA through the RWE initiative, Franklin and Schneeweiss, have published articles with recommendations on real world data analyses[7,11] that are relevant to best practices when conducting externally controlled trials. They are directing a research project called RCT duplicate[12] that seeks to duplicate important randomized controlled trials using prospective longitudinal insurance claims data. The authors have identified "four key use cases in which RWE could support regulatory decision making,"[11] and by extension externally controlled trials in general. The first case is the primary approval of a therapeutic and includes using historically controlled trials and synthetic comparison groups[11]. Historically controlled trials are the focus of this dissertation, however the techniques and approaches evaluated may be applicable to the remaining use cases. The second case is indication expansion, i.e., approval for new populations (pediatrics, more extreme disease severity stages, or additional genotypes). The third case is an adaptive approval pathway or when "an initial conditional marketing authorization was made for a limited population with high unmet medical need on the basis of biomarker data or small clinical trials" and the full approval occurs using data gathered from a registry or other RWD source after the initial approval. The fourth and final case is post-market safety surveillance, which can include "rapid regulatory response to a safety signal."[7]

There are many scenarios in which externally controlled trials and the RWE framework are preferred to traditional randomized controlled trials:

> "When studying a highly promising treatment for a disease with no other available treatments, ethical consideration may preclude randomizing patients to placebo, particularly if the disease is likely to result in severely compromised quality of life or mortality. In these cases, RWE can support product regulation by providing evidence on the safety and effectiveness of the therapy against the typical disease progression observed in the absence of treatment." (pg. 925-916)[7]

Franklin and Schneeweiss specify that the goal in primary approvals is "to quantify the typical disease trajectory in the absence of the new medication to provide a counterfactual to the disease trajectory observed on the experimental treatment."[11] The emphasis of their work is that trials that use real world data must be conducted rigorously to avoid many common pitfalls and biases, and they provide a structured process for implementation of RWD analyses for regulatory decision making. Even when using previously completed randomized controlled trial data, the same important questions apply: 1) Is the completed trial data a quality fit in terms of protocol and definitions? 2) Is this data's use as an external comparator arm feasible? 3) Can this data be used in a way that is responsive to additional analyses requested by regulators (**Figure 1**)?

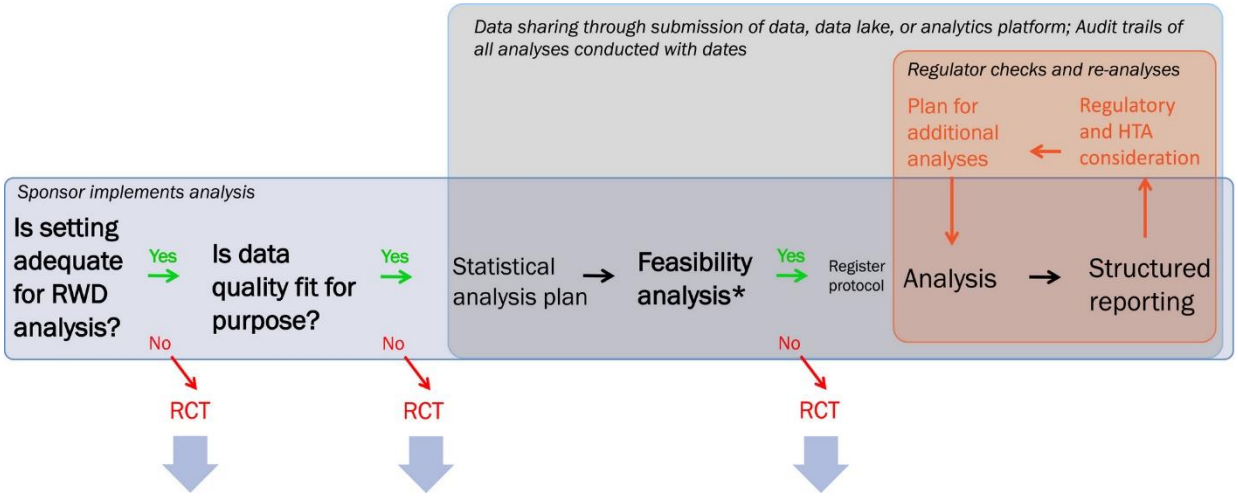


Figure 1. A structured process for implementation of RWD analyses from Schneeweiss 2019[6]. Reprinted from "Evaluating the Use of Nonrandomized Real-world Data Analysis for Regulatory Decision Making," by J. M. Franklin, R. J. Glyn, D. Martin, and S. Schneeweiss[11], 2019, Clinical Pharmacology & Therapeutics, p. 872. © 2019 The Authors Clinical Pharmacology & Therapeutics © 2019 American Society for Clinical Pharmacology and Therapeutics. "*Feasibility analysis can include checking covariate balance after applying the chosen confounding adjustment strategy, checking statistical power, evaluating positive or negative control outcomes, and other analyses without evaluating the study outcomes in the two treatment groups." DOI: 10.1002/cpt.1351. Reprinted with permission.

In a situation where high-quality research data derived from completed studies can be reused, there is an opportunity to design modern clinical trials that are efficient and timely. This approach can expedite the development of new therapeutics. The following chapters will examine analytic and study design approaches in historically controlled trials. These approaches,

when conducted in an appropriate manner, can rigorously meet modern challenges in certain rare disease settings.

Another important note on the importance of externally (historically) controlled trials is how they compare to active comparator trials, which are trials that compare an experimental treatment to another treatment that is typically the standard of care. In the case where there is an available treatment (as in certain cases in CF research), comparing a new treatment to this standard of care treatment can avoid ethical issues of having participants on placebo arms. However, in exchange for this benefit, the active comparator trial must be designed with high assay sensitivity, defined as the ability to distinguish between two treatment effects.[13] This contrasts with comparing to no treatment effect. The trial must have high enough assay sensitivity to detect efficacy in the form of non-inferiority or superiority. In the case of non-inferiority, statistical margins can be small to account for important trial design aspects and this conservatism can lead to large required sample sizes. Even in the case of a superiority trial, "the expected difference between two drugs is always smaller than the expected difference between [a] test drug and placebo, again leading to large sample sizes."[13] A central goal of historically controlled trials is a reduction in sample size, i.e. a gain in statistical efficiency, and these trials are one approach a researcher might choose over an active comparator trial.

1.2 CYSTIC FIBROSIS AND THE NEED FOR MODERN CLINICAL TRIAL DESIGNS

Cystic fibrosis (CF) is a rare, monogenic, autosomal recessive genetic disorder that affects about 70,000 individuals worldwide and 30,000 in the United States[14]. The disorder disrupts the CF transmembrane conductance regulator (CFTR) protein, which is used in many organ systems[15]. In the lungs, the most seriously affected organ, cell tissue slowly breaks down and scar tissue can develop due to lack of clearance of mucus[16]. Additionally, individuals with CF have a greatly increased risk of lung infection for a variety of pathogens but notably *Pseudomonas aeruginosa (Pa)*[14,16–18] which has been associated with an increased risk of pulmonary exacerbations (PEx) requiring antibiotics. CF patients frequently experience PExs and even with modern treatment they still occur with regularity. Prior to 2012, the median life expectancy for individuals with CF

was in their early thirties[14], and therapeutics were focused on symptom management.

Fortunately, in 2012, the first ever disease-modifying therapy (termed a CFTR modulator) was developed called ivacaftor (brand name: Kalydeco[19]). This agent is a small molecule which corrects one class of cellular deficits in the CFTR protein associated with mutations in the CF gene present in less than 10% of the CF population. In subsequent years, three additional combination CFTR modulators[20–22] have been developed and approved, providing effective primary therapies for up to 94% of individuals with CF with modulator responsive mutations. The median life expectancy for Americans with CF improved to around 50 years of age. The most recent modulator (elexacaftor, tezacaftor, ivacaftor, or brand name Trikafta[22]) improves lung function so well that hospitalization time due to PExs decrease on average by 86% from 27 days to 4 days.[23] The benefits can also be seen in improved BMI and self-reported outcomes.[24] This adoption of CFTR modulators into standard of care is a major milestone for the majority of individuals with CF.

The discovery and approval of CFTR modulators has empowered researchers to continue to develop new therapies for CF but has also created new challenges. For example, trials of new treatments for the remaining ~6% of CF patients ineligible for modulators are targeting an extremely small study population of approximately 2,000 people across the US. Feasibility (participant accrual, consent, and retention) presents a great challenge in the development of new therapeutics for these sub populations and makes traditional, large, randomized controlled trials impractical for regulatory approvals. For novel therapies, including nucleic acid-based therapies focusing on the ~6% of individuals with CF with genetic mutations currently not eligible for modulator therapy, the small population size is the greatest challenge to demonstrating safety and efficacy for FDA approval. However, externally controlled trials that leverage historical data can address these new challenges and are encouraged in these situations.

In addition to sample size considerations for new drugs targeting CF sub-populations, development of new primary and secondary therapies in patients currently benefiting from an approved modulator treatment are facing a new challenge. Lengthy placebo-controlled randomized clinical trials may no longer be appropriate. In the US, CFTR modulators have

become the standard of care for many individuals with CF with eligible mutations because they have been shown to greatly improve clinical outcomes. Depending on the length of a potential trial for a new CF therapy, it may not be ethical for participants to be assigned to a placebo study arm that would require withdrawal of standard of care modulator therapy. In fact, a feasibility study surveyed CF treating physicians and found only 62% would consider withdrawal of the standard of care CFTR modulator for up to 4 weeks for a trial of a new CFTR modulator or other CF therapeutic. The survey found CF patients and their caregivers were less willing to interrupt treatment (50%), and willingness was even lower for longer duration trials, falling below 20% as reported by both patients and physicians[25].

Increasing CFTR modulator access is expected to reduce the rates of lung function decline and exacerbation risk as well as increasing quality of life. However, these substantial improvements do not fully restore health and additional treatments may still be beneficial. Ancillary treatment research, including treatments targeting mucociliary clearance, inflammation, infection, and nutrition, has been based on evidence derived from a CF population not yet on modulators[26]. Therefore, it is unclear if chronic secondary treatments, such as hypertonic saline and dornase alfa, are still beneficial for a population on modulators; and investigators are currently studying that question[27,28]. Research on these types of therapies must be repeated and aligned such that they are relevant to populations on modulators. The new standard of care will most likely result in statistically small treatment effect sizes on clinical outcomes in modern clinical trials, requiring even larger sample sizes to detect. However, it is plausible that these treatments may offer incremental benefits on top of modulators and therefore will remain clinically meaningful. Demonstrative evidence of benefit is difficult to obtain but needed. Therefore, alternative trial designs that allow for smaller sample sizes, such as historically controlled trials, are again useful for reassessing current secondary treatments and continuing development of new ones to be used in tandem with modulators.

The development of new primary treatments is also an essential long term goal of many global CF clinical trial networks, including the CF therapeutics development network (CF TDN), European CF Society Clinical Trials Network (ECFS-CTN), and others[26]. These primary treatments include novel nucleic acid-based therapies, such as nonsense-mediated decay

inhibitors (NMD), antisense oligonucleotides (ASOs), and mRNA therapies[29], as well as new CFTR modulators. These new primary therapies are essential to impact individuals with rare CF causing mutations unresponsive to currently available modulators many of whom are from racial minority groups. A recent cross-sectional study found that individuals with CF from racial minority groups were less likely to be eligible for CFTR modulators[30]. New primary therapies may also decrease costs and increase accessibility when considering patients who live in regions without approvals for currently developed drugs and for those with financial barriers to access treatment. Timely development of new modulators could improve equity of health outcomes.

Meeting required efficacy and safety standards for new CF therapeutics will require new approaches to address challenges such as ethical comparator arms (reduced length trials or historical controls) and smaller expected effect sizes. A group of core advisors from multiple international regions and clinical trial networks have suggested that a study design package consisting of multiple components may be most appropriate for approval of new primary and symptomatic CF therapies by regulatory agencies. The combination allows investigators to generate all the required evidence in a logistically feasible and ethical way, avoiding many of the problems already discussed. These components include short term placebo-controlled trials, open-label prescription-reliant active comparator trials, open-label active arm safety studies, and *historically controlled trials*[31]. Each type of study design can provide different desirable characteristics and can collectively comprise robust evidence of efficacy and safety (see Hamblett et al. for details on the other components). Historically controlled trials, the focus of this proposal, take advantage of available data from past clinical trials to create an external control group to replace or augment an active placebo-controlled arm. This component can look at long term clinical efficacy as well as certain key clinical outcomes, notably pulmonary exacerbations.

The CF research community in the United States is well prepared to conduct historically controlled trials because of its centralized CF trial network. The CF therapeutics development network (CF TDN) is a clinical trials network consisting of about 90 clinical research centers across the country that coordinate their efforts in recruiting patients and conducting clinical trials. An extensive archive of completed clinical trial data is maintained through the CF TDN

coordinating center, a central hub that directs the network. The TDN coordinating center stores and analyzes the trial data on a variety of key outcomes and therapeutics. The trial protocols, data dictionaries, case reporting forms, and raw data are all available, and the hope is that this data can be leveraged in the development of new CF therapies in the ways described above. Additionally, the Cystic Fibrosis Foundation maintains the CF patient registry, which includes encounter-based and annual clinical data collection on almost all individuals with CF in the U.S. This real-world registry data is made available to researchers and has already been used in many studies. These characteristics have enabled great scientific advancements in the treatment of individuals with CF, and hopefully will continue to support creative new trial designs.

Historically controlled trials are becoming accepted and recommended in these situations, but current publications focus on the statistical theory of the approaches with a small application in real data. These papers provide a solid statistical foundation, but further research and preparations are required to implement them in CF and other rare disease research. Both Bayesian and frequentist methods have been proposed to incorporate historical controls into trial designs[32], but there are few studies that have directly compared these new methods and apply them with CF trial data. We seek to assess these methods in the context of CF because these types of clinical trials are desirable to meet the evolving demands of CF patients, industry sponsors of CF therapeutics, and regulatory agencies[33]. We will expand upon the literature regarding the proper methods to conduct historically controlled trials in CF.

1.3 RESEARCH OVERVIEW

This dissertation seeks to innovate in clinical trial design by comparing analytic methodologies and adaptive trial designs that reduce the need for a placebo-controlled arm by using external data in the form of historical control arm trial participants. The broad goal is to produce disease and context specific recommendations to facilitate the first use of historically controlled clinical trials in CF[31,33,34] with a level of rigor acceptable by the FDA; however, other disease contexts could also implement these recommendations. This dissertation is separated into three parts, the first two of which are conducted in a simulation study and the final implemented in real trial data (**Figure 2**).

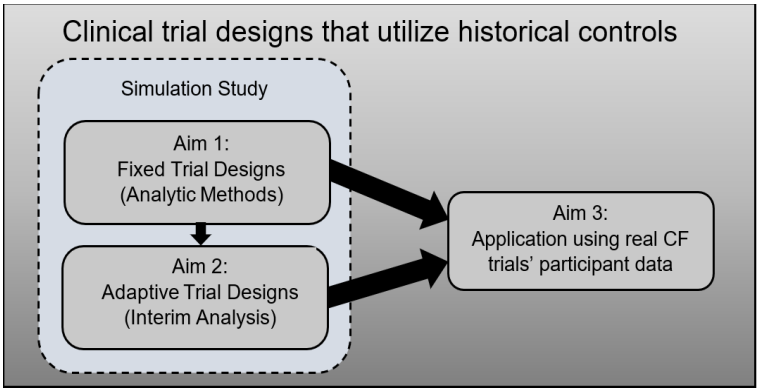


Figure 2. Components of the current study. Aim 1 and 2 are situated within a simulation study, evaluating the selected analytic methods and adaptive study design. Aim 3 will use the results of the previous aims and apply the techniques in real CF trial data.

The simulation study will assess fixed trial designs that use statistical, analytic methods to incorporate the historical control data and an adaptive trial design that uses a planned interim analysis to determine whether the inclusion of the external data is appropriate and improves study power. One specific goal is to consider both situations in which the historical trial data and current trial data are similar and are different in terms of covariate prevalence. Both cases are possible and unknown to investigators in the planning stages of a trial. A set of 64 possible combinations of parameters will be used for the data-generating mechanisms as described in the next section. These sets represent a wide range of scenarios in which the historical trial data may be similar or substantially different from the active trial. Examining performance measures in these scenarios can help determine which factors are most influential on these approaches, thereby providing useful insights for the planning of future trials. Finally, the results are extended to an application in real CF trial data to connect the key results of the simulation study back to disease specific and more complex data.

1.4 SIMULATION DATA GENERATING MECHANISMS, PERFORMANCE MEASURES, AND PROCEDURES

We followed the best practices for evaluating statistical methods using simulation studies described by Morris et al.[35] We generated data sets according to prespecified parametric models, containing a historical control indicator, a treatment arm indicator, two dichotomous covariates,

and a time-to-first event outcome with right censoring. The parameters for the parametric models were either prespecified or dependent on the specific simulation scenario. A total of 64 scenarios were simulated which differed on six specific input parameters: (1) treatment effect size (2) the covariate prevalence in the active trial, (3) agreement between the historical and active trial participant covariate prevalence, (4) the covariance between the two covariates (assumed to be the same for both the active and historical trial), (5) the strength of association between the covariates and the outcomes (i.e. the covariates' β values), and (6) the relative number of historical controls available. Additional details of the scenarios, parametric models, and data generation parameter values are described below:

Table 1.1 Simulation specifications and data-generating mechanisms.

<table>
<tr><th>Simulation characteristic</th><th>Variable(s)</th><th>Distribution(s)</th><th>Notes and parameter value(s)</th></tr>
<tr><td>Number of active trial participants</td><td>n_{active}</td><td>Constant</td><td>• Calculated as if traditional randomized controlled trial with 80% power and 5% type I error
• Analytic methods consider a concurrent control group that is half the size of the treatment arm</td></tr>
<tr><td>Historical control indicator</td><td>H_i</td><td>$H_i \in (0,1)$</td><td>• Only historical controls have a value of 1</td></tr>
<tr><td>Treatment indicator</td><td>E_i</td><td>$E_i \in (0,1)$
$E_i \sim \begin{cases} Bern(0.5), & if\ H_i = 0 \\ 0, & if\ H_i = 1 \end{cases}$</td><td>• Historical controls have value 0
• Active trial treatment assignment is a Bernoulli random variable. Therefore, active trial treatment and control arms may not always be perfectly balanced</td></tr>
<tr><td>Binary covariate, X_1</td><td>$x_{i,1}$</td><td>$X_1 \sim probit(W_1)$
$x_{i,1} \in (0,1)$</td><td>• A probit model produces binary variable with prevalence: $p = \mu_{x_1}$</td></tr>
<tr><td>Binary covariate, X_2</td><td>$x_{i,2}$</td><td>$X_2 \sim probit(W_2)$
$x_{i,2} \in (0,1)$</td><td>• A probit model produces binary variable with prevalence: $p = \mu_{x_2}$</td></tr>
<tr><td>Latent, normal variable used to build X_1, W_1</td><td>$w_{i,1}$</td><td rowspan="2">$W = \{W_1, W_2\}$
$W \sim N(\mu, \Sigma)$
$\mu = \{\mu_{w_1}, \mu_{w_2}\}$</td><td rowspan="2">
<table>
<tr><th>X</th><th>μ</th><th>σ^2</th></tr>
<tr><td>W_1</td><td>$\Phi^{-1}(\mu_{x_1}\|H)$</td><td>1</td></tr>
<tr><td>W_2</td><td>$\Phi^{-1}(\mu_{x_2}\|H)$</td><td>1</td></tr>
</table>
• μ_{x_1}, μ_{x_2} are the desired prevalence for the covariates
• μ_{x_1}, μ_{x_2} depend on whether a participant is part of the historical or active trial, see details below in (2) and (3)
• Σ is defined in (4)</td></tr>
<tr><td>Latent, normal variable used to build X_2, W_2</td><td>$w_{i,2}$</td></tr>
<tr><td>Outcome (Time; event)</td><td>$Y_i = (T_i, y_i)$</td><td>$t_i \sim exp(\beta_0 + x_i^T\beta + E_i^T\theta)$
$T_i \sim \min(t_i, D)$
$y_i \sim I(t_i < D)$</td><td>• t_i follows an exponential distribution and β is varied in (5)</td></tr>
</table>

			• β_0 is constant so that cumulative study event rate is 50% when $X_1 = 0$ and $X_2 = 0$. • $D = 1$, study duration for right censoring
(1) Treatment Effect	θ	Constant	• No effect, $HR = 1$ • Beneficial, $HR = 0.50$
(2) X_1, X_2 prevalence in the active trial	$\mathrm{M} = \#(\mu)$	Constant	• $\mu_1 = (0.75, 0.2)$ • $\mu_2 = (0.85, 0.1)$
(3) Agreement between the historical and active controls' covariate prevalence	A = #(δ)	Constant	Constant added to historical trial μs to vary agreement with active trial • Strong, $\delta = (0,0)$ • Weak, $\delta = (-0.4, +0.2)$
(4) Covariance and correlation (Σ) between covariates	$\mathrm{C} = \#(\Sigma)$	Constant covariance/correlation structure	Defined by covariance matrices: • $cov = \begin{Bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{Bmatrix}$ • $cov = \begin{Bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{Bmatrix}$ • Note that because the variances (diagonals) are 1, then these matrices are both covariance and correlation matrices.
(5) Strength of association (β) between covariates and outcome (Y)	$S = \#(\beta)$	Constant	Log(HR)'s between covariates and outcome: • $\beta = \log(HR) = log\begin{Bmatrix} 1.10 \\ 1.30 \end{Bmatrix}$ • $\beta = \log(HR) = log\begin{Bmatrix} 1.20 \\ 0.65 \end{Bmatrix}$
(6) Number of historical controls relative to expected treatment arm size	$HC = \#(h)$	Constant	• $h = 1$, the number of available historical controls is the same size as the active trial's treatment arm • $h = 2$, the number of available historical controls is twice the size of the active trial's treatment arm
# of scenarios	SCs	$\mathrm{M} * A * C * S * HC * \#(\theta)$ $2 * 2 * 2 * 2 * 2 * 2 = 64$	• 2 treatment effects (1) • 2 covariate prevalence levels (2) • 2 levels of agreement (3) • 2 covariate covariances (4) • 2 levels of association between the covariates and outcome (5) • 2 relative values for the number of available historical controls (6)
# of simulations	N_{sims}	$\mathrm{SCs} * n_{sims}$ $64 * 3600 = 230{,}400$	• $n_{sims} = 3600$ required to control Monte Carlo error, see text below.

#() is the counting function

In the active trial, treatment arm assignment was determined as a random Bernoulli variable and was zero for all the historical controls. The size of the active trial, n_{active}, was calculated using Lachin's formula detailed in Fundamentals of Clinical Trials by Friedman et al.[36]. The formula used a target power of 80%, a controlled type 1 error of 5% and a treatment effect size corresponding to a hazard ratio (HR) of 0.5 or 0.7. The effect size in the formula varied by scenario. In scenarios simulating a beneficial treatment effect (a true HR of 0.5), we used an

effect size corresponding to this hazard ratio of 0.5, but, for scenarios where no treatment effect was simulated (a true HR of 1), an effect size corresponding to a hazard ratio of 0.7 was used. This allowed the results in these scenarios to be interpreted as a trial of a possible treatment with modest but meaningful effect that in truth had no effect. After a fully powered active trial was simulated, half of the control arm was discarded. The focus of the simulation was on reducing the number of needed controls by augmenting them with historical controls (termed the augmenting approach discussed in more detail later). The number of historical controls was set to be either equal to or double the expected number of treatment arm participants.

The parametric models for the two, dichotomous covariates, X_1,X_2, were determined so the level of binary agreement could be controlled with a single parameter, Σ. Specifically, two latent normal variables, (W_1,W_2), were generated with a variance of one, scenario-specific means and correlations, and a covariance matrix, Σ, that is also the correlation matrix because both covariates have a variance of one. These latent variables were transformed into the binary trial covariates, (X_1,X_2), by a probit model[37]. For example, if the latent, normal variables were greater than zero, then the covariate value would be one and zero otherwise. The latent variables' means were selected such that they produced the desired marginal prevalence in the historical and active trials for each scenario. We allowed for the historical controls to differ in prevalence for these covariates from the active controls and treatment arm for some scenarios as detailed earlier (**Table 1.1**). It should be noted that the strength of agreement between binary variables is restricted by their marginal distributions. To see an illustration of the covariate data generating mechanism and for further discussion, see **Appendix A**.

For the outcome variable, we sought to make the simulation conditions align as closely as possible to pulmonary exacerbations in preparation for the implementation of the methods in real trial data. We used an exponential distribution for the time to event variable, thereby implying a constant hazard rate (λ) dependent on the covariates and treatment arm[38]. A constant hazard rate has been observed in preliminary work with real trial data[33] and is described in more detail in section 4.1. Modeling time to event with an exponential variable creates a Poisson process where events occur continuously and independently at a constant rate. The outcome's rate can then be modeled with a generalized linear model (GLM) with a Poisson likelihood function and a natural

log link function[38], essentially a Poisson regression. The mean model assumption of GLMs is supported by real trial data in planning the outcome's data generating mechanism (see chapter 4.1), but the variance may be underestimated due to the lack of true count data (the time-to-event data only contains the first observed event per participant rather than counting all events in a time span). However, variance estimates can be corrected used the Huber-White sandwich estimator[37] that is able to estimate the true variance even when model assumptions are not perfectly met. We also chose this model for the outcome because the historical control analytic methods we selected were designed using the GLM framework and work well for time-to-event data with a constant hazard ratio.

The simulation's critical estimand was the treatment effect, measured as a hazard ratio and denoted θ. All methods estimated both a point and interval estimate of the treatment effect. The target of the simulation is to evaluate the effect of the six varying characteristics that comprise the 64 parameter combination scenarios and to compare the selected methods. Examining estimates from each of these scenarios can help determine which elements of the scenarios most impact the methods. Based on the results, we crafted recommendations for future trials that wish to use these methods.

When utilizing historically controlled trials, a natural question is whether to fully replace the control arm in the active trial, to retain a small set of active control participants augmented with historical controls, or to recruit a full active control arm that is pooled with the historical controls for the analysis. These questions were considered in our earlier work[39], but the focus of the current study will be the second case. We used the sample size needed for a traditional randomized trial but then halved the number of active controls. For all the methods, there was: a) a small active control arm – half the size of what the control arm would be in a fully powered traditional trial – b) a set of historical control data of two possible sizes discussed below, and c) a full-size treatment arm. This situation is most applicable to the development of future trials as it provides benefit over traditional designs by reducing total size. It is assumed that participants are randomized using a 2:1 strategy in the active trial to reflect a control arm half the size of what it would be in a fully powered traditional randomized controlled trial.

In each scenario, we assessed model convergence, bias, percent bias, empirical standard error, mean squared error (a combination of accuracy and precision), coverage, average confidence/credible interval length, power (when $\theta = 0.5$) or type I error (when $\theta = 1$), and the Monte Carlo introduced errors, which can be controlled through the number of iterations.

Model convergence is defined as whether or not the statistical method produces estimates. Sometimes certain methods use numerical computation that will fail to converge to a final estimate and the frequency of this failure defines model convergence.

Bias and percent bias are measures of accuracy, meaning how far an estimate is from the true, known data generation mechanism's parameter calculated as a difference. Empirical standard error is a measure of precision and is the variance of the statistical method's estimate (which may be biased or unbiased). If a method produces roughly the same estimate on many generated datasets of the same size (even if there is some bias), then it is considered precise. Mean squared error (MSE) is like empirical standard error but instead of being variance it is the mean squared difference between the estimate and the true parameter. MSE is inflated both by bias and by lack of precision by the method.

Coverage is the frequency that a method's interval estimate contains the true known value, and a high percentage is desirable. Average confidence/credible interval length is descriptive of the performance of the interval estimates of the methods and is self-explanatory. Power and type 1 error are rejection probabilities (under the null and alternative hypothesis) that the null hypothesis is rejected. Power is when the null hypothesis is rejected when the alternative hypothesis is true and is normally targeted to be 80%. Type 1 error is when the null hypothesis is rejected when it is true and is normally controlled to be less than 5%.

Finally, Monte Carlo errors are statistical errors introduced by aggregating across simulated datasets. Statistical methods produce estimates using data generated by known mechanisms in the simulation, and then these estimates are aggregated across the simulation. These estimates may be biased or unbiased but either way some random error is introduced when aggregating across the simulation. Note that this is not the variance of the estimator caused by the statistical

method, i.e. the empirical standard error which is method specific and estimated in the simulation. This is error introduced only in the aggregation process of combining estimates from the large number of simulated datasets. Monte Carlo error approaches zero as the number of simulations go up, and therefore can be controlled by the researcher.

To control errors introduced by the simulation process, we will require the following Monte Carlo standard errors to be as follows: For bias, we required that two standard errors be less than 0.01 on the generalized linear model's link function's scale, represented in the inequality (1.3.1) below. For coverage, type 1 error, and power, we required that two standard errors be less than 1.3%, represented in inequality (1.3.2) and (1.3.3). Note that two standard errors for a normally distributed random variable are associated with the critical value 1.96. This criterion allows the calculation of the minimum number of required simulations with formulas available in Morris et al.[35] Assuming the standard deviation of $\hat{\theta}$ on the link scale is less than or equal to 0.2 (or variance less than or equal to 0.04), we can solve this system of linear inequalities.

$$1.96 * Monte\ Carlo\ SE(Bias) = 1.96 * \sqrt{\frac{Var(\hat{\theta})}{n_{sims}}} = 1.96 * \sqrt{\frac{0.04}{n_{sims}}} \leq 0.01 \quad (1.3.1)$$

$$1.96 * Monte\ Carlo\ SE(Coverage\ and\ Type\ I\ error) = 1.96 * \sqrt{\frac{0.95 * 0.05}{n_{sims}}} \leq 0.013 \quad (1.3.2)$$

$$1.96 * Monte\ Carlo\ SE(Power) = 1.96 * \sqrt{\frac{0.8 * 0.2}{n_{sims}}} \leq 0.013 \quad (1.3.3)$$

When considering inequalities (1.3.1), (1.3.2), and (1.3.3), any value of n_{sims} that meets (1.3.3) will meet (1.3.1) and (1.3.2). Therefore, solving for n_{sims} in (1.3.3) derives the needed number minimum of simulations:

$$n_{sims} \geq 3637 \quad (1.3.4)$$

However, for the sake of simplicity, we determined that 3,600 was acceptable, yielding two Monte Carlo SEs (error introduced when aggregating simulation wide estimates) for power at

about 1.307%. 3,600 simulations control the Monte Carlo error such that two standard errors equal to 0.7% for Coverage/Type 1, 1.3% for power, and 0.0066 for bias. For percent bias, this is roughly 0.7% in the case that $\theta = 1$ and roughly 1% in the case when $\theta = 0.5$. Two standard Monte Carlo errors ensure that the aggregated simulation value of the performance measures (bias, empirical standard error, etc.) are likely within these amounts of the value that would be obtained with we used an infinite number of simulations.

The simulation was run in R version 4.0.2 (2020-06-22)[40], using a set of packages: broom 0.7.2[41], extraDistr 1.9.1[42], ggplot2 3.3.2[43], HDInterval 0.2.2[44], lubridate 1.7.9[45], magrittr 1.5[46], MASS 7.3-53[47], mvtnorm 1.1-1[48,49], parallel 4.0.3[40], Optmatch 0.9-17[50] (Note: this older version was needed to run on AWS servers), R2jags 0.7-1[51], remotes 2.4.0[52], RItools 0.3-1[53,54], rjags 4-10[55], rstream 1.3.6[56], sandwich 3.0-1[57–59], and tidyverse 1.3.0[60]. The parallel program utilized socketed clusters in base R's parallel package. For reproducibility, please ensure these versions are used when the code is run. All code that may be shared is available at Github (**Appendix B**).

When parallel programming is used, it is important to carefully manage the random number generation. As discussed in Morris et al.[35], streams of random numbers generated in each of the processing cores can overlap and cause undesired correlation between simulated data sets and unwanted behavior in statistical procedures that need random numbers (such as Bayesian Markov Chain Monte Carlo methods). For the current study, the random number generation method was set to use the method of L'Ecuyer[61,62] rather than R's default, the Mersenne twister[63]. This random generation mechanism has a seed vector of 6 (signed) integers and a period of around $2^{191} \approx 3.1 \times 10^{57}$. Each 'stream' is a subsequence of the period of length $2^{127} \approx 1.7 \times 10^{38}$ which is in turn divided into 'sub streams' of length $2^{76} \approx 7.6 \times 10^{22}$. These streams of random numbers are all statistically independent but reproducible, making them perfect for parallel programming. The simulation can be replicated using one primary seed and unique random sub streams for each core, ensuring the simulation avoids this possible issue.

In addition to random number generation, the simulation was split into computation jobs and run on two Amazon Web Services virtual instances (c5.18xlarge, each with 72 virtual CPUs). Specifically, the simulation was separated by parameter set and then broken down into four

computation jobs of 900 simulations (which combine into 3600 total simulations). Then, 50 simulations were run and timed for each parameter set to estimate the computation time required for each job. Once calculated, linear programming was used to determine an optimal job assignment to the processing cores to balance computation load and minimize the time required to complete the full simulation. The linear programming optimization was done in the Rglpk package[64] in R.

1.5 A REVIEW OF RELATED LITERATURE

To select the statistical and study design methods for historically controlled trials used in this dissertation, we conducted a non-systematic, snowballing literature review. The theoretical literature is still developing, and new techniques are still being published. During the planning stages of this dissertation, we discovered many candidates for the simulation but screened them out for a variety of reasons. The final selected methods are discussed in detail in section **1.6**. However, it is valuable to review some of the related literature, including methods which we did not select but may hold promise for further investigation in the future. The following discussion is a brief discussion of a subset of relevant literature.

Perhaps the first biostatistician to consider historically controlled trials seriously was Stewart Pocock. He developed a set of criteria in his 1976 article[65], which determines if candidate historical controls are high enough quality to be used as an external comparator arm in a clinical trial. In fact, his criteria are applied and discussed in **Chapter 4** when real trial data from completed pulmonary exacerbation trials are examined. In that same article, he suggested a simplified Bayesian statistics approach (applying Bayes' theorem discussed in section 2.2) with the historical data being treated as a conjugate normal prior that suggests a closed form posterior distribution in terms of the mean and variances of the historical and active trials. He goes on to describe how one might extend the ideas to binary and exponentially distributed data (such as our time to event outcome) and applies these ideas to real trial data. Since the publication of this article, more advanced techniques have been developed, but it is good to acknowledge his foundational work.

Power priors were one of the first Bayesian techniques proposed for historically controlled trials. They were first proposed and developed by Ibrahim et al.[66] and also use the Bayesian paradigm (see section 2.2) to incorporate the historical trial information. All the historical trial's information is summarized in a prior distribution. The power prior, $\pi(\theta|D_0, \alpha_0)$, is defined:

$$\pi(\theta|D_0, \alpha_0) \propto L(\theta|D_0)^{\alpha_0}\pi_0(\theta) \tag{1.3.5}$$

Where $L(\theta|D_0)$ is the likelihood of the parameter of interest given the historical data, D_0, α_0 is a scalar that controls the informational weight of the historical data's likelihood, and $\pi_0(\theta)$ is an

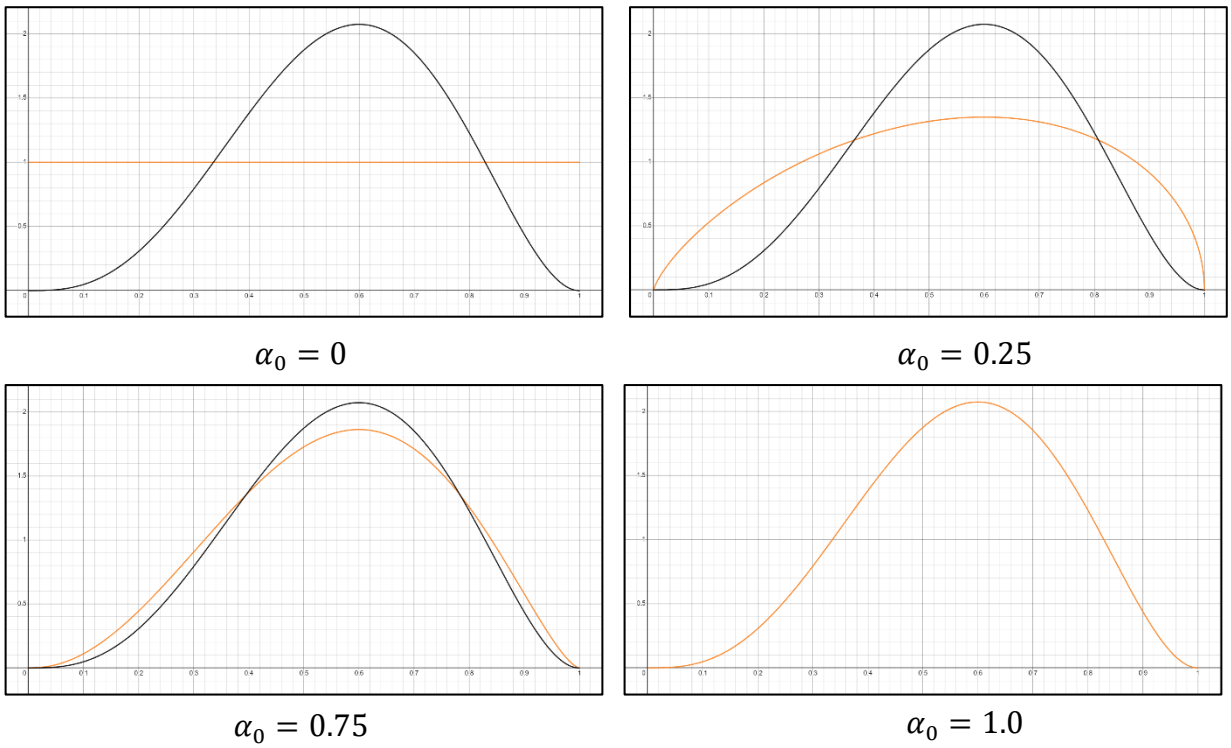


Figure 3. Power prior's α_0 parameter's influence visualized. Shown in black is an example of a historical data's likelihood scaled to be a probability density. Shown in orange is the same likelihood raised to the power parameter, α_0, and scaled to be a probability density. In this example, when $\alpha_0 = 0$, the historical data likelihood is discounted to a non-informative, uniform distribution that has no effect on the analysis. As α_0 goes to 1, the historical data's likelihood gains information weight until it is valued as heavily as the active trial data.

initial prior, typically a non-informative one. Note that $\propto$ means "proportional to," and is used to

avoid explicitly writing the Bayesian normalizing constant that ensures the posterior is a probability distribution. An example of the influence of α_0 is shown in **Figure 3**.

The formulation of the power prior is almost like conducting a Bayesian analysis on the historical data with a non-informative prior, generating a posterior, and then using that posterior as the prior for the new clinical trial. There is one important deviation from this description and that is that the historical likelihood function is raised to the power of α_0, a value between 0 and 1. This exponentiation causes the contributions of the historical data's likelihood function to range from contributing almost nothing (when α=0) to being treated as equal in informational value to the new trial (α=1). This can be expressed mathematically. Let D represent the active trial's data. Then the posterior distribution, $q(\theta)$, of the parameter of interest, θ, is as follows:

$$q(\theta|D, D_0, \alpha_0) \propto L(\theta|D)L(\theta|D_0)^{\alpha_0}\pi_0(\theta) \qquad (1.3.6)$$

Note that the components of expression 1.3.5 are contained in 1.3.6. The important concept is that the α_0 parameter can control the informational weight of the historical data. In fact, there are even methods for estimating the information weight that the historical power prior represents relative to the active trial (akin to comparing sample sizes to determine levels of informational precision) based on selected values or ranges of values of α_0. See Morita et al. for details.[67]

The ingenuity of this formulation is that it allows researchers to discount historical trial information in cases where there is concern that the historical trial and active trial differ in key aspects that could bias the primary analysis. In the case where the historical data quality is excellent (meets Pocock's criteria, gathered under the same protocol, etc.), then researchers can set α_0 equal to 1. However, in cases of concern, researchers can discount the historical data to an appropriate amount where it can still be valuable but not bias the primary analysis results.

Power priors leave room for criticism in that specification of the value of α_0 can appear subjective and leave room for investigator introduced bias. Ibrahim et al.[66] have gone on to propose using a hierarchical model that assigns a hyperprior to α to avoid this criticism, and

there are other ways to avoid this problem. Propensity score-based power priors are further developments of power priors, discussed in section **1.6**.

1.6 METHODS CONSIDERED BUT NOT IMPLEMENTED

Another Bayesian technique to incorporate data from historical trials that has been implemented in trial data[68] is the Meta Analytic Predictive prior or MAP. The Meta-Analytic-Predictive Prior follows a very similar idea to the random effects meta-analysis. It assumes there is a grand hyper model parameter that is the true population parameter of interest, indicated by Ψ. Like in the random effects meta-analysis, the trial specific parameters, θ, are assumed to be normally distributed around this true parameter with a between trial variance of τ. This is the assumed hierarchical model on which the MAP prior is built and what is presented in a paper by Neuenschwander[69]. The model has interesting connections with the normal random effects meta-analysis model. The meta-analytic-predictive prior (MAP) can be used for many types of outcomes, including normally distributed outcomes, binary outcomes, and Poisson outcomes. It should be noted that this approach shares many similarities with the commensurate prior approach we selected.

The MAP prior estimates the components of this hierarchical model, (Ψ, τ), using a fully Bayesian approach and the historical trial parameters, θ. It requires a specified outcome model that produces θ in the trials and a prior for τ, the between trial variance. From the historical study trials, it estimates the posterior distribution of Ψ. Then, the posterior *predictive* distribution of Ψ, an extension of the posterior that models the probability of observing a specific trial parameter, θ, rather than the probability of Ψ, is calculated. This is also where the name derives. The posterior predictive distribution represents our beliefs about the active trial's control group's outcome. Bayesian updating is used to estimate the control group's and experimental group's posteriors, and then a convolution of these distributions to describe the treatment effect is derived from which inferences are drawn. Additionally, the MAP technique was further refined by Schmidili through a robustification process[70]. What this process entails is to take the MAP prior and produce a mixture prior of a non-informative prior and the MAP prior. This lessens the

informational weight of the prior to an appropriate amount, thereby making it robust (in a sense) by heavily relying on the active trial.

The biggest restriction, which led us to screen out this technique, was the inability to adjust directly for covariate data. The MAP prior also works best with multiple historical trials and was developed under a meta-analytic framework. We want to focus on methods that work well even with a single historical trial. However, this framework has potential to be expanded upon.

The idea of a synthetic control group was alluded to by Franklin and Schneeweiss in their paper on real world evidence and by extension externally controlled trials. Therefore, we investigated proposals for creating synthetic control groups. Originally, a synthetic control method was developed by Abadie et al. in the context of economic research[71]. In their work, they used state-level data to create a synthetic state: a counterfactual version of California that did not enact a particular social policy. This synthetic comparator comprised of weighted averages of other states such that its demographics and key characteristics aligned with California's values for as many years prior to the enactment of the social policy as possible. The beauty of this work is that the authors offer strong theoretical justification and argue convincingly that the analysis can constitute casual inference.

Synthetic control/comparator groups are often statistically created in casual inference to estimate various counterfactual effects. Casual inference techniques are discussed in detail in the book *What If*, by Miguel Hernán and James Robins[72]. Although, the method of Abadie could likely be adapted to our goal with guidance from casual inference literature, we decided it was too complex a task and did not necessarily take advantage of the high-quality data gathered in previous trials to warrant further investigation.

In addition to analytic techniques, trial designs were also considered, and one approach made the final cut and is discussed in **Chapter 3**. However, there were other study design level approaches we considered. One such technique was a Bayesian Adaptive design proposed by Williamson et al.[72]. This trial design used a randomized, response-adaptive design that seeks to maximize the total number of patient successes in the trial while ensuring a minimum number of patients are

recruited to each treatment arm. The modeling approach uses Bayesian inference procedure and is closely related to what is known as a finite-horizon Bayesian Bernoulli two-armed bandit problem. Like the other Bayesian methods, historical information can be incorporated into the procedure using priors in the planning stages. Then, the adaptive trial design can be conducted, randomizing participants are they are recruited and obtaining optimal operating characteristics. This design was intriguing but had requirements that could not be met in our setting. The worst violation was the requirement that the outcome of treatment had to be known before study arm assignment (treatment or control arm) of the next participant. For pulmonary exacerbations, this can never be met in a regular trial.

The final method to review is one proposed by Chen et al. called the propensity score-integrated composite likelihood approach for augmenting the control arm of RCTs by incorporating real-world data[73]. First, a propensity score of being in the active trial is estimated, and the participants (both from previous trials and active participants) are stratified by this metric. In each stratum, a composite likelihood function is specified and used to appropriately weight the participant data derived from previous trials. Interestingly, this directly models the possible differences between the active and historical trials in a single step and allows for frequentist-like inferences to be made in these scenarios. Ultimately, we decided not to pursue this method due to its complexity and lack of available scripts or software to adapt to our purposes. However, this technique may be one to investigate in future research.

1.7 DISSERTATION STRUCTURE

In this dissertation, the primary goal is to review the selected techniques, evaluate them in the CF research setting, and prepare for a future therapeutic trial that features historical controls. These robust and practical statistical analysis methods and study designs are described and assessed in Chapter 2 and Chapter 3. In Chapter 4, the appropriateness and quality of existing randomized controlled trial data used for a historically controlled trial in pulmonary exacerbations in CF will be presented. We will demonstrate that a high quality historically controlled trial could have been conducted for the assessment of the effect of azithromycin in reducing pulmonary exacerbations among children with newly acquired *Pseudomonas aeruginosa (Pa)*. Finally, we will conclude with future research directions and limitations in Chapter 5.

Chapter 2. ANALYTIC METHODS FOR CLINICAL TRIALS THAT UTILIZE HISTORICAL CONTROLS

2.1 INVERSE PROBABILITY WEIGHTING

Randomized controlled trials (RCTs) are the gold standard in casual inference for estimating treatment effects because of crucial factors: 1) treatment arm assignment is determined through a random mechanism, providing a strong case for exchangeability of the treatment and comparison arms. 2) Strict protocols ensure that the observed treatment effect is consistent, meaning exposure (treatment) is specific and well-defined and ensuring the casual path of interest is being examined 3) All levels of covariates can be observed in both groups due to strict inclusion and exclusion criteria. When focusing on exchangeability in a two-arm RCT, a participant's probability of being assigned to the treatment arm is typically 50%. By controlling this probability and with a large enough sample size, the two arms will be balanced in terms of observed covariates, unobserved covariates, and expected risk of the outcome in the absence of a treatment effect. When the causal treatment is estimated comparing the trial arms, randomization removes any potential associations between the exposure (treatment of interest) and risk factors for the outcome (potential confounders).

Observational studies lack the control of randomized controlled trials; however, in the words of Cochran (1965), "the objective [of observational studies] is to elucidate cause-and-effect relationships…[when] it is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover, or to assign subjects at random to different procedures."[74] It is crucial to identify statistical techniques to mimic scientifically rigorous properties of the randomized controlled trial in observational data. In 1983, Rosenbaum and Rubin, published a seminal paper titled, *The central role of the propensity score in observational studies for causal effects*[75], that established a statistical framework to remove bias due to observed covariates from casual effect estimates in observational studies. The work was better summarized in Rosenbaum's book, the *design of*

observational studies[76]. A propensity score is an example of a balancing score. Balancing scores are the first key concept in Rosenbaum's work and have the following theoretical definition:

$$x \perp\!\!\!\perp z \mid e(x) \tag{2.1.1}$$

Where x are the observed covariates, z is the treatment assignment, and $e(x)$ is the balancing (i.e. propensity) score, typically estimated with logistic regression. Recall that the nomenclature was developed for use in observational studies, and in practice it is not always possible to estimate an acceptable balancing score. However, 2.1.1 is a testable assumption. In the case of historically controlled trials, the "treatment assignment" is an indicator of whether the data are obtained from the historical ($z = 0$) or the active ($z = 1$) trial, but the theory is just as applicable. The propensity score is the coarsest balancing score because it is a single value. In contrast, adjusting for all observed covariates is the finest balancing score because it could consist of many variables on which to balance (which has disadvantages).

The second key concept is that of "strongly ignorable treatment assignment given covariates, x." Strongly ignorable treatment assignment means that all of these are independent of the treatment assignment: a) the observed treatment response, b) the unobserved treatment response, and c) unobserved covariates. It is a strong assumption because it is almost never true in practice. However, that does not mean the assumption is not useful as will be explained below. Strongly ignorable treatment assignment is explicitly defined as follows:

$$Z \perp\!\!\!\perp (r_T, r_c, u) \mid x, \tag{2.1.2}$$

where Z is treatment assignment, (r_T, r_c) is the response (outcome) under treatment and under control, and u is the unobserved covariates. It should be noted that only one of (r_T, r_c) can be observed. Therefore, strongly ignorable treatment assignment is an untestable assumption about unobserved aspects of trial participants. Additionally, if 2.1.2 holds, then Rosenbaum et al. showed that the following also holds for the propensity score:

$$Z \perp\!\!\!\perp (r_T, r_c, u) \mid e(x) \tag{2.1.3}$$

Under strongly ignorable treatment assignment, at any value of the balancing score, an unbiased estimate of the casual treatment can be obtained with traditional statistical techniques.

Rosenbaum further explains that strongly ignorable treatment assignment is almost never true in observational studies (nor in historically controlled trials). However, it is still useful because it separates the scientific inquiry into two parts. The first is creating two treatment groups that are comparable on all observed covariates. The second is scrupulously and scientifically addressing the concern of violations of strong ignorability of treatment assignment. The second task can be controversial and requires assessment of evidence outside of the statistical model. In the case of historically controls, the second task of assuring that violations of strong ignorability are small and inconsequential largely depends on the quality of the historical data as assessed by Pocock's criteria (see section 1.5) and is investigated with sensitivity analyses.

The strengths of propensity score analyses[77] include the control of confounding through balancing while retaining the option of using a simple outcome model. The propensity score model can be as complex (over-parameterized) as desired even if the outcome model for the primary analysis is simple. Knowledge of the form of associations between the outcome and confounders is not required to balance them, which starkly contrasts with direct adjustment in the regression model (analogous to using the 'finest' balancing score). Also, the outcome is not used in the propensity score model, meaning that investigators do not accidentally induce bias when they use these methods. Some also consider these methods as a dimension reduction technique for the primary model[77]. Rosenbaum and Rubin[25] showed that the results from both full variable adjustment and using the propensity score should lead to the same conclusions. Other authors have noted that this smaller model may allow the investigator to perform diagnostic checks on the fit of the model more reliably than if there were many covariates included in the model.[78]

There are many limitations of propensity scores; therefore, crucial caveats will be summarized into a few ideas[77]. The greatest of these is that propensity score analyses only balance on the measured covariates. They may also balance on unmeasured covariates, but this is only directly attempted in randomized trials through the randomization process. Next, like most statistical techniques, large samples improve performance while missing data and small samples can lead

to a poor propensity score model. In the casual inference framework, matching on observed covariates may open backdoor paths of association between unrelated covariates and cause bias (see Hernán and Robins[72], pg. 89, for more details). Therefore, in the planning stages of a study, the form of the outcome model must be determined with consideration for causal pathways, typically with the use of directed acyclic graphs.

The propensity score can be used in many ways, and Austin (2011) describes how four methods can be utilitized[79]. Propensity score matching creates matched sets of the treatment groups or other binary groups. After the matched sample is formed, the primary outcome analysis can be conducted without concern for bias from the observed covariates. Stratification on the propensity score is another method, dividing participants into equal size strata defined by the propensity score. Although residual confounding may still exist, Rosenbaum and Rubin showed that this technique can eliminate 90% of the bias due to confounding in observed covariates[80]. The propensity score can also simply be used as a covariate that is adjusted for in the regression analysis. This approach also controls bias but requires specification of a model relating the propensity score to the outcome. The final technique, inverse probability weighting using the propensity score, creates "a synthetic sample in which the distribution of measured baseline covariates is independent of treatment assignment[79]" that is used in the outcome analysis.

Inverse probability weighting is a commonly used propensity score method for historically controlled trials. The propensity score for membership in the active trial can be estimated using logistic regression or other predictive modeling approach like machine learning. Then, the primary analysis can be estimated with weighted maximum likelihood estimation using inverse probability weights, described by Austin[81] and defined as below.

$$\frac{Z}{e(x)} + \frac{1 - Z}{1 - e(x)} \tag{2.1.4}$$

Recall that $z = 0$ for a historical trial participant and $z = 1$ for an active trial participant. Therefore, the expression in 2.1.4 takes the value of $\frac{1}{e(x)}$ for active trial participants and $\frac{1}{1-e(x)}$ for historical trial participants. Using weighted maximum likelihood, the primary analysis is

conducted as if two synthetic samples were used for the trial arms. The control arm would be treated as a sample from a synthetic historical population with a specific covariate distribution derived from the active trial. Then, the treatment arm would be treated as a sample from a synthetic active trial with that same specific covariate distribution (which is approximately the same as the actual active trial's covariate distribution). In essence, participants are upweighted or down weighted so that the two trial arms have roughly the same distribution in expectation on all observed covariates.

An important consideration when using inverse probability weighting is that variables with weak associations with the outcome but strong associations with trial membership can decrease accuracy in inverse probability weighting.[82] Care must be taken when determining the propensity model and such variables sometimes must be excluded. We discuss this in more detail in **section 4.2** in the application to real trial data. For the simulation, the covariates are equally associated with trial membership and the outcome and the forementioned decrease in accuracy is avoided.

In the simulation, analytic datasets were generated that contained an experimental arm, a concurrent control arm half the size of what it would be in a traditional trial, and set of historical controls per scenario described in **section 1.4**. A treatment indicator, E_i, historical control indicator, H_i, two binary covariates, X_1,X_2, and the outcome summarized as a time to event or study end variable and event or right-censored indicator as defined in section 1.4. The propensity score was estimated using logistic regression adjusting for covariates X_1,X_2 and modeling active/historical trial membership as the outcome. All data, including simulated active trial and historical trial participants, was used when fitting this propensity model:

$$Study_i \sim Bernoulli(p_i) \tag{2.1.6}$$

$$logit(p_i) \sim \omega + \gamma_1 X_{1,i} + \gamma_2 X_{2,i} \tag{2.1.5}$$

Where ω (the intercept), γ_1, and γ_2 are parameters on the logit scale, modeling the probability, p_i, of being in the active trial ($Study_i$=1) vs. the historical trial ($Study_i$=0). Using the notation in chapter 1.4, note that $Study_i = 1 - H_i$. These symbols were used to distinguish the propensity score model from the outcome model. The likelihood function is derived from the

Bernoulli distribution. Inverse probability weights were calculated using the expression in 2.1.4. Then, a weighted generalized linear model using the Poisson likelihood was used to estimate the treatment effect. The following outcome model was used with a log link function:

$$Y_i \sim Poisson(\mu_i) \tag{2.1.7}$$

$$\log(\mu_i) \sim \alpha + \beta_E E_i + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \log(t_i) \tag{2.1.8}$$

Where $Y_i \in (0,1)$ indicates if the first pulmonary observation was observed in the study. E_i is the treatment indicator variable, and X_1,X_2 are the covariates. μ_i is the mean of a generalized linear model with Poisson likelihood, corresponding to the event rate (sometimes written as λ). The variable t_i is the minimum of time to event or time to the end of the study (right censoring). The parameters α, β_E, β_1, and β_2 are parameters in the linear model on the scale of the log-link function to be estimated.

By the simulation's design, the mean model in 2.1.8 is correctly specified because it perfectly aligns with the data generating mechanism. There is also strong evidence that the model presented in 2.1.7 and 2.1.8 suitably fits the real trial data that inspired the simulation. Model fit and appropriateness in the real trial data is examined in **Chapter 4**. Therefore, the treatment effect (β_E) and covariate (β_1, β_2) parameters can be estimated correctly. However, the standard errors of these statistics may not be correctly estimated because the outcome is assumed to be distributed as a Poisson random variable, placing assumptions on the form of the variance that may be incorrect. The event is either observed, $Y_i = 1$, or not, $Y_i = 0$, during the study period, and Y_i is not a standard count of events. To solve this problem and accurately estimate the standard errors, the Huber-White sandwich estimator was used. This estimator of the variance is unbiased even when the variance model is misspecified or when there is heteroskedasticity of unknown form[83]. We estimated traditional point estimates and confidence intervals.

Recall that the target estimand is β_E, and the inverse probability weight's performance under six characteristics was assessed. Only percent bias, power, and type I error are presented in this chapter. See Github (**Appendix B**) for additional figures examining the other performance measures of the methods.

In **Figure 4** and **Figure 5** below, percent bias of the treatment effect was measured and plotted. The two figure-level columns represent the case where there is no treatment effect (θ=1) vs. one when there is a treatment effect (θ=0.5). The figure-level rows represent the simulation parameters we wish to assess, which were numbered and discussed in section 1.4. These include (2) the covariate prevalence in the active trial, (3) the agreement between the historical and active trial participant covariate prevalence, (4) the covariance between the two covariates, (5) the strength of association between the covariates and the outcomes (i.e. the covariates' β values), and (6) the relative number of historical controls available.

The goal was to evaluate the effect of the treatment effect and of these 5 parameters (2-6) on the performance measures for inverse probability weighting. In each of figure 5 and 6's panels, labeled A-J, one simulation parameter is isolated and examined at each of its two possible values. Recall that in total, there were $2^6 = 64$ possible combinations of (1) the treatment effect and (2-6) the simulation parameter. Therefore, when the treatment effect is held constant (as is done in the columns), there are 32 remaining combinations of the simulation parameters. When a single parameter is examined, this set is further divided into 2 groups of 16 scenario combinations which correspond to all scenarios with the specified value of the simulation parameter.

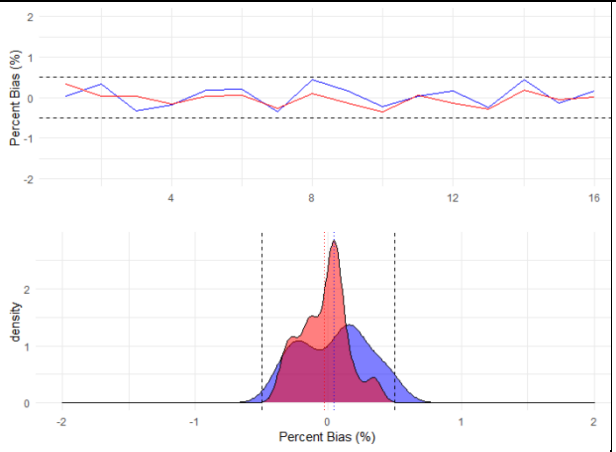
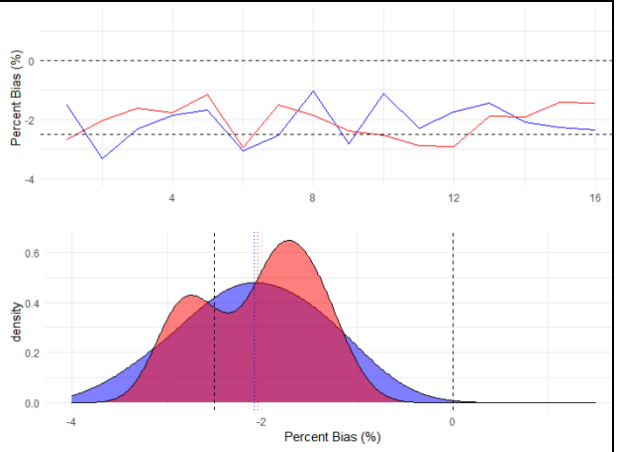
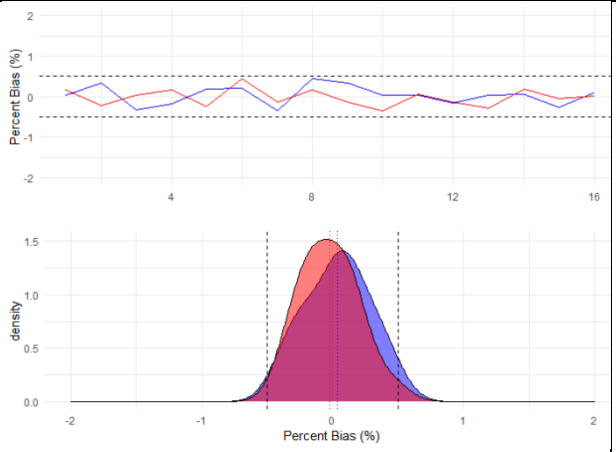
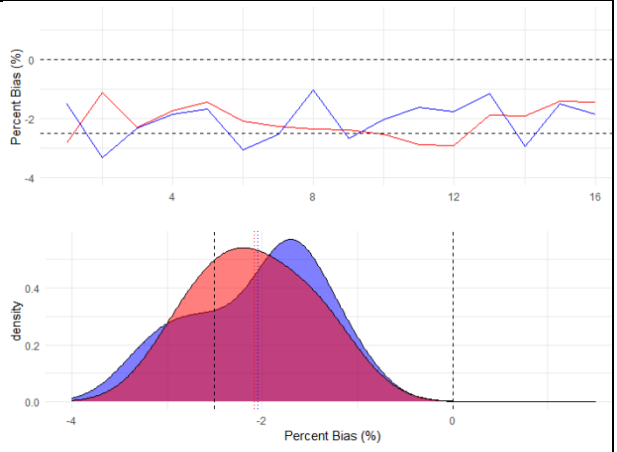
For example, in **Figure 4** panel A, the treatment effect HR is set at $\theta = 1$ and the simulation parameter that is being isolated and examined is (2), covariate prevalence in the active trial. The two possible values of this parameter are $(\mu_1, \mu_2) = (0.75, 0.20)$ shown in blue and $(\mu_1, \mu_2) = (0.85, 0.10)$ shown in red. Each of these groups include the 16 combinations of the other simulation-level parameters for which the prevalence of the covariates in the active trial is held at one of the two values.

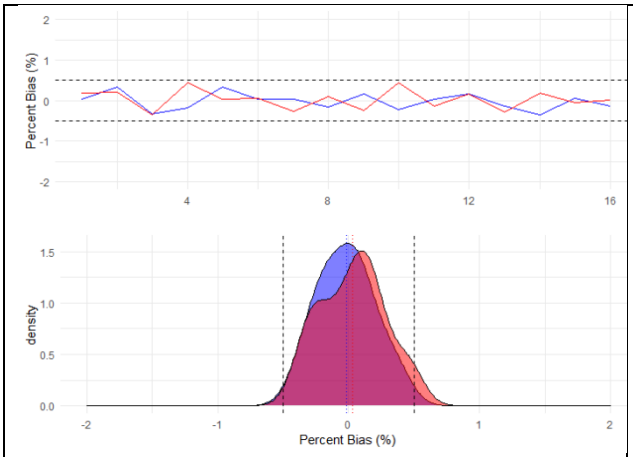
By varying one simulation parameter, the distribution of the performance measure over all the other parameters can be examined. The 16 scenarios are both directly plotted in the upper portion of each panel in a line plot and the distribution of the group is estimated with kernel density estimation techniques. It is possible to identify patterns by examining these figures to see if the

selected simulation parameter was influential on the analytic method's performance.

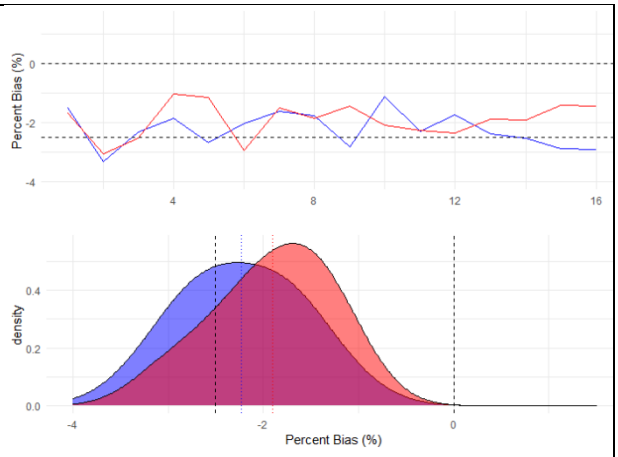
When looking at kernel density estimation, we identified potential patterns that could indicate that inverse probability weighting (IPW) does not work well in certain circumstances. Panels where there appears to be separation between the densities could provide insight into which characteristics affect the performance of IPW. For example, in panel F, the distribution of the red group seems to be slightly less biased because more of the density is closer to 0 as compared to the blue group. Bimodality (or sometimes multi-modality) occurs when a simulation parameter other than the one being examined is exerting influence. For example, in panel B, the bimodality observed in the red group is likely due to the influence (4) covariance/correlation between covariates. The line plots in the upper part of each panel directly compare the specific values for each of the 16 scenarios, when only the specified parameter is changed. Through careful examination of all the panels, influential and less important simulation parameters can be identified, allowing us to focus on essential parameters only.

Figure 4. Inverse probability weighting (IPW), percent bias by simulation characteristic
In the figures below, the panel columns represent the two simulated (1) treatment effects (null versus alternative hypothesis). The panel rows represent the simulation characteristics (2)-(6) that were assessed, outlined in **Table 1.1**. The red and blue represent the two considered values of the selected characteristic. The upper part of the figure is a line graph, which compares each of the 16 subsets of scenarios where the (1) treatment effect and one of the characteristics (2)-(6) are held constant. Within each line, scenarios are arranged from 1 to 16 based on a specific ordering with details given in Appendix B. The lower part is a kernel estimated density[84] of the 16 scenarios to identify patterns. Dotted lines show percent bias within $\pm 0.5\%$ or 0% & -2.5% for their respective columns.

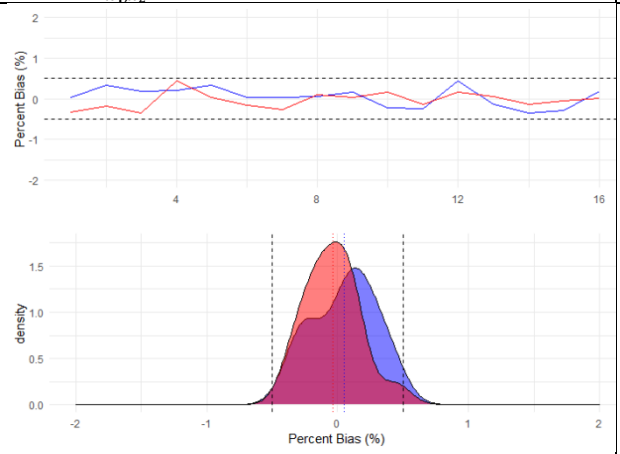
(1) Treatment effect HR, θ=1	(1) Treatment effect HR, θ=0.5
	
A) Varying (2) covariate prevalence in the active trial. **Blue:** $(\mu_1, \mu_2) = (0.75, 0.20)$ **Red:** $(\mu_1, \mu_2) = (0.85, 0.10)$	**B) Varying (2) covariate prevalence in the active trial.** **Blue:** $(\mu_1, \mu_2) = (0.75, 0.20)$ **Red:** $(\mu_1, \mu_2) = (0.85, 0.10)$
	
C) Varying (3) agreement of trials' covariate prevalence. **Blue:** $\delta = (-0.4, +0.2)$ **Red:** $\delta = (0, 0)$	**D) Varying (3) Agreement of trials' covariate prevalence.** **Blue:** $\delta = (-0.4, +0.2)$ **Red:** $\delta = (0, 0)$



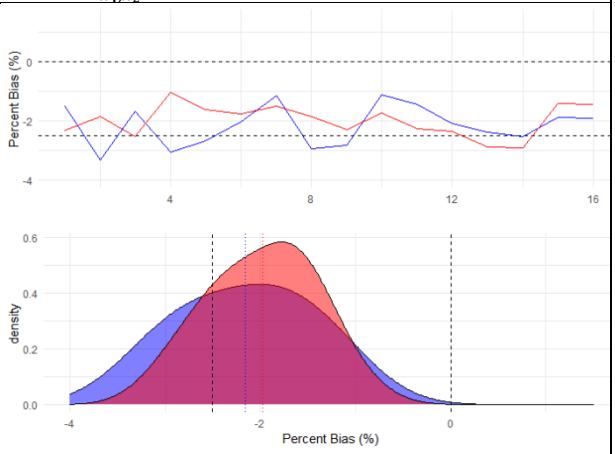
E) Varying (4) covariance/correlation between covariates
Blue: $Cov_{x_1,x_2} = 0.2$
Red: $Cov_{x_1,x_2} = 0.6$



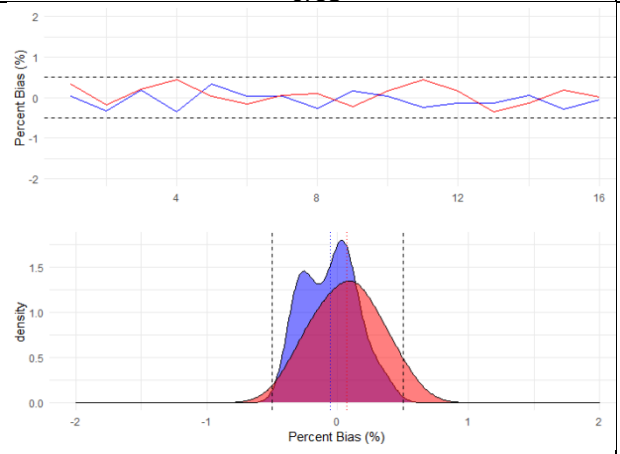
F) Varying (4) covariance/correlation between covariates
Blue: $Cov_{x_1,x_2} = 0.2$
Red: $Cov_{x_1,x_2} = 0.6$



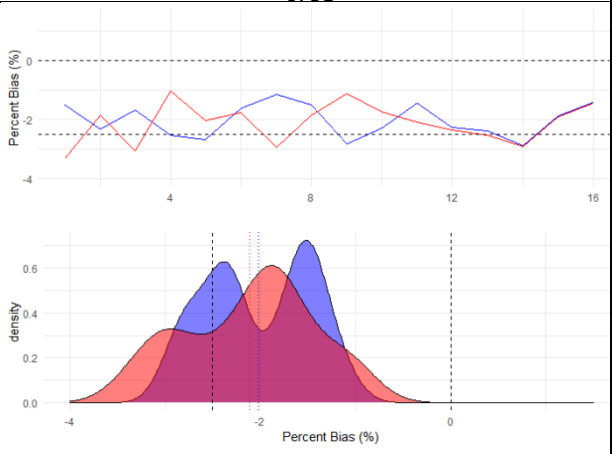
G) Varying (5) strength of association between covariates and outcome.
Blue: $\beta = \log(HR) = log\{1.10\}$
Red: $\beta = \log(HR) = log\{1.20\}$



H) Varying (5) strength of association between covariates and outcome.
Blue: $\beta = \log(HR) = log\{1.10\}$
Red: $\beta = \log(HR) = log\{1.20\}$



I) Varying (6) number of historical controls relative to the expected treatment arm size.
Blue: The same size
Red: Double the size



J) Varying (6) number of historical controls relative to the expected treatment arm size.
Blue: The same size
Red: Double the size

As seen in **Figure 4**, when the treatment had no effect ($\theta = 1$), inverse probability weighting (IPW) produced unbiased estimates of θ within 0.5% of the true value (panels A, C, E, G, I). Recall from **section 1.4** that the estimate for percent bias in each individual scenario when $\theta = 1$ can vary up to 0.7% or two Monte Carlo standard errors just because of estimation due to repeated simulation (rather than the method itself). All the percent bias estimates were tightly clustered around 0 and certainly within 0.5%. Even adding on the possibility of up to 0.7% from the Monte Carlo error, this easily puts the bias of IPW within 1% of the true value. No obvious additional patterns emerged in this set of panels, indicating that when no treatment effect was present, IPW's performance was consistent across a wide range of outcome frequencies and associations with covariates.

When there was a substantial treatment effect ($\theta = 0.5$), there was a small but systematic bias of 2% below the true value in all scenarios (panels B, D, F, H, J). However, the percent bias estimates always fell between -4% and 0%. Two Monte Carlo standard errors when $\theta = 0.5$ amounted to about 1% for these estimates. Therefore, 2% systematic bias is likely attributable to IPW in this simulation rather than to simulation induced variation, but the magnitude is small.

There appeared to be slightly less bias when the covariance (Σ) of X_1, X_2 was larger, indicating stronger agreement as seen in panel F. This also appeared to be the case when the covariate β values were more extreme, as seen in the red group for $\beta = \log(HR) = log\begin{Bmatrix}1.20\\0.65\end{Bmatrix}$ in panel H. There is bimodality seen in the density estimates in panel B's red group, when $(\mu_1, \mu_2) = (0.85,0.10)$, and panel J's blue group, when the number of the historical controls is the same as the treatment arm size. The cause is likely when high levels of binary agreement are seen in the covariates, X_1, X_2. As discussed in detail in Appendix A, certain combinations of δ and Cov_{x_1,x_2}, lead to the strong or weak levels (large or small Cramer's ϕ) of overall association between the binary covariates, X_1,X_2, and likely leads to the two observed modes. The β values magnify these strong or weak associations. A scientific interpretation is that the propensity score can be better fit in these situations. A stronger association reduces the proportion of random errors without changing any systematic error. It should be noted that in these panels with bimodality, the

difference between the two modes lies within 1%, or 2 Monte Carlo standard errors, therefore we cannot be sure this observation is not just an artifact of the simulation.

When examining power and type I error below in **Figure 5**, the structure of the figures and guidelines for interpretation are the same as for **Figure 4**. However, there are a few important differences to keep in mind when examining the panels. In the left panel column (when $\theta = 1$) the type I error is the performance measure being plotted. In the right panel column (when $\theta = 0.5$), the power is plotted. Together these are called the rejection percentages. The simulation was powered for 0.5% type I error and 80% power and the method should perform at least at these levels or better. To assist in interpretation, dotted lines are plotted at 0% and 5% for type I error and 80% and 100% for power, indicating target regions for the respective rejection percentages.

All scenarios when the treatment had no effect ($\theta = 1$) achieved a type 1 error of around 5% that was always within about 0.7% of this value, which corresponds to two Monte Carlo standard errors for each scenario. When there was no treatment effect, type 1 error was well controlled.

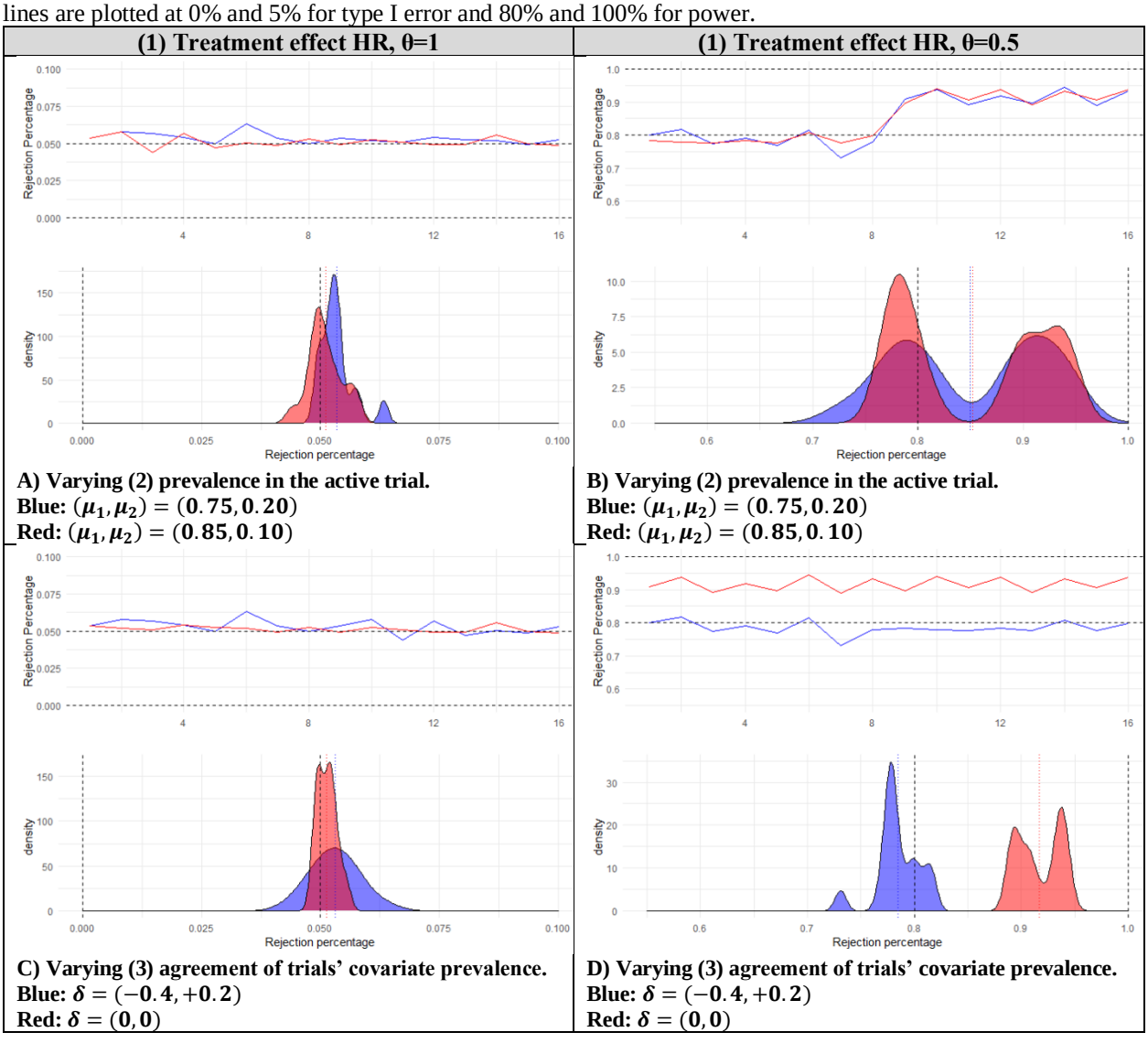
However, a clear trend was identified with regards to power (when the treatment effect HR, $\theta = 0.5$). Power suffered slightly (77%) when there were differences between the historical and active controls as seen in panel D, and the power to detect a treatment effect was higher (92%) when the trial covariate prevalence was the same. This phenomenon leads to the bimodality seen in panels B, F, H, and J. Another key observation was that having more available historical controls improved power in both situations where the trials' prevalence were the same and different as seen in panel J.

In summary, IPW produced an unbiased estimate and had power much greater than the requirement of 80% when the trials were similar in terms of covariate prevalence. However, when the trials differed in terms of covariate prevalence, the method's estimate became slightly biased away from the null and power suffered slightly, dipping below 80%. A more predictive propensity score may improve this method as suggested by theory and hinted at by

Figure 4's panels F and H. IPW could be recommended in situations where there is good evidence of agreement between the historical and active trial when examining descriptive statistics before the propensity score is even estimated. It appears that the more similar the trials (same target populations, approximately the same time, similar inclusion/exclusion criteria), the better IPW will perform.

Figure 5. Inverse probability weighting (IPW), rejection percentage (type 1 error for HR=1 and power for HR=0.5) by simulation characteristic

In the figures below, the panel columns represent the two simulated (1) treatment effects (null versus alternative hypothesis). The panel rows represent the additional simulation characteristics (2)-(6) that were assessed. Red and blue represent the two considered values of the selected characteristic. The upper part of the figure is a line graph, which compares each of the 16 subsets of scenarios where the (1) treatment effect and one of the characteristics (2)-(6) are held constant. Within each line, scenarios are arranged from 1 to 16 based on a specific ordering with details given in Appendix B. The lower part is a kernel estimated density[84] of the 16 scenarios to identify patterns. Dotted lines are plotted at 0% and 5% for type I error and 80% and 100% for power.

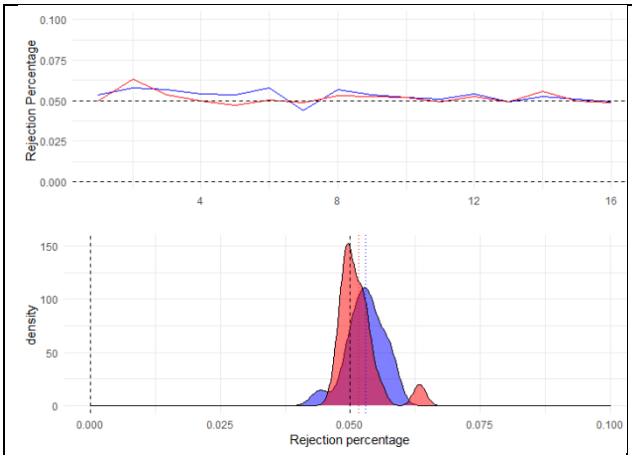


A) Varying (2) prevalence in the active trial.
Blue: $(\mu_1, \mu_2) = (0.75, 0.20)$
Red: $(\mu_1, \mu_2) = (0.85, 0.10)$

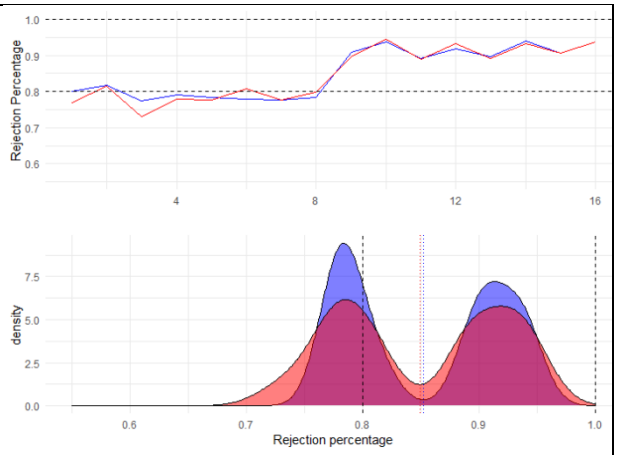
B) Varying (2) prevalence in the active trial.
Blue: $(\mu_1, \mu_2) = (0.75, 0.20)$
Red: $(\mu_1, \mu_2) = (0.85, 0.10)$

C) Varying (3) agreement of trials' covariate prevalence.
Blue: $\delta = (-0.4, +0.2)$
Red: $\delta = (0, 0)$

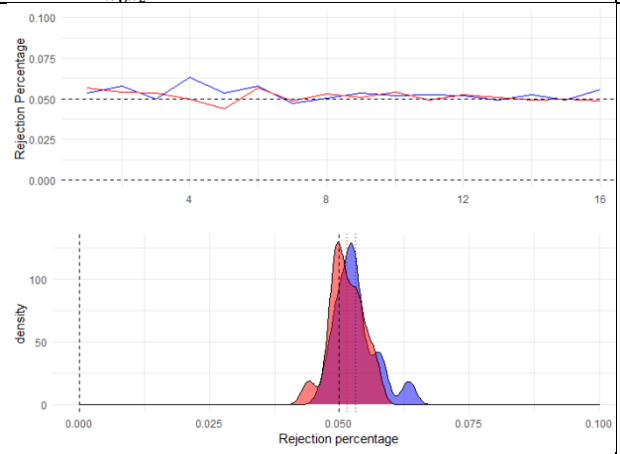
D) Varying (3) agreement of trials' covariate prevalence.
Blue: $\delta = (-0.4, +0.2)$
Red: $\delta = (0, 0)$



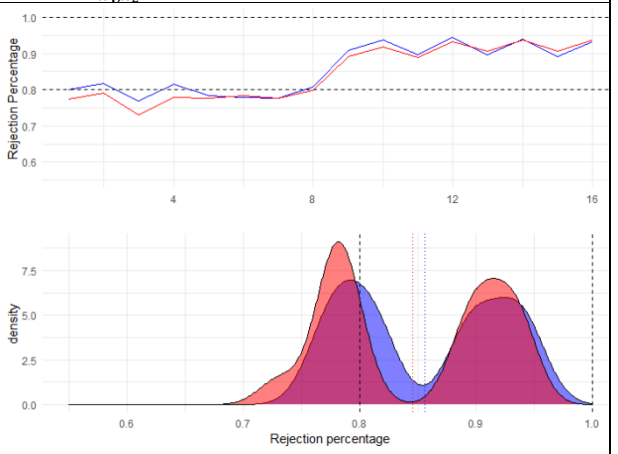
E) Varying (4) covariance/correlation between covariates
Blue: $Cov_{x_1,x_2} = 0.2$
Red: $Cov_{x_1,x_2} = 0.6$



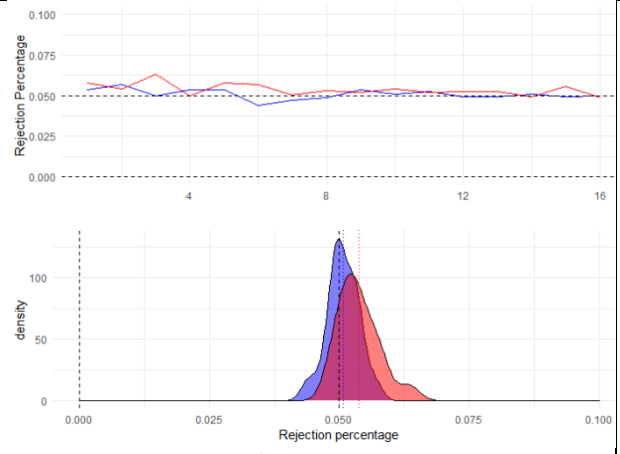
F) Varying (4) covariance/correlation between covariates
Blue: $Cov_{x_1,x_2} = 0.2$
Red: $Cov_{x_1,x_2} = 0.6$



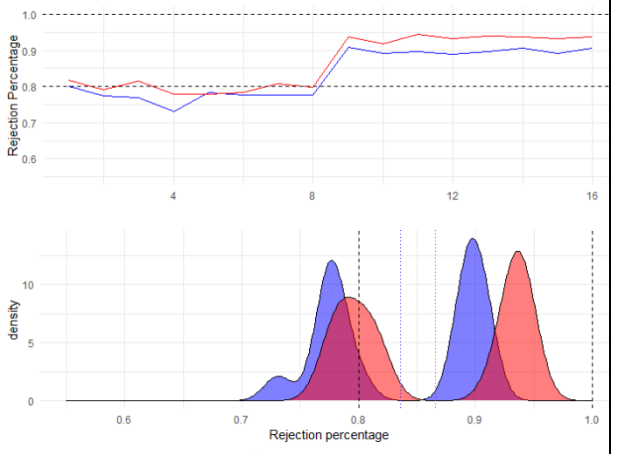
G) Varying (5) strength of association between covariates and outcome.
Blue: $\beta = \log(HR) = log\begin{Bmatrix}1.10\\1.30\end{Bmatrix}$
Red: $\beta = \log(HR) = log\begin{Bmatrix}1.20\\0.65\end{Bmatrix}$



H) Varying (5) strength of association between covariates and outcome.
Blue: $\beta = \log(HR) = log\begin{Bmatrix}1.10\\1.30\end{Bmatrix}$
Red: $\beta = \log(HR) = log\begin{Bmatrix}1.20\\0.65\end{Bmatrix}$



I) Varying (6) number of historical controls relative to the expected treatment arm size.
Blue: The same size
Red: Double the size



J) Varying (6) number of historical controls relative to the expected treatment arm size.
Blue: The same size
Red: Double the size

2.2 PROPENSITY SCORE-BASED POWER PRIORS

Bayesian methods are abundant in the literature on historically controlled trials. A strength they possess over the traditional, frequentist philosophy-based statistics is that they offer a mathematically formal way of incorporating external, scientific information. Frequentist methods typically rely only on the data at hand and make minimal scientific assumptions, but Bayesian philosophy-based statistics can use expert opinion and beliefs to improve statistical models. The ability to systematically incorporate known, external information is useful in cases where using historical information could improve trials as discussed in previous chapters. Some authors have even called Bayesian statistics "the statistics of small data[85]," because when data are scarce or difficult to obtain this statistical paradigm can compensate by incorporating information external to the sample. However, critics can argue that this paradigm's subjective nature can cause researchers to unwittingly introduce bias by poor prior specification or selection. Therefore, all modeling choices for these methods must be scientifically justified and defensible to appropriately take advantage of their strength in small data situations.

The statistical framework of Bayesian statistics is mathematically consistent and coherent, though it has strengths and weaknesses when compared with frequentist statistics[86]. In his book, A Comparison of the Bayesian and Frequentist Approaches to Estimation[87], Samaniego concludes that "neither the Bayesian nor the frequentist approach… is universally superior and the context of the statistical problem at hand should therefore guide one's decision about the approach that promises to give better performance."

The fundamental theorem in Bayesian statistics is Bayes' theorem derived from Bayes' rule, named after the reverend Thomas Bayes who started the theoretical development of this statistical philosophy in the 1700s[88]. Bayes' rule is a probability identity defined as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) * P(A)}{P(B)} \tag{2.2.1}$$

The probability of A given B is equal to the probability of B given A times the probability of A and divided by the probability of B. The rule allows P(A|B) to be rewritten in terms of some

event B and a general knowledge of the probability of A. This can be extended further into Bayes' theorem, which provides the basis for all of Bayesian statistics and inferences:

$$P(A_i|B) = \frac{P(B|A_i) * P(A_i)}{\sum_{j=1}^{n} P(B|A_j) * P(A_j)} \quad (2.2.2)$$

The theorem says that when a sample space is partitioned into n disjoint events, A_j, then the probability of one event, A_i, given event B, can be calculated by 2.2.2. This theorem is powerful and useful because of how it can be applied to a statistical model's parameter space given data:

$$p(\theta|X) = \frac{p(X|\theta) * p(\theta)}{p(X)} = \frac{p(X|\theta) * p(\theta)}{\int_{\Theta} p(X|\theta) * p(\theta)\, d\theta} \quad (2.2.3)$$

Where θ is the statistical model's parameter space and X is the observed data. Note that the integral is analogous to the summation in 2.2.2 except for continuous events rather than discrete ones. This formulation provides probabilities about the statistical model's parameters, described in a probability distribution, and derived from the data's probability, $p(X)$, and an initial probability distribution, $p(\theta)$, of the model parameter space (known as a "prior").

The prior is the reason Bayesian methods are useful for incorporating historical information or any other sort of external information. Researchers construct priors in specific ways to mathematically represent external information such as historical data, expert opinion, or other scientifically justified beliefs about the model parameters. In fact, there are many possible prior distributions that can be used in a single analysis and careful selection of an appropriate one is a critical duty of the researcher.

The data's probability is derived from the statistical model averaged over all possible parameters as seen in the integral, $\int_{\Theta} p(X|\theta) * p(\theta)\, d\theta$. Finally, a posterior distribution, $p(\theta|X)$, is calculated from which inferences about the model parameters are made given the observed data. From the posterior distribution, we can derive point estimates and interval estimates for model parameters. The point estimate we used in this simulation was the posterior mode, or the

parameter value with the largest probability density. For interval estimates, we used the highest density credible interval. This interval is defined as the narrowest interval in the posterior distribution that contains 95% of the probability. This contrasts with frequentist confidence intervals in that credible intervals do not need to be symmetric and are interpreted markedly differently. However, we will compare them and their corresponding precision as one of the factors when contrasting the statistical methods. For details on comparing Bayesian and Frequentist methods, please see Samaniego's book[87] for a discussion of point estimation as well as the article by Neath and Langenfeld[89] for a discussion on interval estimation.

The likelihood function is a grand product of the probability of each data point under the statistical model for an input set of parameters. Bayesian statistics uses the likelihood function to serve as the basis for the probability of the data by normalizing it so that it is a probability distribution and averaging it over all the entire parameter space.

Propensity score-based power priors fall under this framework, which requires a statistical model and the associated likelihood function. The prior in the Bayesian statistical analysis takes the form of a power prior, which was discussed in section 1.5, and can be summarized:

$$\pi(\theta|D_0, \alpha_0) \propto L(\theta|D_0)^{\alpha_0}\pi_0(\theta) \tag{2.2.4}$$

Where θ are the model parameters (maybe be a vector of multiple parameters), $\pi(\theta|D_0, \alpha_0)$ is the power prior, $L(\theta|D_0)$ is the likelihood function, D_0 is the historical trial data, α_0 is the power parameter, and $\pi_0(\theta)$ is typically a non-informative, initial prior for the model parameters, θ. Then, the posterior is derived using the Bayesian procedure described mathematically in 2.2.3.

$$q(\theta|D, D_0, \alpha_0) \propto L(\theta|D)L(\theta|D_0)^{\alpha_0}\pi_0(\theta)\,, \tag{2.2.5}$$

where $q(\theta|D, D_0, \alpha_0)$ is the posterior distribution from which we can make all model parameter inferences and D is the active trial's data. The parallels between 2.2.5 and 2.2.3 are not obvious at first, but each component in one has an analogous component in the other. First, $q(\theta|D, D_0, \alpha_0)$ is analogous to $p(\theta|X)$. $L(\theta|D)$, the likelihood of the data given a set of model

parameters, is analogous to $p(X|\theta)$, the probability of the data, if it is properly normalized to be a probability distribution. This normalization factor is analogous to $p(x)$ and its presence is indicated with the "proportional to" symbol, $\propto$. Finally, the prior $\pi(\theta|D_0, \alpha_0)$ from 2.2.4 is seen as $L(\theta|D_0)^{\alpha_0}\pi_0(\theta)$ in 2.2.5 and is analogous to $p(\theta)$ in 2.2.3, completing the Bayesian statistical procedure.

Lin et al.[90] further develop the power prior by prespecifying the power parameter, α_0, as a derivative of a propensity score. Specifically, a propensity score is estimated using logistic regression with covariates X_1,X_2 as defined in section 1.4. In fact, we used the same propensity score as used in the inverse probability weighting method of section 2.1. Then, the propensity score-based power prior (abbreviated 'psbpp'), $\pi_{psbpp}(\theta|D_0, \alpha)$, is defined as follows:

$$\pi_{psbpp}(\theta|D_0, \alpha) \propto \pi_0(\theta) \prod_{j \in D_0} L(\theta|d_i)^{\alpha_j} \tag{2.2.6}$$

Where $L(\theta|d_i)^{\alpha_j}$ is the likelihood function applied to a single historical participant's data raised to that participant's propensity score. The grand product of all the historical data multiplied by a non-informative, initial prior, $\pi_0(\theta)$, creates the prior used in the final Bayesian analysis:

$$q(\theta|D, D_0, \alpha_0) \propto L(\theta|D) * \pi_{psbpp}(\theta|D_0, \alpha) \tag{2.2.7}$$

In their article, the authors choose to keep the Bayesian model simple, letting the outcome's statistical model be a generalized linear model with two parameters as seen in 2.2.8 and 2.2.9. For the Bayesian procedure, this means that there must be a prior for both model parameters.

$$Y_i \sim Bernoulli\left(P_{Z_i}\right) \tag{2.2.8}$$

$$logit\left(P_{Z_i}\right) = b_0 + b_1 Z_i \tag{2.2.9}$$

The parameters of interest are b_0 and b_1; Z_i is the treatment indicator; and the likelihood function is specified as the Bernoulli likelihood. With this formulation, priors must be specified for b_0 and

b_1. For b_1, a non-informative prior is used that has little to no effect on the analysis (a normal distribution with variance equal to 1000). For the b_0 prior, 2.2.6 can be applied using the historical data except b_0 replaces θ:

$$\pi_{psbpp}(b_0|D_0, \alpha) \propto \pi_0(b_0) \prod_{j \in D_0} L(b_0|d_i)^{\alpha_j} \quad (2.2.10)$$

Note that the historical data does not offer any information on the treatment effect, b_1, only the intercept term, b_0. The historical data's likelihood informational weight in estimating the base event rate (the intercept) is larger for high propensity scores and smaller for low propensity scores by the nature of the power prior shown in **Figure 3**. A more complex version of 2.2.7 is used to create posterior distributions for the base event rate (intercept) and the treatment effect. It is more complex because it is a two-parameter posterior with a two parameter, joint likelihood and the non-informative prior for b_1 is included. For inferences, the marginal posterior distribution of b_1 is examined.

Given only an indicator for treatment, this simple, two parameter model does not allow adjustment for covariates. The authors avoid this issue by using matching on covariates using the propensity score. Matching is one technique which the author's argue create exchangeable trial arms in terms of all observable covariates.[90] In this context, exchangeability means that the risk of the event without treatment would be the same in both trial arms such that the only meaningful difference in risk of the event is the treatment itself. At least in terms of the observable covariates, both arms should have the same risk of outcome in the absence of treatment. The authors use the optmatch[50] to form pair-matches (1:1) or set-matches (1:2), minimizing the average distance between propensity scores of the matched individuals (active vs. historical trial participants).

In this simulation, we sought to follow Lin et al.'s example as accurately as possible. However, the matching step created an issue, the historical pool to match from had to be larger to allow for the creation of a well-matched set. This would mean that this method would have to have access to more simulated data than the inverse probability weighting or the commensurate prior

methods. We determined that the best way to address this issue was to generate a large historical control pool to be used in the matching, but the final number of historical controls used in the Bayesian model would be either the same size as the expected treatment arm or double it – the same as the other methods. Hypothetically, this might give Lin et al.'s method a small edge in that it could select a comparison group that was even closer in terms of covariates. However, fortunately, there are only two covariates, X_1,X_2. Even in the presence of this hypothetical small edge, the gains would be inconsequential. Also, for the Bayesian step, the analysis would be on a fair playing field as the other methods with the same number of participants.

For this simulation, we allowed the propensity score-based power prior method access to a historical pool that was four times larger than the expected size of the treatment arm. Recall, the active control arm is half the size of a traditional trial's control arm. We restricted the matched sets to be either the same size or double the size of the expected treatment arm. After pair-matching a selection of controls, the Bayesian analysis was conducted as written.

Another important note to address is the outcome model's specification as a Bernoulli random variable in Lin et al.'s method. The method requires a simple, parametric model. Poisson regression has been used throughout and was found acceptable as an alternative to survival analysis with a Cox proportional hazard model. Therefore, modeling the cumulative hazard rate ratio (as done in 2.2.8) is similar to using a simple Poisson regression. This is because the simulated hazard rate (and likely the hazard rate in the real trial data described in **Chapter 4**) is constant. A constant hazard rate implies that a ratio of the cumulative hazards at the end of the study period (the relative risk) will be equal to the hazard ratio:

$$\frac{H_{tr}(t)}{H_c(t)} = \frac{\int_0^t h_{tr}(u)du}{\int_0^t h_c(u)du} = \frac{\int_0^D U_{tr}dt}{\int_0^D U_c\,dt} = \frac{D * U_{tr}}{D * U_c} = \frac{U_{tr}}{U_c} = \frac{h_{tr}(t)}{h_c(t)} \qquad (2.2.11)$$

Cumulative hazard is defined as the integral of the hazard rate over the study period. If the treatment arm's hazard rate, $h_{tr}(t)$, is a constant, U_{tr}, and the control arm's hazard rate, $h_c(t)$, is constant, U_c, then the cumulative hazard risk ratio, $\frac{H_{tr}(t)}{H_c(t)}$, is equivalent to the hazard ratio, $\frac{h_{tr}(t)}{h_c(t)}$.

Therefore, a generalized linear model with a Bernoulli likelihood function and logit link function can produce estimates of the cumulative hazard rates at study end for the two comparison groups. Then, a ratio of the estimated cumulative hazard rates is analogous to the desired hazard ratio.

Specifically, after estimation of b_0 and b_1 in the statistical model 2.2.8 and 2.2.9, the statistical software calculates the probability of having the event in the study period for both the treatment arm (P_1) and the control arm (P_0) that includes both active and historical controls. In the survival analysis nomenclature, these values equal $F(D) = P(T \leq D)$, the cumulative distribution function of the time to event random variable at the study end, D. These values can be converted into the cumulative hazards rates at the study end, $H(D)$, with the following known formula[91]:

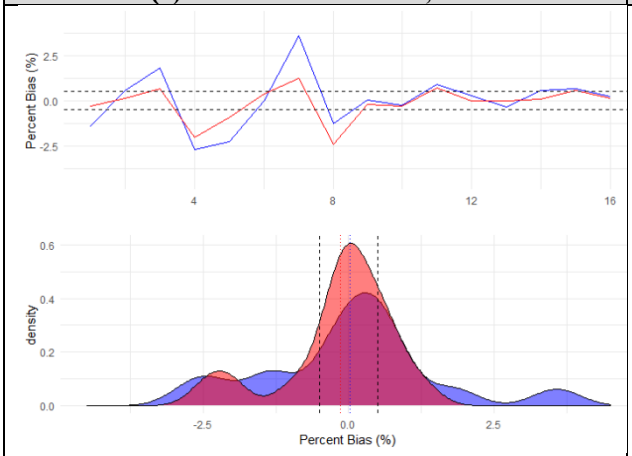
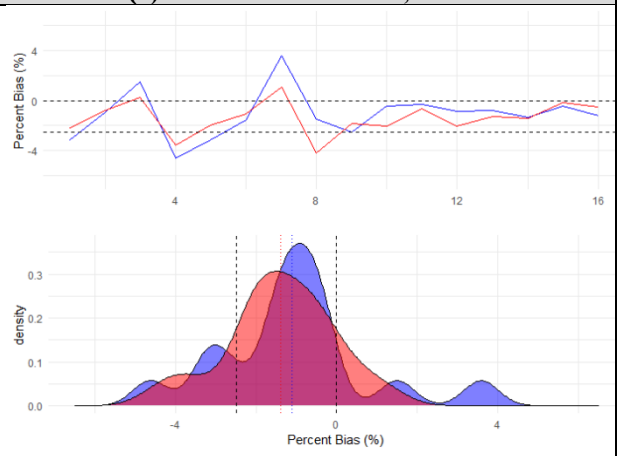
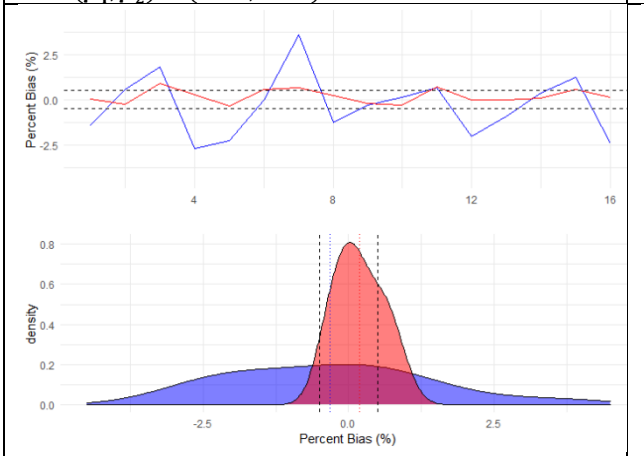
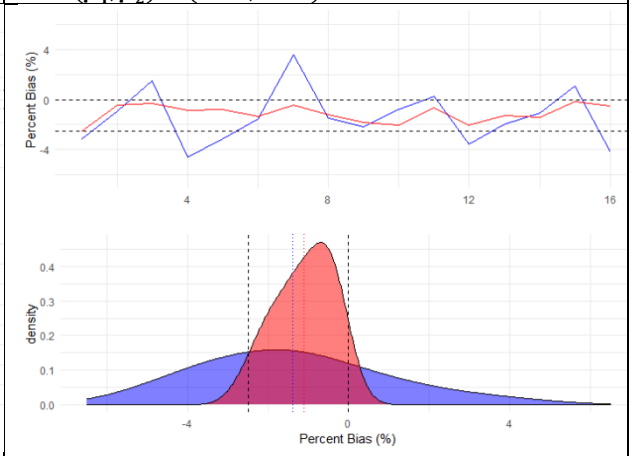
$$H(D) = -\ln\big(1 - F(D)\big) \tag{2.2.12}$$

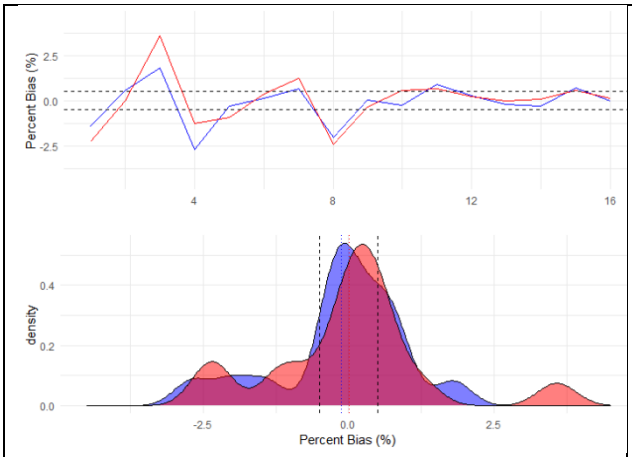
Then the ratio shown in 2.2.11 that corresponds to the hazard ratio can be directly calculated. For easy comparison with the other methods, the log is also taken of the estimated hazard ratio such that it is put on the same scale as the estimate produced by the other methods.

Using the method proposed by Lin et al. has a significant limitation in that the effect on performance measures of the matching step versus the Bayesian analysis cannot be separated. Further investigation should be conducted to address this limitation in future research.

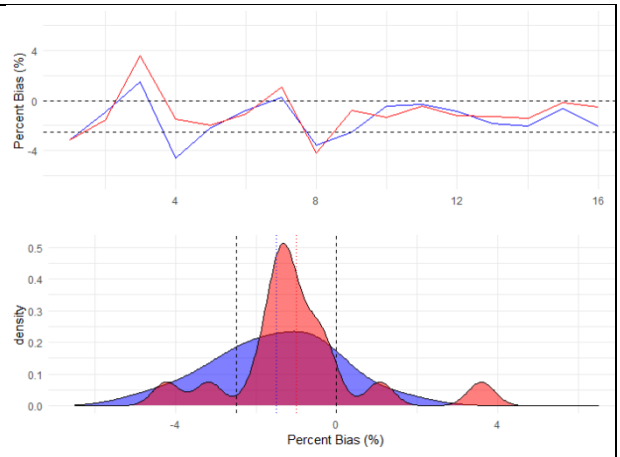
Note that the target estimand is the treatment effect, the log hazard ratio, and propensity score-based power priors' performance under six characteristics was assessed. Only percent bias, power, and type I error are presented in this chapter. See **Appendix B** for additional tables examining the other performance measures of the methods. **Figure 6** is interpreted in the same manner as was done for **Figure 4**, and the same goes for and **Figure 5** and **Figure 7**. Please pgs. 35-36 for guidance.

Figure 6. Propensity score-based power priors, percent bias by simulation characteristic
In the figures below, the panel columns represent the two simulated (1) treatment effects (null versus alternative hypothesis). The panel rows represent the simulation characteristics (2)-(6) that were assessed, outlined in **Table 1.1**. The red and blue represent the two considered values of the selected characteristic. The upper part of the figure is a line graph, which compares each of the 16 subsets of scenarios where the (1) treatment effect and one of the characteristics (2)-(6) are held constant. Within each line, scenarios are arranged from 1 to 16 based on a specific ordering with details given in Appendix B. The lower part is a kernel estimated density[84] of the 16 scenarios to identify patterns. Dotted lines show percent bias within $\pm 0.5\%$ or 0% & -2.1% for their respective columns.

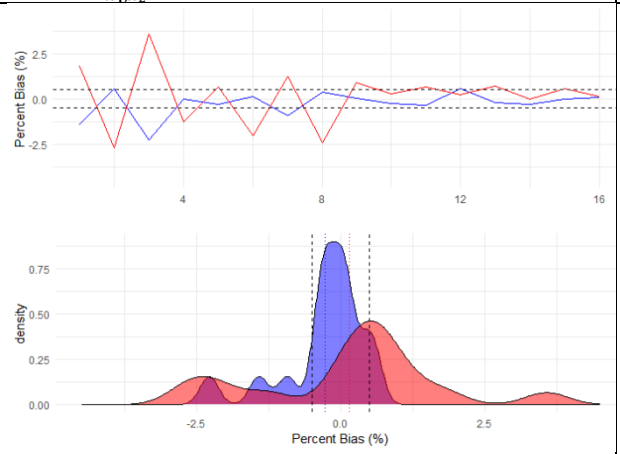
(1) Treatment effect HR, θ=1	(1) Treatment effect HR, θ=0.5
	
A) Varying (2) prevalence in the active trial. **Blue:** $(\mu_1, \mu_2) = (0.75, 0.20)$ **Red:** $(\mu_1, \mu_2) = (0.85, 0.10)$	**B) Varying (2) prevalence in the active trial.** **Blue:** $(\mu_1, \mu_2) = (0.75, 0.20)$ **Red:** $(\mu_1, \mu_2) = (0.85, 0.10)$
	
C) Varying (3) agreement of trials' covariate prevalence. **Blue:** $\delta = (-0.4, +0.2)$ **Red:** $\delta = (0, 0)$	**D) Varying (3) agreement of trials' covariate prevalence.** **Blue:** $\delta = (-0.4, +0.2)$ **Red:** $\delta = (0, 0)$



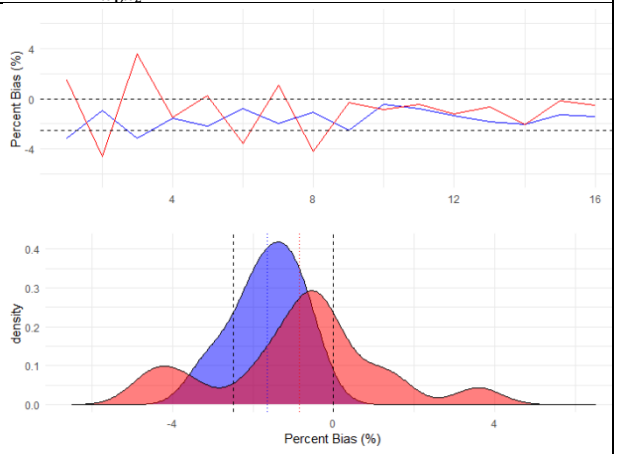
E) Varying (4) covariance/correlation between covariates
Blue: $Cov_{x_1,x_2} = 0.2$
Red: $Cov_{x_1,x_2} = 0.6$



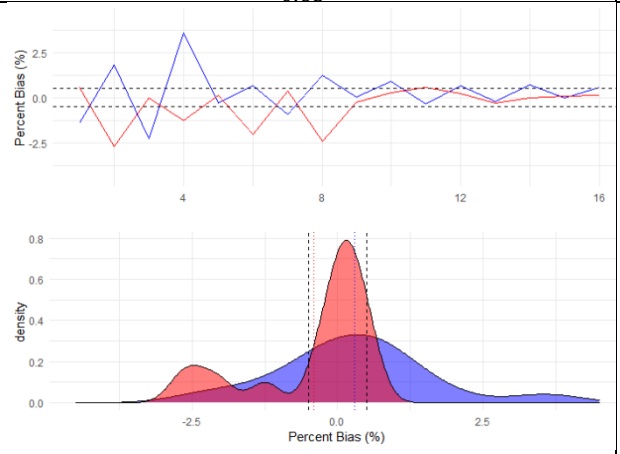
F) Varying (4) covariance/correlation between covariates
Blue: $Cov_{x_1,x_2} = 0.2$
Red: $Cov_{x_1,x_2} = 0.6$



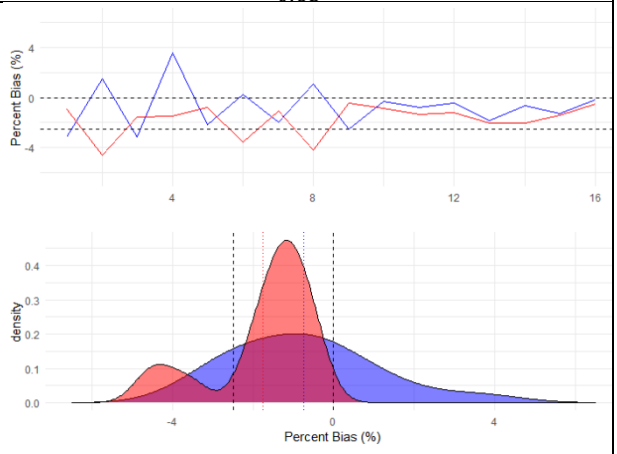
G) Varying (5) strength of association between covariates and outcome.
Blue: $\beta = \log(HR) = log\{1.10\}{1.30}$
Red: $\beta = \log(HR) = log\{1.20\}{0.65}$



H) Varying (5) strength of association between covariates and outcome.
Blue: $\beta = \log(HR) = log\{1.10\}{1.30}$
Red: $\beta = \log(HR) = log\{1.20\}{0.65}$



I) Varying (6) number of historical controls relative to the expected treatment arm size.
Blue: The same size
Red: Double the size



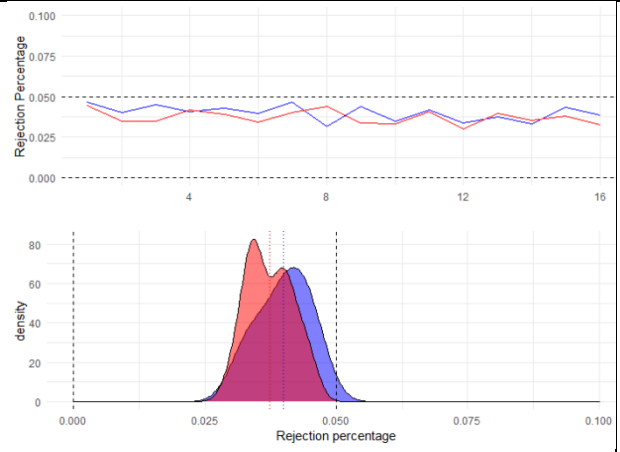
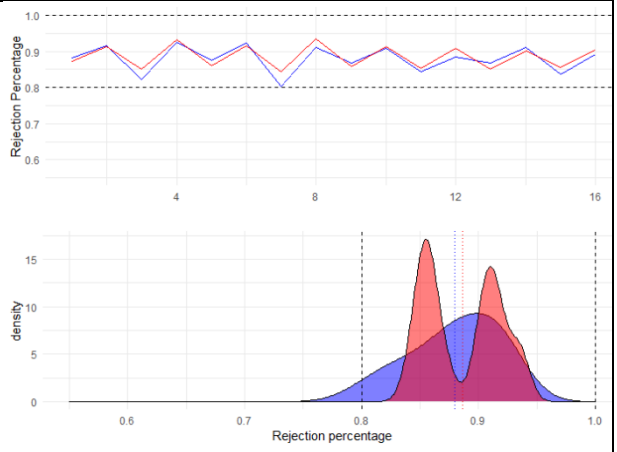
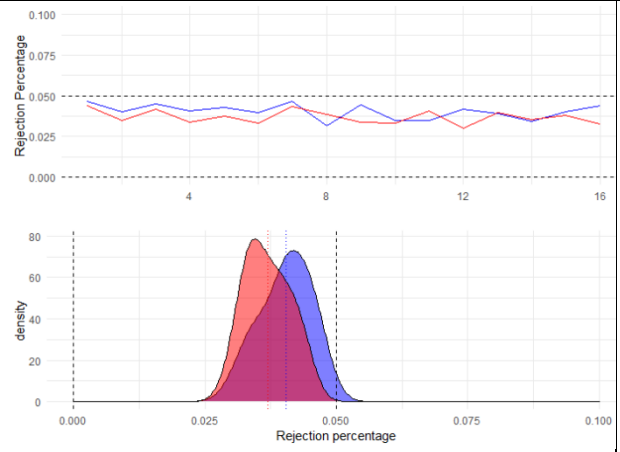
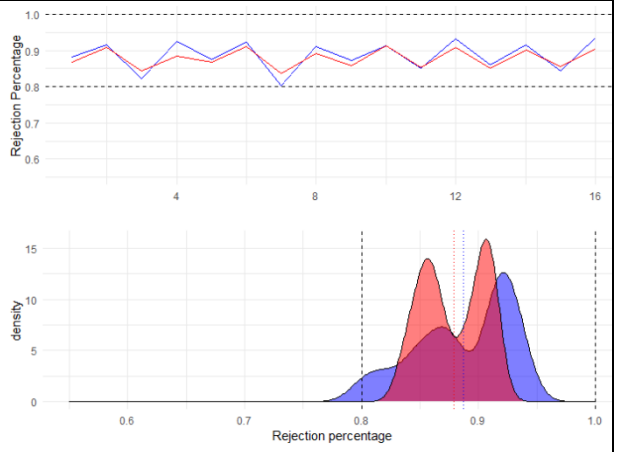
J) Varying (6) number of historical controls relative to the expected treatment arm size.
Blue: The same size
Red: Double the size

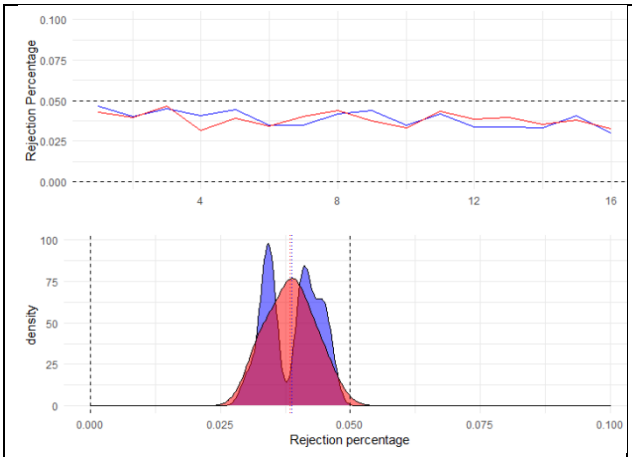
In contrast to the inverse probability weighting method, much greater variability was seen in the percent bias estimates in all the figure panels. The estimates ranged from ~ -4% to 4%, which is much larger than two Monte Carlo standard errors which were 0.7% and 2.1% for $\theta = 1$ and $\theta = 0.5$, respectively. Closer examination revealed that this was likely caused by scenarios in which the trial covariates were distributed identically. Looking at **Figure 6**'s panels C and D, when the trial covariate prevalence were different, a tighter range of percent bias estimates were observed. However, when they were the same, much more variability was introduced. This spread is reflected in all the panels likely due to this simulation characteristic. A similar observation was seen in panel F where having a higher covariance/correlation between covariates seemed to lead to less variability versus when the covariance/correlation was lower. It is also seen in panels I and J, where having more historical controls seemed to help reduce the variability.

Additional differences can be seen. When examining **Figure 6**'s panels G and H, the more extreme β values seemed to create slightly different distributions of the percent bias. As discussed with inverse probability weighting and in Appendix B, certain combinations of δ and Cov_{x_1,x_2}, lead to the strong or weak levels (large or small Cramer's ϕ) of overall association between the binary covariates, X_1,X_2. A better fit propensity score may improve the percent bias estimates in these cases.

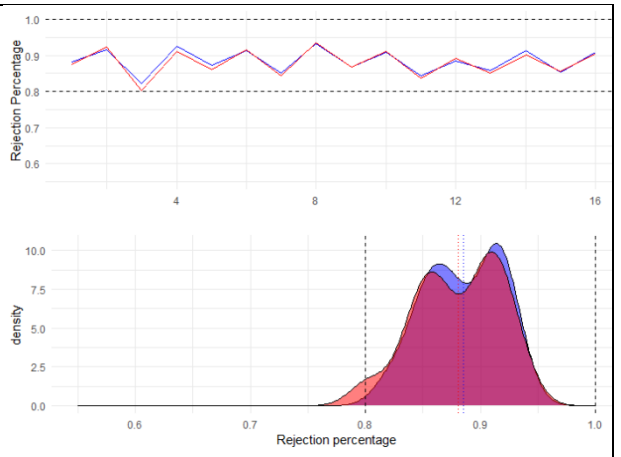
Figure 7. Propensity score-based power priors, rejection percentage (type 1 error for HR=1 and power for HR=0.5) by simulation characteristic

In the figures below, the panel columns represent the two simulated (1) treatment effects (null versus alternative hypothesis). The panel rows represent the simulation characteristics (2)-(6) that were assessed, outlined in **Table 1.1**. The red and blue represent the two considered values of the selected characteristic. The upper part of the figure is a line graph, which compares each of the 16 subsets of scenarios where the (1) treatment effect and one of the characteristics (2)-(6) are held constant. Within each line, scenarios are arranged from 1 to 16 based on a specific ordering with details given in Appendix B. The lower part is a kernel estimated density[84] of the 16 scenarios to identify patterns. Dotted lines are plotted at 0% and 5% for type I error and 80% and 100% for power.

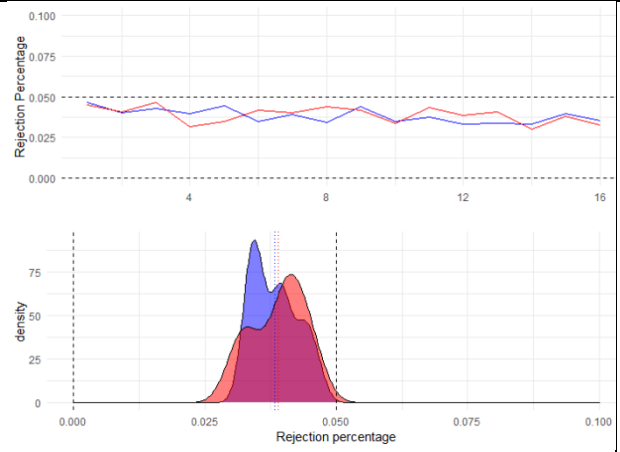
(1) Treatment effect HR, θ=1	**(1) Treatment effect HR, θ=0.5**
	
A) Varying (2) prevalence in the active trial. **Blue:** $(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) = (\mathbf{0.75}, \mathbf{0.20})$ **Red:** $(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) = (\mathbf{0.85}, \mathbf{0.10})$	**B) Varying (2) prevalence in the active trial.** **Blue:** $(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) = (\mathbf{0.75}, \mathbf{0.20})$ **Red:** $(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) = (\mathbf{0.85}, \mathbf{0.10})$
	
C) Varying (3) agreement of trials' covariate prevalence. **Blue:** $\boldsymbol{\delta} = (\mathbf{-0.4}, \mathbf{+0.2})$ **Red:** $\boldsymbol{\delta} = (\mathbf{0}, \mathbf{0})$	**D) Varying (3) agreement of trials' covariate prevalence.** **Blue:** $\boldsymbol{\delta} = (\mathbf{-0.4}, \mathbf{+0.2})$ **Red:** $\boldsymbol{\delta} = (\mathbf{0}, \mathbf{0})$



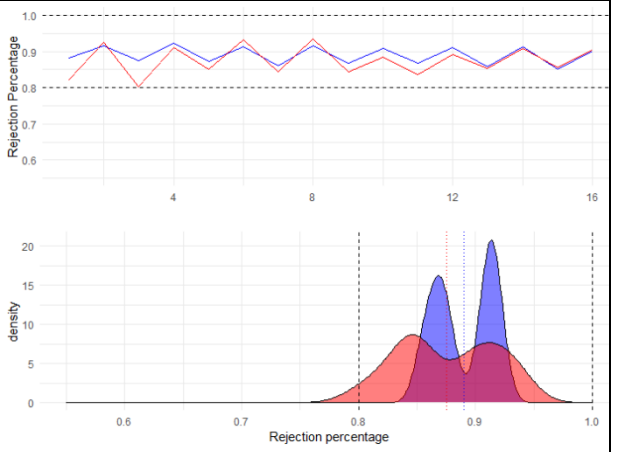
E) Varying (4) covariance/correlation between covariates
Blue: $Cov_{x_1,x_2} = 0.2$
Red: $Cov_{x_1,x_2} = 0.6$



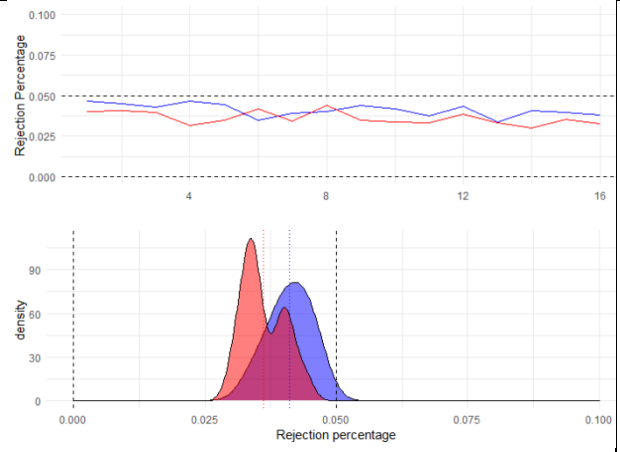
F) Varying (4) covariance/correlation between covariates
Blue: $Cov_{x_1,x_2} = 0.2$
Red: $Cov_{x_1,x_2} = 0.6$



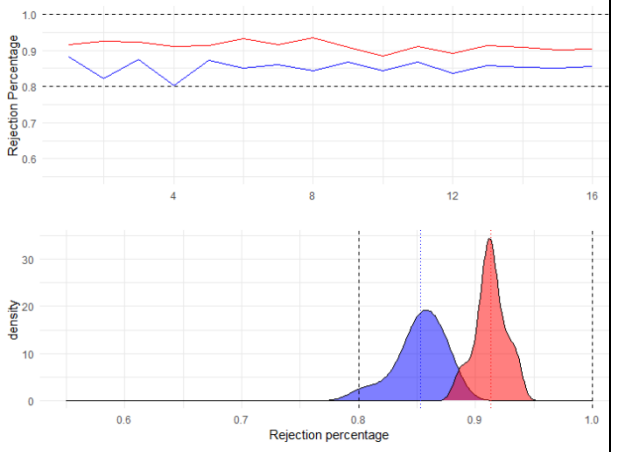
G) Varying (5) strength of association between covariates and outcome.
Blue: $\beta = \log(HR) = log\{1.10\}{1.30}$
Red: $\beta = \log(HR) = log\{1.20\}{0.65}$



H) Varying (5) strength of association between covariates and outcome.
Blue: $\beta = \log(HR) = log\{1.10\}{1.30}$
Red: $\beta = \log(HR) = log\{1.20\}{0.65}$



I) Varying (6) number of historical controls relative to the expected treatment arm size.
Blue: The same size
Red: Double the size



J) Varying (6) number of historical controls relative to the expected treatment arm size.
Blue: The same size
Red: Double the size

The target power for the computed sample size was 80% and was 5% for type I error. In the case when the treatment had no effect ($\theta = 1$), all scenarios had a type I error $< 5\%$, hovering around 3.75%. In the case where there was a treatment effect ($\theta = 0.5$), all scenarios had power greater than 80%, hovering around 87%. Having more historical controls clearly improved the power of this method, increasing it from roughly 85% to 92% as seen in **Figure 7**'s panel J. No other simulation characteristics significantly affected type 1 error or power.

2.3 COMMENSURATE PRIORS

In 2011, Brian Hobbs, Bradley Carlin, Sumithra Mandrekar, and Daniel Sargent proposed a Bayesian clinical trial design that incorporates historical information based on a measure of commensurability.[92] Their original approach used a hierarchical Bayesian model adapted from the traditional power prior approach introduced by Ibrahim and Chen[93] discussed earlier. The authors introduce the location commensurate power prior (LCPP). In a 2012 paper, the authors changed the model formulation and dropped the power prior component and provided an easier framework extendable to generalized linear models[94]. The result is the version of commensurate priors implemented in the current study. We will briefly review the LCPP model without going into extensive detail to understand the evolution of the concept of trial 'commensurability' and then closely examine the 2012 methodology used in the current study. A reason the old method is worth reviewing is because a theoretical relationship with power priors exists even with the revision and new method.

Location commensurate power priors were derived first from modified power priors (MPP) that addressed some of the concerns of the original power prior. Utilizing the power prior with an independent prior for α_0 raised some concerns that the formulation violates the likelihood principal[95] (the proposition that, given a statistical model, all the evidence in a sample relevant to model parameters is contained in the likelihood function) and did not produce marginal posteriors for α_0 that are proportional to products of familiar probability distributions[92]. Therefore, a modified power prior was proposed that normalizes the power prior and adds a proper prior for the power parameter:

$$\pi^{MPP}(\theta, \alpha_0 | D_0) \propto \frac{L(\theta|D_0)^{\alpha_0}\pi_0(\theta)}{\int L(\theta|D_0)^{\alpha_0}\pi_0(\theta)\, d\theta}\pi(\alpha_0) \qquad (2.3.1)$$

Where θ is a model parameter of interest, α_0 is a parameter that controls the amount of borrowing (note that π^{MPP} is a joint probability distribution), D_0 is the historical data, and π_0 is an initial, typically non-informative prior. The MPP was the foundation for the location commensurate power prior. The goal of the LCPP was to directly parameterize the commensurability of the historical and new data. Commensurability is defined as the similarity

of historical and active trial populations. It is left purposefully vague because the exchangeability of trial data is much more complex than what can be handled analytically (see Pocock for a greater discussion[65]). In the 2011 paper[92], the authors propose the following framework: Suppose that both the historical and current data depend on a common parameter, θ. Assume that the historical group and current group have different parameters, θ and θ_0. Assume that θ is normally distributed around θ_0 with a precision parameter, τ (Note that precision is simply the inverse of the variance). Then τ now parameterizes analytic commensurability and is referred to as the commensurability parameter. The 2011 paper emphasizes that as $\tau \to 0$, the historical information is discarded because the normal distribution's variance gets huge. This indicates that the historical trial and active trial have very different model parameters such as baseline hazard rates. As $\tau \to \infty$ full borrowing is performed because the normal distributions variance approaches 0, meaning that the historical and active trials have almost identical estimates for key model parameters such as a baseline hazard rate. This allows for the definition of the location commensurate power prior (LCPP), called that because commensurability is defined in terms of the *location* of θ and θ_0. Here is the expression that defines the LCPP:

$$\pi^{LCPP}(\theta, \alpha_0, \tau|D_0) \propto \int N\left(\theta\middle|\theta_0, \frac{1}{\tau}\right) * \frac{[L(\theta_0|D_0)]^{\alpha_0}}{\int [L(\theta_0|D_0)]^{\alpha_0} d\theta_0} d\theta_0 * Beta(\alpha_0|g(\tau), 1) * \pi(\tau) \quad (2.3.2)$$

The MPP in 2.3.1 includes an initial prior for the power prior α_0, $\pi(\alpha_0)$. In 2.3.2, this initial prior takes the form of a Beta distribution with the first shape parameter defined as a function of τ (the commensurability parameter). This makes sense because the amount of borrowing from the historical data (controlled by the power parameter α_0) is determined by the level of commensurability measured by τ. In 2.3.1, the historical data defines the modified power prior and is not included in the likelihood. In 2.3.2, the historical data again is part of the prior, although it is more difficult to see directly. Ignoring how τ changes the expression for a moment, the following expression is the relevant parts of 2.3.2 that align with the MPP from 2.3.1:

$$\frac{[L(\theta_0|D_0)]^{\alpha_0}}{\int [L(\theta_0|D_0)]^{\alpha_0} d\theta_0} * Beta(\alpha_0|g(\tau), 1) \quad (2.3.3)$$

Greater commensurability (agreement between the historical and active trial model parameters) leads to greater informational borrowing and less commensurability leads to no borrowing. In contrast to 2.3.1, a normal distribution is used for estimating the active trial's parameters, θ, centered as θ_0 with precision τ (note that $\sigma^2 = \frac{1}{\tau}$) . This τ parameter also requires an initial prior, $\pi(\tau)$. The parameter τ now directly models the commensurability as discussed above.

After receiving feedback and criticism about the LCPP introduced in 2011, Hobbs et al. changed the framework and proposed new methods for the estimation of τ. First, they consider the historical data to be part of the *likelihood* in the Bayesian formula rather than as a *prior*. The new framework makes many changes from (2.3.2). It does not include *any* power prior parameter (α_0) nor the usual normalized likelihood as seen in 2.3.3. Instead, the historical data is now captured in a likelihood function that relates to normally distributed data and only three priors are now required: a prior for the historical trial's mean, $p(\mu_0)$, and priors for the historical trial's and active trial's variances, $p(\sigma, \sigma_0)$. Additionally, the original paper considered a single parameter of interest, but the authors redefined θ to represent a vector as described in the next paragraph.

Consider a Gaussian-distributed outcome variable *y*, one historical trial and the active trial, and a model with parameters $\theta = (\lambda, \mu, \mu_0, \sigma^2, {\sigma_0}^2)$. λ is the treatment effect, μ, μ_0 are the mean outcomes from the active (μ) and historical trial (μ_0), and $\sigma^2, {\sigma_0}^2$ are the respective variances of those outcomes. τ is the precision of the normal distribution that connects the two means (a measure of commensurability). The commensurate prior can then be defined as follows:

$$q(\theta|\tau, y, y_0) \propto \underbrace{N\left(\mu\middle|\mu_0, \frac{1}{\tau}\right) p(\mu_0) p(\sigma, \sigma_0)}_{\text{Commensurate prior}} \prod_{j=1}^{n_0} \underbrace{N\left(y_{0j}\middle|\mu_0, \sigma_0^2\right) \prod_{i=1}^{n} N(y_i|\mu + d_i\lambda, \sigma^2)}_{\text{Likelihoods of the Active and Historical Trials}}. \quad (2.3.4)$$

Posterior

It should be noted that "The commensurate prior assumes that μ is a non-systematically biased representation of μ_0."[94] The y_s are data, τ is the commensurability parameter, and d_i is the

treatment indicator. In the above formulation, τ is given rather than incorporated as a parameter.

To estimate τ, the authors suggest using either empirical bayes (i.e. using the maximum marginal likelihood estimate for τ) or a fully Bayesian approach (treating τ as a parameter). For our purposes, we will be considering the fully Bayesian approach for which the authors suggest using a "spike and slab" distribution for the initial prior of τ. The "spike and slab" distribution was introduced by Mitchell and Beauchamp in 1988 for variable selection in Bayesian methods.[96] The "spike and slab" distribution is generally defined as follows:

$$\Pr(\tau < S_l) = 0, \tag{2.2.5}$$

$$\Pr(\tau < u \ \cap \{S_l \leq u \leq S_u\}) = p_0\{(u - S_l)/(S_u - S_l)\} \tag{2.2.6}$$

$$\text{and } \Pr(\tau > S_u) = \Pr(\tau = \text{K}) = 1 - p_0. \tag{2.2.7}$$

Where p_0 determines the amount of density in the spike versus the slab. The slab (2.2.6) is simply a uniform distribution between two values, S_l and S_u, scaled by p_0, and the spike (2.2.7) is a single value, K, (like the Dirac delta function) for the remaining density:

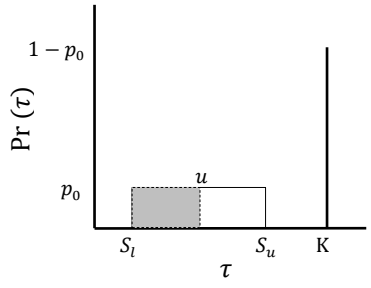


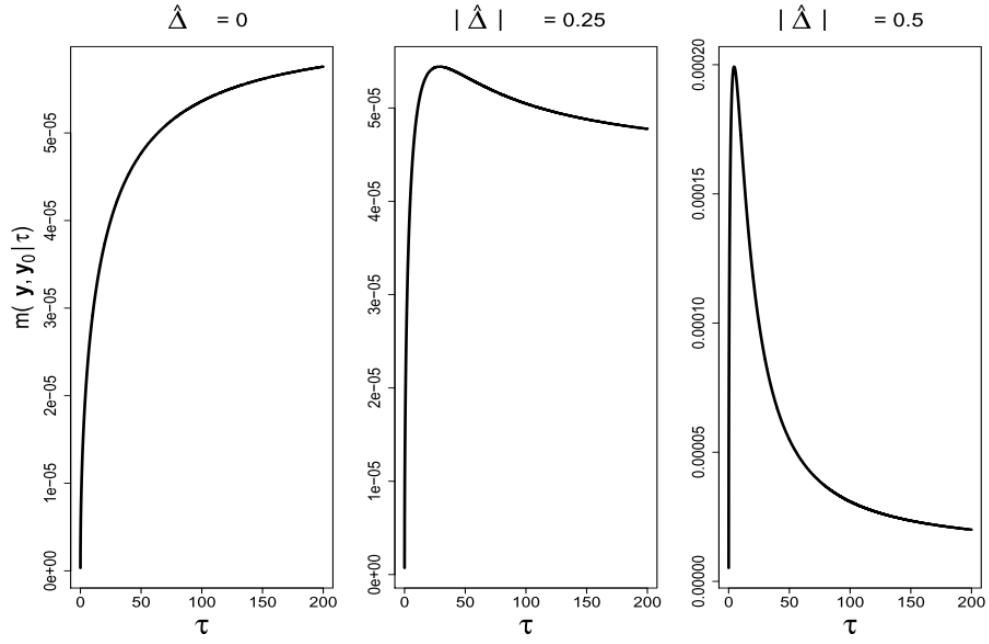
Figure 8. Spike and slab distribution. The y-axis is the probability density. The x-axis are values of τ. S_u and S_l are the upper and lower limits of the "slab." K is the τ value of the "spike." p_0 is the amount of density allocated to the "slab" with $1 - p_0$ allocated to the "spike." The shaded region is $Pr(\tau \leq u)$.

Why the "spike and slab" prior is useful for the initial prior for τ is explained most prominently by a figure in Hobbs's 2012 paper[94], reprinted in **Figure 9**. The $\widehat{\Delta}$ at the top of the three panels is

a measure of the difference between the historical and active trial mean outcomes. The x-axes are the values of the commensurability parameter, τ, which characterizes precision (lower values mean they are less similar, higher values mean they are more similar). When there is no difference between the trials (leftmost panel), you will see that the marginal density of the data likelihood plateaus at the higher values of τ. In fact, there is little information to help the model determine whether a value of 170 vs 200 for τ is better or worse because the likelihood values are almost the same. In contrast, as the difference between the historical and active trial increases, a clear peak in the marginal density of the data emerges (rightmost panel).

In the case of similarity, the Bayesian model needs assistance in picking a large value of τ when information is sparse. This is where the spike of the "slab and spike" prior distribution becomes useful. The spike is located at a large value of τ. Therefore, when there is little information, i.e. the likelihood of the data is relatively flat as in the leftmost panel in **Figure 9**, then the spike in the prior of τ will help the model converge to a reasonable, large value of τ. The slab portion of the prior is a simple, uniform vague initial prior that does not influence τ if there is a lot of information for a small value of τ as in the rightmost panel in **Figure 9**. This situation corresponds to when there is heterogeneity among the historical and active trials.

Figure 9. Marginal density of $y, y_0|\tau$ as a function of τ for three values of $|\widehat{\Delta}|$ from Hobbs 2012[94] where $\widehat{\Delta} = \hat{\mu} - \hat{\mu}_0$ and $\sigma^2 = 1, v_0 = 1/n_0, n_0 = 60, n = 180, n_d = 90$.



The 2012 paper finishes by expanding the commensurate prior framework to multivariable, generalized linear models and goes on to briefly discuss log-linear models for time to event data. The outcome can be written in terms of a normal distribution with the mean represented as a linear equation of variables and parameters and with a homoscedastic variance, both standard assumptions in statistical modeling. Note that in the following equations, $X_{0,g}$, $\beta_{0,g}$, X_g, and β_g are now *vectors* of parameters for $g \in (1, \dots, p)$ where p is the number of parameters in the multivariable linear model:

$$y_0 \sim N_{n_0}\left(X_{0,g}\beta_{0,g}, \sigma_0^2\right) \tag{2.2.8}$$

$$y \sim N_n\left(X_g\beta_g + d\lambda, \sigma^2\right) \tag{2.2.9}$$

Let λ (a constant) be the treatment effect and d the treatment indicator. The multiple covariates ($X_{0,g}$) in this model are baseline covariates *not* covariates through time. Note the change from a univariate model to a multivariable model while still having only one historical trial and one active trial:

$$X_{0,g}\beta_{0,g} = X_{0,1}\beta_{0,1} + X_{0,2}\beta_{0,2} + \cdots + X_{0,p}\beta_{0,p} \tag{2.2.10}$$

$$X_g\beta_g = X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p \tag{2.2.11}$$

In this case, the commensurate prior model is formulated by replacing a single commensurability parameter, τ, with the vector $\tau = (\tau_1, \dots, \tau_p)$ for all beta coefficients in the $\beta_{0,g}$ and β_g vectors. For each pair of parameters β_g and β_{0g} for $g \in (1, \dots, p)$ the commensurate priors, $p(\beta_g|\beta_{0g})$,

follow a normal distribution with variance τ_g^{-1} (note that τ_g is the precision) and are assumed to be independent:

$$p(\beta_g|\beta_{0g}) \propto N(\beta_g|\beta_{0g}, \tau_g^{-1}) \tag{2.2.12}$$

The active trial parameters, β_g, are modeled as if they are normally distributed around the historical trial's parameters, β_{0g}, with variance equal to the level of commensurability between the trials, τ_g^{-1}. The level of commensurability is allowed to vary for each parameter. This then allows us to define the joint posterior of $\theta|\tau, y, y_0$ as proportional to the following:

$$p(\lambda, \sigma, \sigma_0)\underbrace{N_{n_0}(y_0|X_0\beta_0, \sigma_0^2 I_{n_0})N_n(y|X\beta + d\lambda, \sigma^2 I_n)}\prod_{g=1}^{p}\underbrace{N(\beta_g|\beta_{0g}, \tau_g^{-1})p(\beta_{0g})} \tag{2.2.13}$$

Vague initial prior for the treatment effect and trial variances

Active and historical trial likelihoods

Commensurate priors with vague initial priors for β_{0g}

These models can be estimated using Markov Chains Monte Carlo (MCMC) methods[97], such as a Gibbs sampler. MCMC is a technique for estimating the posterior distribution using Bayes' theorem, an initial prior, and data likelihood. Its strength is that it can be used to estimate the posterior well no matter the complexity of the Bayesian model. In fact, before the development of MCMC, Bayesian statistics was limited to situations in which posteriors had known analytic forms after the Bayesian updating process, called conjugate priors.

In this work, we used Rjags, the R implementation of the C program JAGS (Just another Gibbs Sampler). All JAGS code is implemented by representing models using directed acyclic graphs (DAGS). For the simulation study, a representation of the commensurate prior hierarchical model is shown in **Figure 10** below.

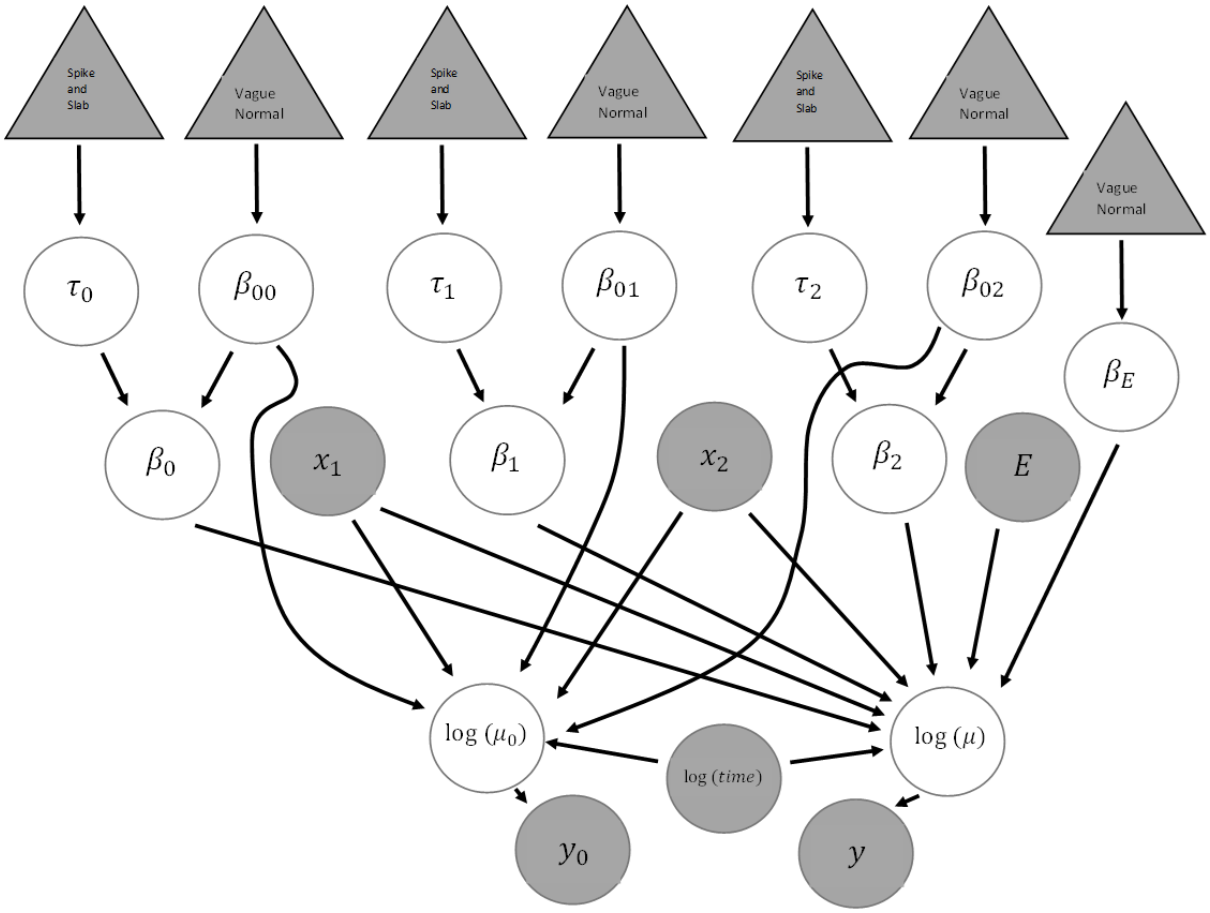


Figure 10. Directed Acyclic Graph (DAG) representation of the commensurate prior Bayesian hierarchical model used in the simulation. White circles represent model parameters. Grey circles represent observed variables and outcomes. Grey triangles represent prespecified priors on model parameters, including spike and slab and vague normal priors for the treatment effect.

The parameters τ_0, τ_1, τ_2,p_0, β_{00}, β_{01}, β_{02}, and β_E have the following prior specifications:

$$\tau_0 \sim Spike\ and\ Slab(S_l = 0.005, S_u = 30, \mathrm{K} = 30, p_{0,1}) \quad (2.2.14)$$

$$\tau_1 \sim Spike\ and\ Slab(S_l = 0.005, S_u = 100, \mathrm{K} = 300, p_{0,2}) \quad (2.2.15)$$

$$\tau_2 \sim Spike\ and\ Slab(S_l = 0.005, S_u = 100, \mathrm{K} = 300, p_{0,3}) \quad (2.2.16)$$

$$For\ i = 1,2,3;\ p_{0,i} \sim Bernouli(0.5) \quad (2.2.17)$$

$$For\ i = 1,2,3;\ \beta_{0i} \sim N(0, 10000) \quad (2.2.18)$$

$$\beta_E \sim N(0, 10000) \quad (2.2.19)$$

The values for β_{00}, β_{01}, β_{02}, and β_E all have non-informative normal priors (normal distributions with a huge variance of 10,000 do not affect the posterior estimation hence why they are called non-informative).

The values for the spike and slab priors were calibrated prior to the full simulation by considering a small simulation to determine their effect on the model. First, we considered whether $p_{0,i}$ should simply be set as a constant for all three values or allowed to be a parameter estimated in the model. We found that allowing $p_{0,i}$ to be a parameter led to most flexible model requiring fewer assumptions. Therefore, we modeled $p_{0,i}$ as a parameter. Next, we examined the scale of the β_g and $\beta_{0,g}$ parameters. The intercept terms, β_0 and $\beta_{0,0}$ are on a different scale than the two log hazard ratios β_1,β_2, $\beta_{0,1}$, and $\beta_{0,2}$ associated with X_1,X_2. Therefore, the value of K and S_u (as introduced in equations 2.2.6 and 2.2.7 and **Figure 8**) differ for τ_0 as seen in 2.2.14 because this precision (and corresponding variance) is on a different scale than τ_1 and τ_2. As mentioned above, calibration of the spike and slab priors are important for the fully Bayesian modeling approach.

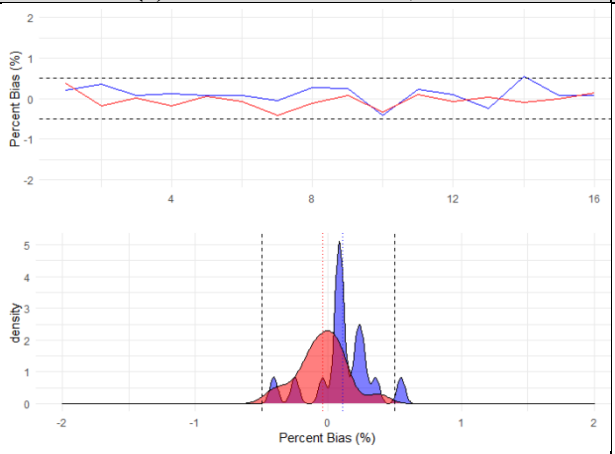
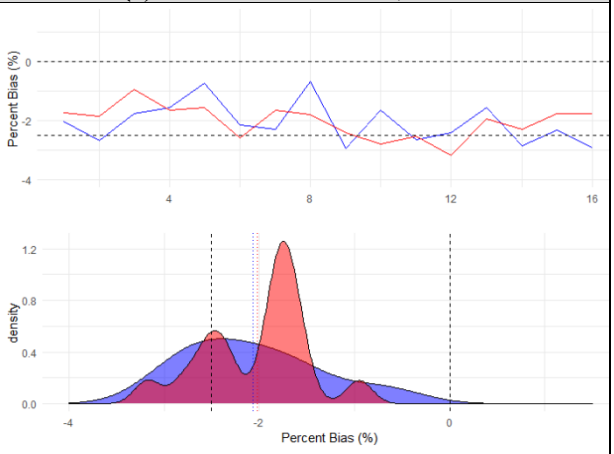
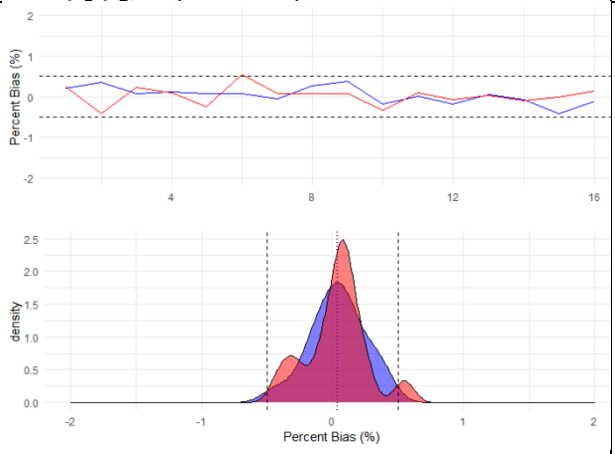
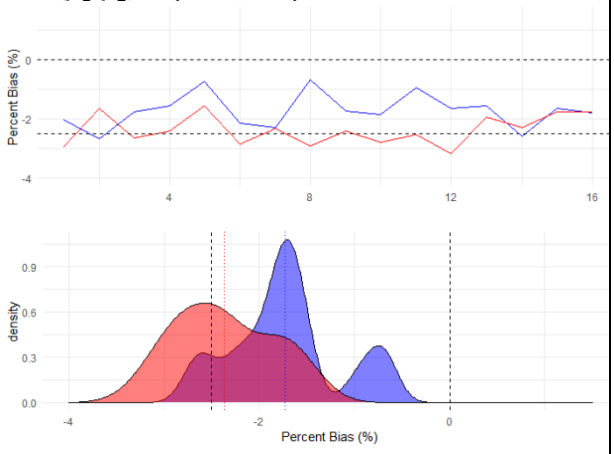
The Gibbs sampler utilizes the posterior distribution for each of the parameters (white circles in **Figure 10**) conditioned on the data and all other parameters. The result of the sampling process produces the joint posterior distribution and any desired marginal distributions.

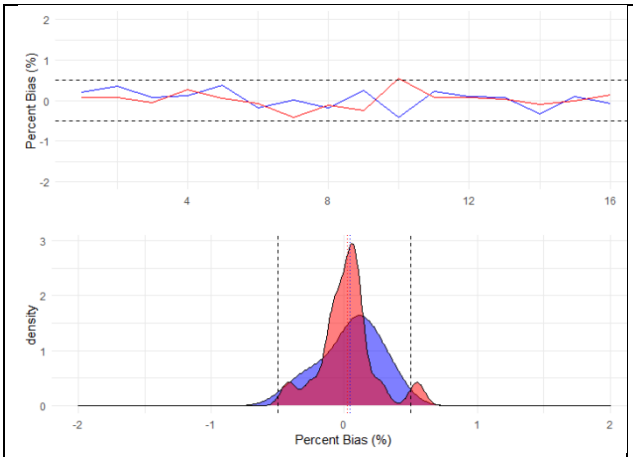
The parameter of interest is β_E, which is the treatment effect on the log-link scale because

exponentiated it is θ, the target estimand of the simulation. Therefore, the marginal posterior for β_E is extracted and 95% highest density credible intervals are derived. For brevity and because results were consistent across assumptions and parameters, a subset of figures and results will be presented in this chapter. However, the full simulation results on all performance measures are available on Github with details in appendix B.

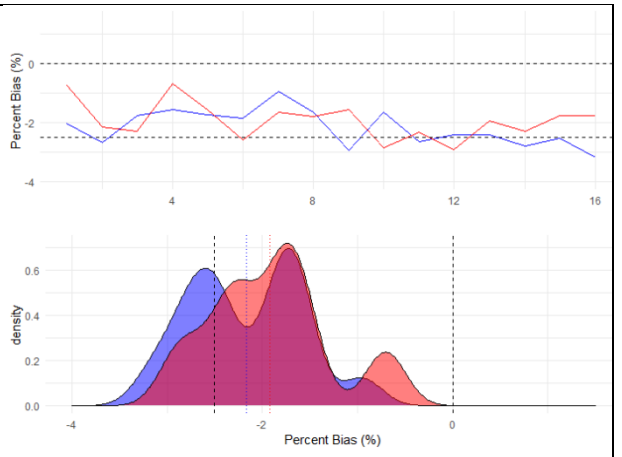
Figure 11. Commensurate priors, percent bias by simulation characteristic

In the figures below, the panel columns represent the two simulated (1) treatment effects (null versus alternative hypothesis). The panel rows represent the simulation characteristics (2)-(6) that were assessed, outlined in **Table 1.1**. The red and blue represent the two considered values of the selected characteristic. The upper part of the figure is a line graph, which compares each of the 16 subsets of scenarios where the (1) treatment effect and one of the characteristics (2)-(6) are held constant. Within each line, scenarios are arranged from 1 to 16 based on a specific ordering with details given in Appendix B. The lower part is a kernel estimated density[84] of the 16 scenarios to identify patterns. Dotted lines show percent bias within $\pm 0.5\%$ or 0% & -2.1% for their respective columns.

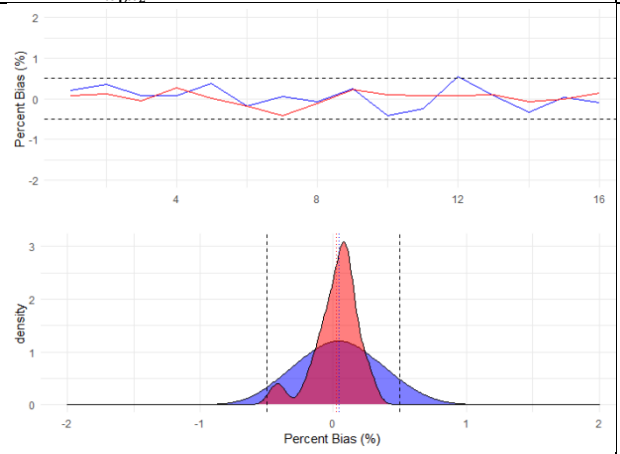
(1) Treatment effect HR, θ=1	**(1) Treatment effect HR, θ=0.5**
	
A) Varying (2) prevalence in the active trial. **Blue:** $(\mu_1, \mu_2) = (0.75, 0.20)$ **Red:** $(\mu_1, \mu_2) = (0.85, 0.10)$	**B) Varying (2) prevalence in the active trial.** **Blue:** $(\mu_1, \mu_2) = (0.75, 0.20)$ **Red:** $(\mu_1, \mu_2) = (0.85, 0.10)$
	
C) Varying (3) agreement of trials' covariate prevalence. **Blue:** $\boldsymbol{\delta} = (-0.4, +0.2)$ **Red:** $\boldsymbol{\delta} = (0, 0)$	**D) Varying (3) agreement of trials' covariate prevalence.** **Blue:** $\boldsymbol{\delta} = (-0.4, +0.2)$ **Red:** $\boldsymbol{\delta} = (0, 0)$



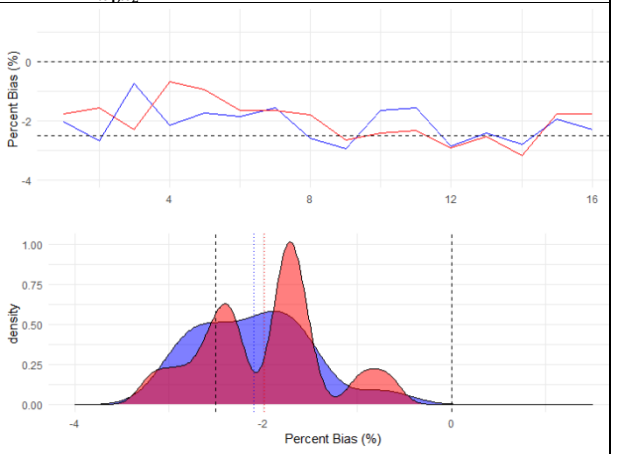
E) Varying (4) covariance/correlation between covariates
Blue: $Cov_{x_1,x_2} = 0.2$
Red: $Cov_{x_1,x_2} = 0.6$



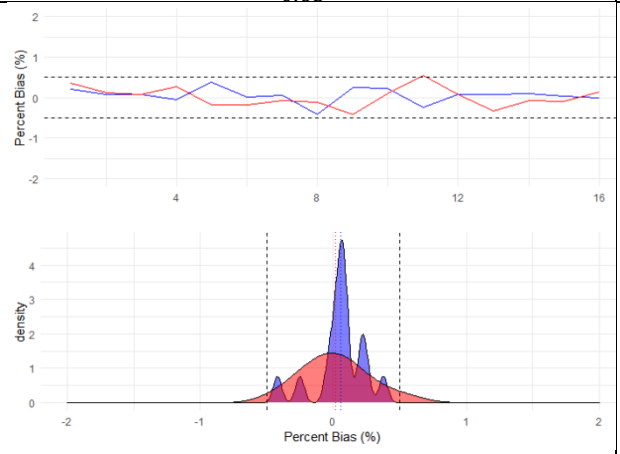
F) Varying (4) covariance/correlation between covariates
Blue: $Cov_{x_1,x_2} = 0.2$
Red: $Cov_{x_1,x_2} = 0.6$



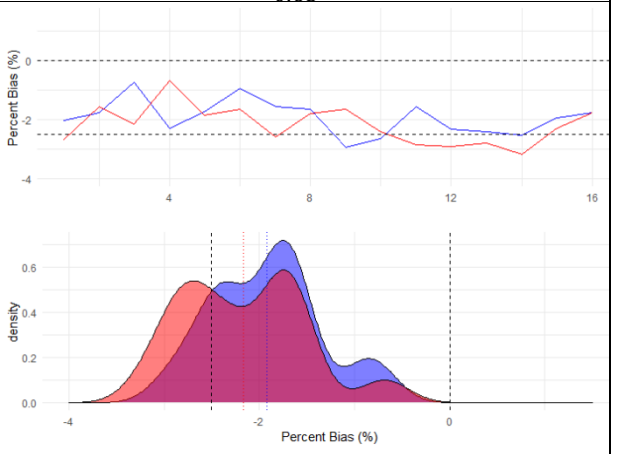
G) Varying (5) strength of association between covariates and outcome.
Blue: $\beta = \log(HR) = log\{1.10\}{1.30}$
Red: $\beta = \log(HR) = log\{1.20\}{0.65}$



H) Varying (5) strength of association between covariates and outcome.
Blue: $\beta = \log(HR) = log\{1.10\}{1.30}$
Red: $\beta = \log(HR) = log\{1.20\}{0.65}$



I) Varying (6) number of historical controls relative to the expected treatment arm size.
Blue: The same size
Red: Double the size



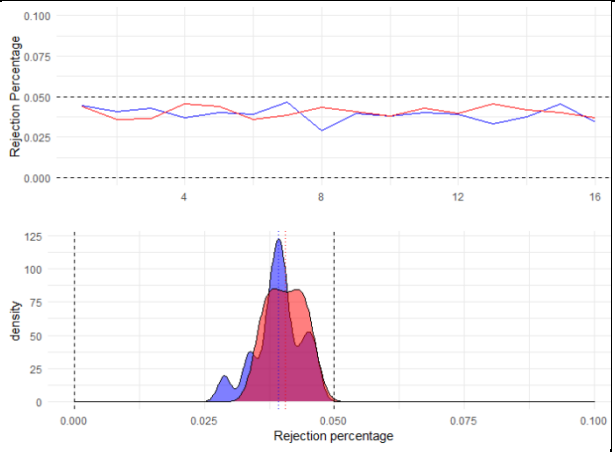
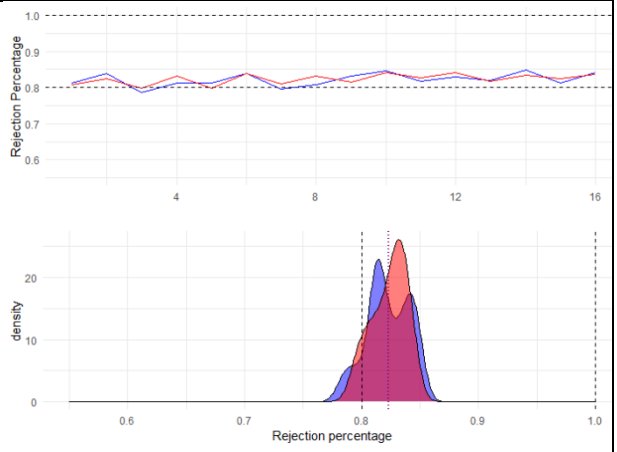
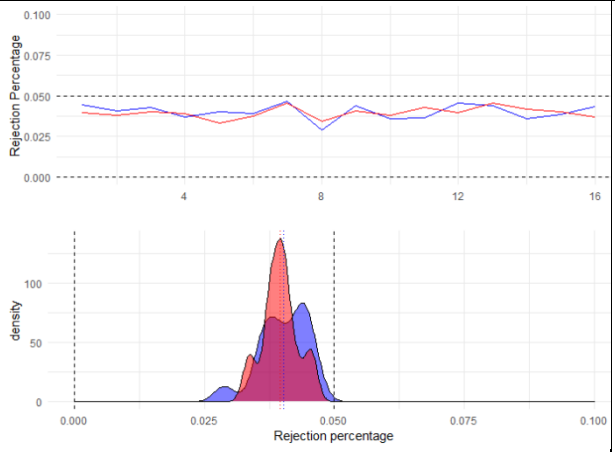
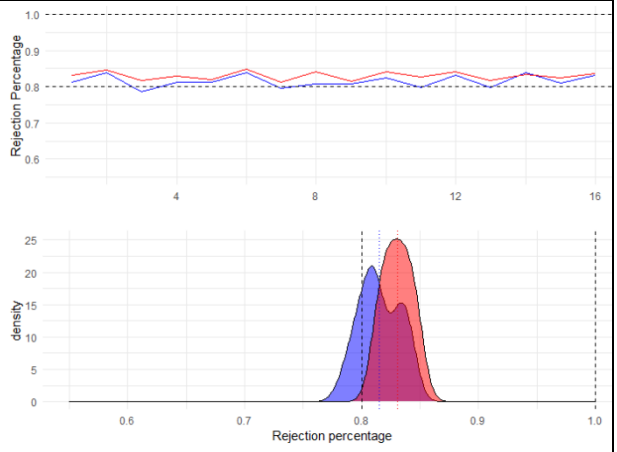
J) Varying (6) number of historical controls relative to the expected treatment arm size.
Blue: The same size
Red: Double the size

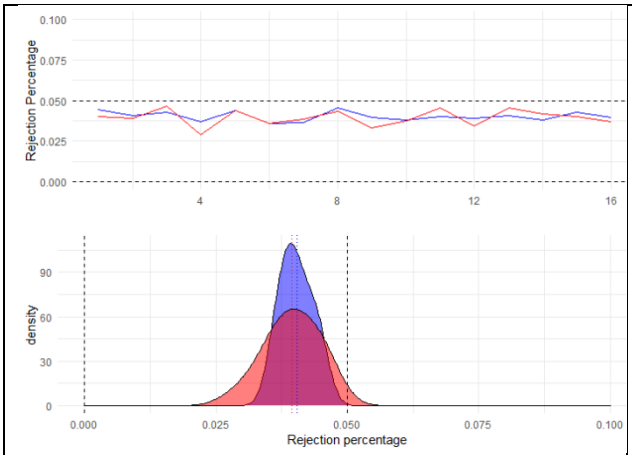
When the treatment had no effect, i.e. $\theta = 1$, all scenarios were unbiased with estimated within 0.5% of the truth, well within two Monte Carlo standard errors of 0.7%. When there was a treatment effect, i.e. $\theta = 0.5$, there appeared to be a slight systematic bias of -2%, however no bias exceeded 4%. When there was a treatment effect, the model seemed to perform slightly worse when the trials had different covariate prevalence (**Figure 11**, panel D). Similar to the propensity score-based power prior method, greater variation was observed when there was a treatment effect (**Figure 11**, panels B, D, F, H, and J).

Analysis of type I error and power showed that when there was no treatment effect, i.e. θ=1, all scenarios had type I error less than 5% hovering around 4%. When there was a treatment effect, i.e. $\theta = 0.5$, all scenarios achieved power slightly around 83%, with a target power of 80% and type I error of 5% under the model that includes no historical data. Two factors seemed to improve power further: When the relative size of the historical controls was double, power was improved by an additional ~3% (**Figure 12**, panel J). Also, when there was difference in covariate prevalence between the trials, there may be evidence of improvement in power (**Figure 12**, panel D). This slight improvement could be connected to the observed biased discussed in the paragraph above and seen in **Figure 11** panel D, possibly related to the bias-variance tradeoff of the Bayesian model.

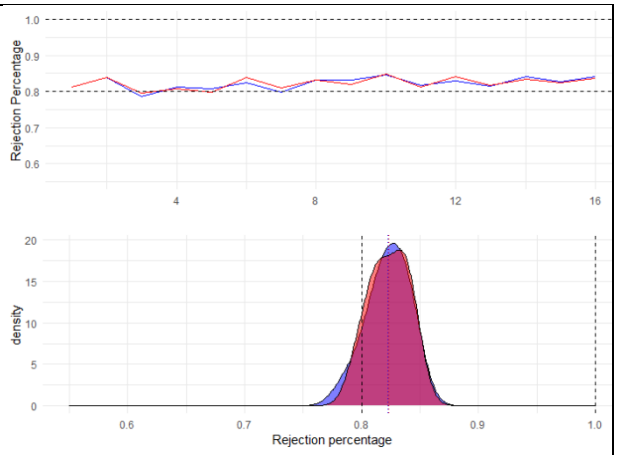
Figure 12. Commensurate priors, rejection percentage (type 1 error for HR=1 and power for HR=0.5) by simulation characteristic

In the figures below, the panel columns represent the two simulated (1) treatment effects (null versus alternative hypothesis). The panel rows represent the simulation characteristics (2)-(6) that were assessed, outlined in **Table 1.1**. The red and blue represent the two considered values of the selected characteristic. The upper part of the figure is a line graph, which compares each of the 16 subsets of scenarios where the (1) treatment effect and one of the characteristics (2)-(6) are held constant. Within each line, scenarios are arranged from 1 to 16 based on a specific ordering with details given in appendix B. The lower part is a kernel estimated density[84] of the 16 scenarios to identify patterns. Dotted lines are plotted at 0% and 5% for type I error and 80% and 100% for power.

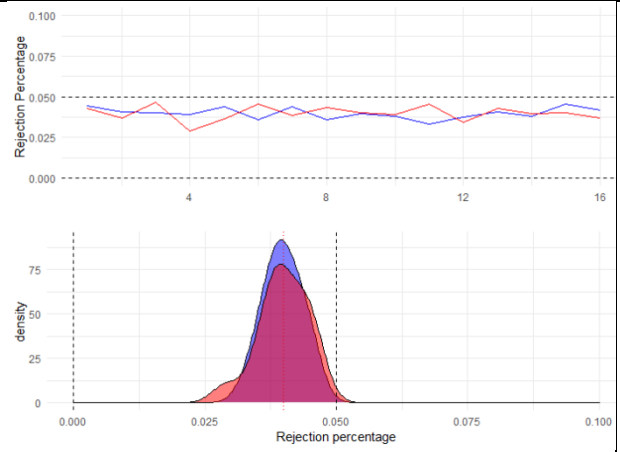
(1) Treatment effect HR, θ=1	(1) Treatment effect HR, θ=0.5
	
A) Varying (2) prevalence in the active trial. **Blue:** $(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) = (\mathbf{0.75}, \mathbf{0.20})$ **Red:** $(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) = (\mathbf{0.85}, \mathbf{0.10})$	**B) Varying (2) prevalence in the active trial.** **Blue:** $(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) = (\mathbf{0.75}, \mathbf{0.20})$ **Red:** $(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) = (\mathbf{0.85}, \mathbf{0.10})$
	
C) Varying (3) agreement of trials' covariate prevalence. **Blue:** $\boldsymbol{\delta} = (\mathbf{-0.4}, \mathbf{+0.2})$ **Red:** $\boldsymbol{\delta} = (\mathbf{0}, \mathbf{0})$	**D) Varying (3) agreement of trials' covariate prevalence.** **Blue:** $\boldsymbol{\delta} = (\mathbf{-0.4}, \mathbf{+0.2})$ **Red:** $\boldsymbol{\delta} = (\mathbf{0}, \mathbf{0})$



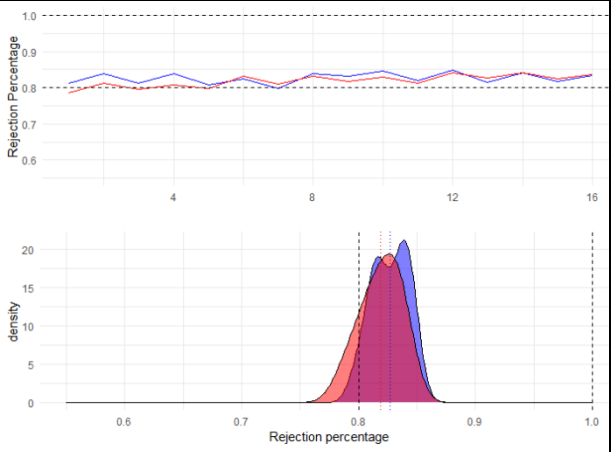
E) Varying (4) covariance/correlation between covariates
Blue: $Cov_{x_1,x_2} = 0.2$
Red: $Cov_{x_1,x_2} = 0.6$



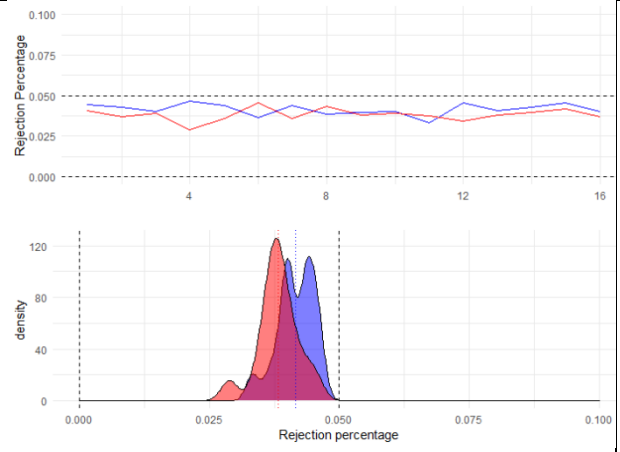
F) Varying (4) covariance/correlation between covariates
Blue: $Cov_{x_1,x_2} = 0.2$
Red: $Cov_{x_1,x_2} = 0.6$



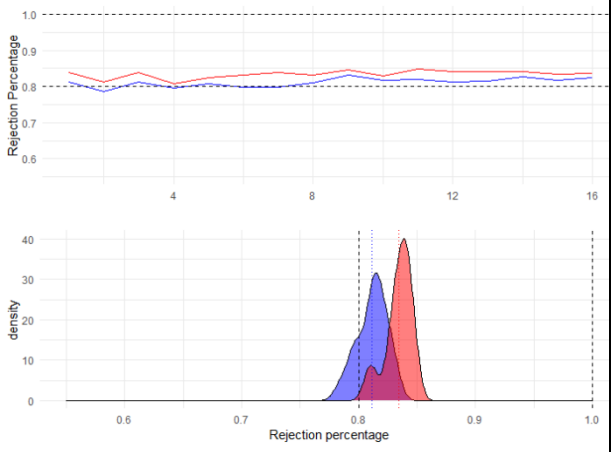
G) Varying (5) strength of association between covariates and outcome.
Blue: $\beta = \log(HR) = log\begin{Bmatrix}1.10\\1.30\end{Bmatrix}$
Red: $\beta = \log(HR) = log\begin{Bmatrix}1.20\\0.65\end{Bmatrix}$



H) Varying (5) strength of association between covariates and outcome.
Blue: $\beta = \log(HR) = log\begin{Bmatrix}1.10\\1.30\end{Bmatrix}$
Red: $\beta = \log(HR) = log\begin{Bmatrix}1.20\\0.65\end{Bmatrix}$



I) Varying (6) number of historical controls relative to the expected treatment arm size.
Blue: The same size
Red: Double the size



J) Varying (6) number of historical controls relative to the expected treatment arm size.
Blue: The same size
Red: Double the size

2.4 COMPARISON OF METHODS' SIMULATION PERFORMANCE MEASURES

We determined that the two simulation characteristics, taken from **Table 1.1**, that made the largest impact on performance measures were (3) agreement of trials' covariate prevalence and (6) number of historical controls relative to the expected treatment arm size (also seen in the respective panel rows when examining each analytic method individually). Therefore, **Table 2.1** below presents a subset scenarios that vary only these two characteristics and the value of θ while the remaining simulation characteristics take the following representative values: (2) prevalence in the active trial: $(\mu_1, \mu_2) = (0.75, 0.20)$, (4) covariance/correlation between covariates: $Cov_{x_1,x_2} = 0.6$, and (5) strength of association between covariates and outcome: $\beta = \log(HR) = log\begin{Bmatrix}1.20\\0.65\end{Bmatrix}$. Additionally, we examined all scenarios by (3) agreement of trials' covariate prevalence and the value of θ visually using three figures (shown below). The bimodality seen in the figures directly corresponds to (6) number of historical controls relative to the expected treatment arm size.

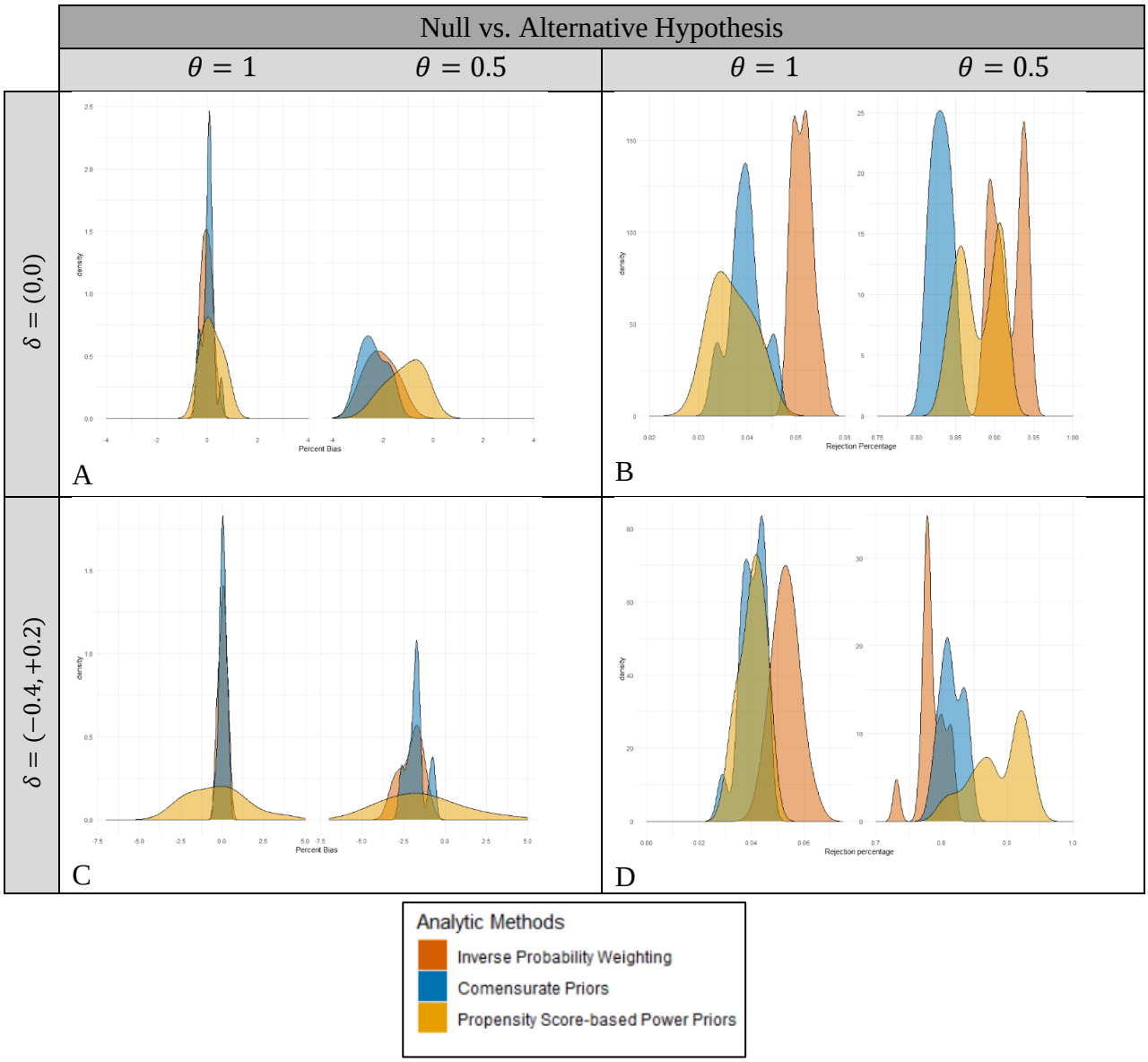
As seen in **Figure 13**, Inverse probability weighting (IPW) and commensurate priors may have a small empirical systematic bias around 2%, but was within Monte Carlo standard error of the simulation. When there is no treatment effect, $\theta = 1$, inverse probability weighting and commensurate priors did well in all scenarios and were within 0.5% of the true value. Propensity score-based power priors had more variability when the trial covariate prevalence were different (up to ~4%). As discussed in the earlier section, there may be slight evidence that a better fitting propensity score (caused by greater binary agreement and Cramer's ϕ between covariates, X_1,X_2) would improve both inverse probability weighting and propensity score-based power priors.

In terms of power and type I error, the two Bayesian methods had excellent type I error and had power around 80% or better. Propensity score-based power priors had the greater power of the two Bayesian methods and performed very well when the trials were different. IPW had the lowest power when the trial covariate prevalence were different and the highest power when the trial covariate prevalence was the same. More historical controls improved the type 1 error and power for all methods (**Figure 13**).

All methods achieved good coverage of the true treatment effect in all scenarios. The two, propensity score-based methods had smaller CI interval estimates than the commensurate prior method (**Figure 14**), although these intervals are interpreted differently and are applied to different quantities. It should be noted that the Bayesian credible intervals need not be symmetric (see section 2.2).

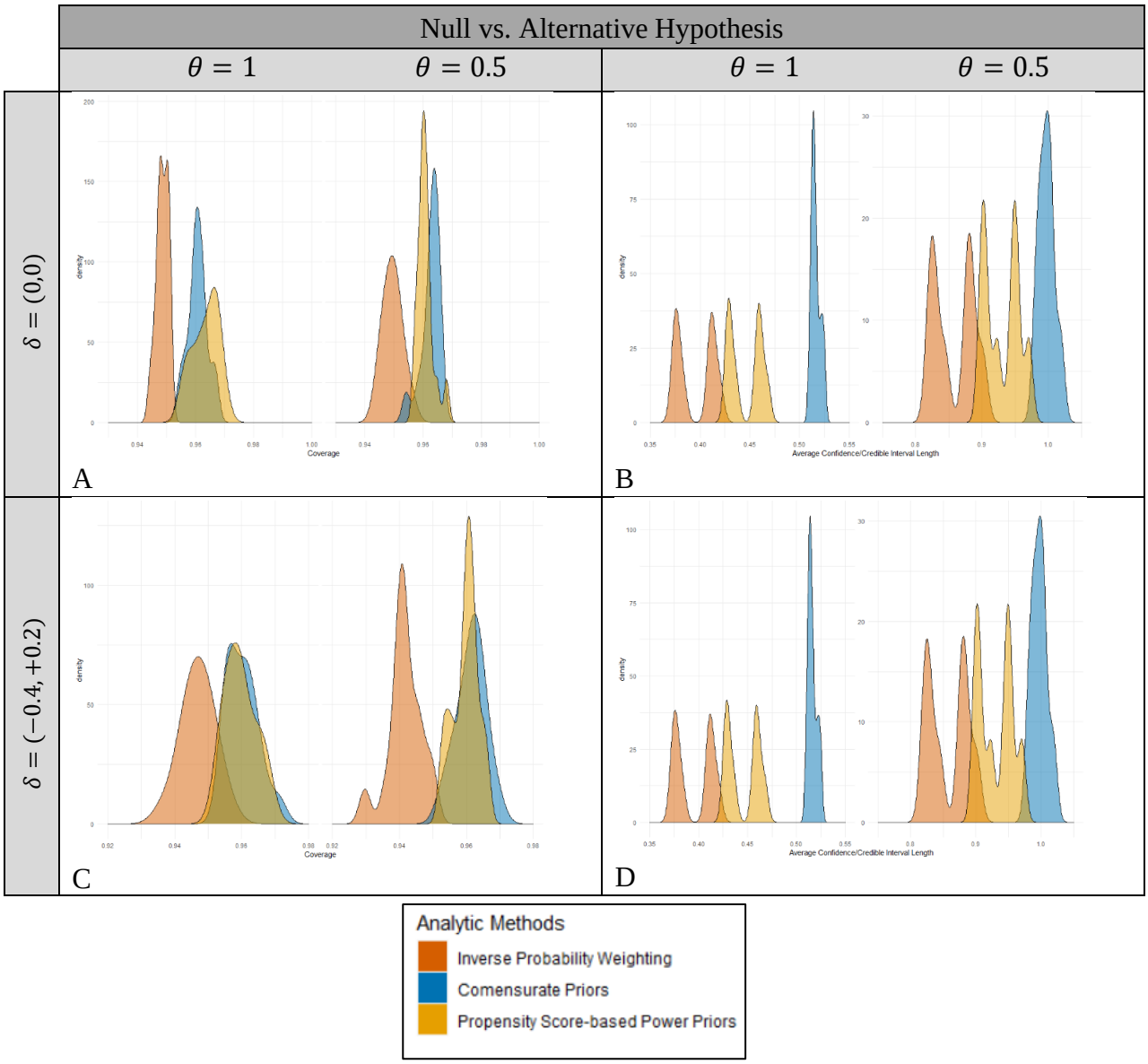
When examining empirical standard error and mean squared error, the Bayesian methods had similar levels of these performance measured regardless of whether the trial prevalence were similar or different. IPW had lower power when trials were different and higher power when trials were the same in terms of empirical standard error and mean squared error (like the other performance measures) (**Figure 15**).

Figure 13. Percent bias, type I error, and power by difference in covariate prevalence



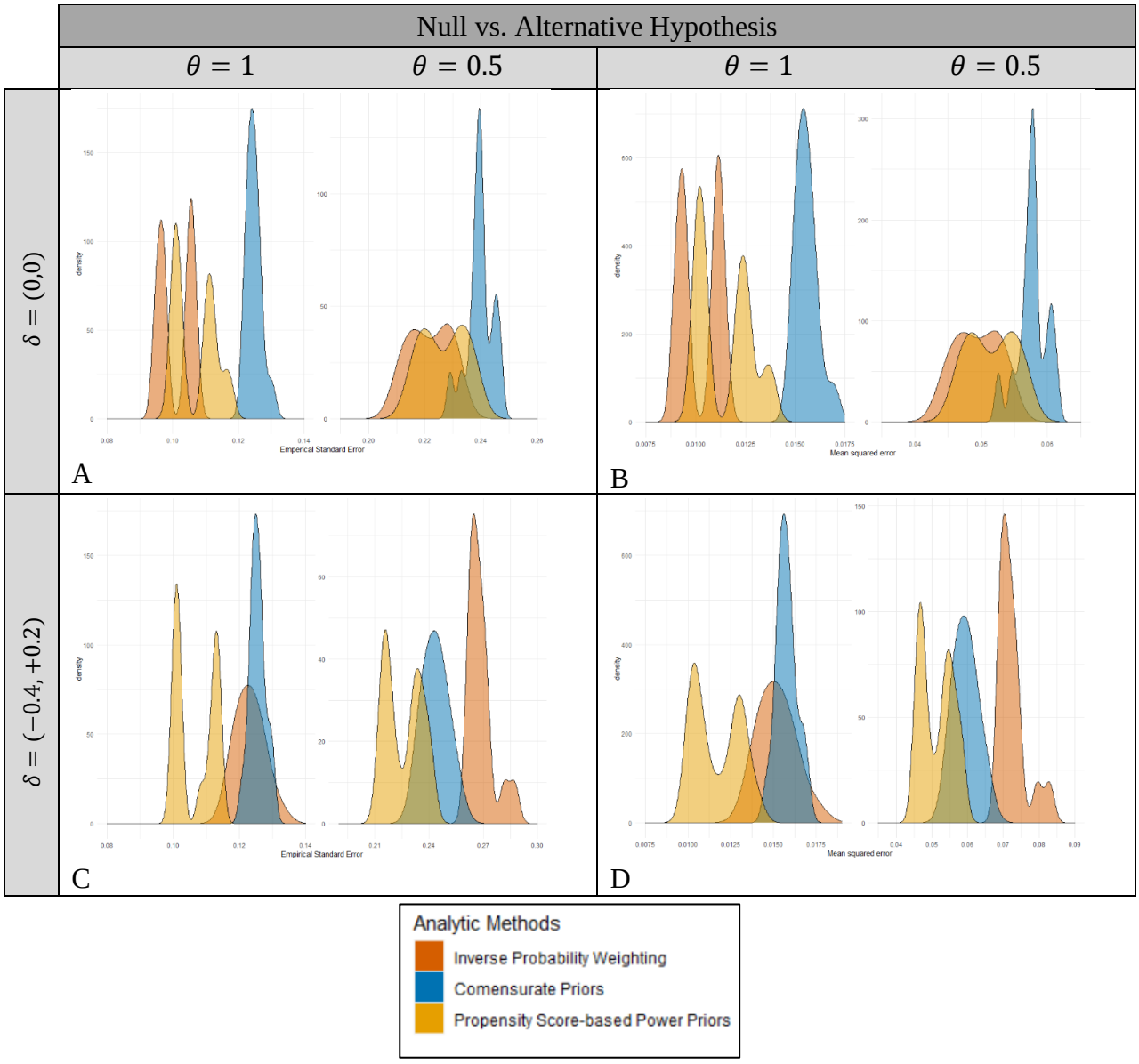
When the trial covariate prevalence were similar, $\delta = (0, 0)$, propensity score-based power priors may perform slightly better in terms of bias and type I error and inverse probability weighting has the greatest power. When the trial covariate prevalence were different, $\delta = (-0.4, +0.2)$, propensity score-based power priors had greater variation in terms of percent bias and the Bayesian methods performed in the best in terms of type I error and power with propensity-score based power priors being the most powerful. Bimodality observed in the figures directly correspond to the size of the historical controls relative to the expected size of the treatment arm.

Figure 14. Coverage and average confidence/credible interval length by difference in covariate prevalence



When the trial covariate prevalence were similar, $\boldsymbol{\delta = (0, 0)}$, or different, $\boldsymbol{\delta = (-0.4, +0.2)}$, the Bayesian methods may perform slightly better in terms of coverage. Commensurate priors had the largest average credible interval length, while IPW had the shortest. Bimodality observed in the figures directly correspond to the size of the historical controls relative to the expected size of the treatment arm.

Figure 15. Empirical standard error and mean squared error by difference in covariate prevalence



Commensurate priors had the largest empirical standard error and mean squared error when the trials were the same, $\delta = (0, 0)$. However, when trials were different and there was a treatment effect, the performance of commensurate priors improved and inverse probability weighting underperformed. Bimodality observed in the figures directly correspond to the size of the historical controls relative to the expected size of the treatment arm.

Table 2.1 Comparison of analytic methods for utilizing historical controls

H	Hx size	δ	Percent bias[a], % (MC SE)			Type I error or power, % (MC SE)		
			IPW[b]	PSPP[b]	CP[b]	IPW[b]	PSPP[b]	CP[b]
H_0	1	$(0,0)$	-0.1 (0.2)	0.6 (0.2)	0.1 (0.2)	5 (0.4)	4 (0.3)	5 (0.4)
H_0	1	$(-0.4,+0.2)$	-0.3 (0.2)	3.6 (0.2)	0.0 (0.2)	5 (0.4)	5 (0.4)	5 (0.4)
H_0	2	$(0,0)$	0.2 (0.2)	0.2 (0.2)	0.1 (0.2)	5 (0.4)	4 (0.3)	3 (0.3)
H_0	2	$(-0.4,+0.2)$	0.4 (0.2)	-1.2 (0.2)	0.3 (0.2)	5 (0.4)	3 (0.3)	3 (0.3)
H_A	1	$(0,0)$	-2.3 (1.3)	-0.4 (1.3)	-2.3 (1.4)	89 (0.5)	84 (0.6)	81 (0.7)
H_A	1	$(0.4,+0.2)$	-2.5 (1.6)	3.6 (1.3)	-2.3 (1.4)	73 (0.7)	80 (0.7)	80 (0.7)
H_A	2	$(0,0)$	-2.3 (1.2)	-1.2 (1.2)	-2.9 (1.3)	93 (0.4)	89 (0.5)	84 (0.6)
H_A	2	$(-0.4,+0.2)$	-1.0 (1.5)	-1.5 (1.2)	-0.7 (1.3)	78 (0.7)	91 (0.5)	81 (0.7)

H	Hx size	δ	Coverage, % (MC SE)			Average CI length[c]		
			IPW[b]	PSPP[b]	CP[b]	IPW[b]	PSPP[b]	CP[b]
H_0	1	$(0,0)$	95 (0.4)	96 (0.3)	96 (0.3)	0.42	0.47	0.52
H_0	1	$(-0.4,+0.2)$	95 (0.4)	95 (0.4)	96 (0.3)	0.52	0.47	0.53
H_0	2	$(0,0)$	95 (0.4)	96 (0.3)	97 (0.3)	0.38	0.44	0.52
H_0	2	$(-0.4,+0.2)$	95 (0.4)	97 (0.3)	97 (0.3)	0.47	0.43	0.52
H_A	1	$(0,0)$	95 (0.4)	96 (0.3)	96 (0.3)	0.90	0.97	1.02
H_A	1	$(0.4,+0.2)$	94 (0.4)	95 (0.4)	96 (0.3)	1.08	0.98	1.03
H_A	2	$(0,0)$	95 (0.4)	96 (0.3)	96 (0.3)	0.85	0.92	1.01
H_A	2	$(-0.4,+0.2)$	94 (0.4)	96 (0.3)	96 (0.3)	1.01	0.91	1.01

H	Hx size	δ	Empirical standard error[d]			Mean squared error[c]		
			IPW[b]	PSPP[b]	CP[b]	IPW[b]	PSPP[b]	CP[b]
H_0	1	$(0,0)$	0.11	0.12	0.13	0.012	0.014	0.017
H_0	1	$(-0.4,+0.2)$	0.13	0.11	0.13	0.009	0.014	0.017
H_0	2	$(0,0)$	0.10	0.10	0.12	0.017	0.010	0.015
H_0	2	$(-0.4,+0.2)$	0.12	0.10	0.12	0.014	0.010	0.015
H_A	1	$(0,0)$	0.23	0.24	0.25	0.054	0.057	0.060
H_A	1	$(0.4,+0.2)$	0.29	0.24	0.25	0.048	0.058	0.064
H_A	2	$(0,0)$	0.22	0.22	0.24	0.083	0.050	0.058
H_A	2	$(-0.4,+0.2)$	0.26	0.22	0.24	0.069	0.047	0.057

MC SE = Monte Carlo standard error, calculated with the formulas from Morris et al.[35], H=Hypothesis, Hx size = size of the historical control group relative to the expected size of the experimental arm, δ is a simulation characteristic (3) - the difference between the historical and active trial covariate prevalence ,

[a]Percent Bias is calculated on the HR scale for H_0 to avoid dividing by 0. Note that the unit is percentages

[b] IPW=inverse probability weighting, PSPP=Propensity score-based power priors, CP=commensurate priors,

[c]Monte Carlo standard error not reported because in all cases it was ≤ 0.001

[d]Monte Carlo standard error not reported because in all cases it was ≤ 0.01

Comparing three analytic methods on performance measures: percent bias, type I error or power, coverage, average confidence/credible interval length, empirical standard error, and mean squared error. Ordered in terms of null ($\boldsymbol{\theta = 1}$) or alternative ($\boldsymbol{\theta = 0.5}$) hypothesis, number of available historical controls, and difference in covariate prevalence between the historical and active trials. The remaining simulation characteristics were set to the following values: $(\boldsymbol{\mu_1, \mu_2}) = (\mathbf{0.75, 0.20})$, $\boldsymbol{Cov_{x_1,x_2} = 0.6}$, $\boldsymbol{\beta = \log(HR) = log\left\{\frac{1.20}{0.65}\right\}}$.

2.5 SUMMARY AND RECOMMENDATIONS

In this simulation, all methods performed well in terms of type I error and power when incorporating historical controls into their analyses and only using half the number of active controls as a traditional trial. However, when directly compared, inverse probability weighting had greater power when the trials were similar and less power when the historical and active trials differed in terms of covariate prevalence as compared to the Bayesian methods. This is also reflected in the coverage rates of IPW. In contrast, the Bayesian methods performed at or above target in terms and power while controlling type 1 error below 5%. Regardless of the scenario, propensity-score based power prior seemed to have a slight edge on the commensurate prior method in these performance measures.

In terms of bias, all methods were within a good range of the target. The propensity score-based power prior method exhibited greater variability in terms of a point estimate ($\pm 5\%$). However, this variability may have led to the higher observed rates of power and type I error as a sort of bias-variance trade off. This variability may have been caused by the matching step in Lin et al.'s method or it may be a consequence of modeling the percent of events in the study and converting them into a hazard ratio rather than modeling the hazard ratio directly.

The commensurate prior method had slightly larger average credible interval lengths but comparable coverage to propensity score-based power priors. IPW had smaller confidence intervals on average and worse coverage than the Bayesian methods. This may be a result of comparing credible intervals to confidence intervals because credible intervals directly model the treatment effect parameter's distribution and therefore may be more conservative. In contrast, confidence intervals are derived from asymptotic distribution results that may improve with greater samples than seen in this simulation. The Bayesian methods empirical standard error (the variance of the method estimates) were very consistent between simulation conditions, indicating robustness regardless of the parameter scenario. This also seemed to be true for mean squared error, a measure of accuracy that accounts for both the empirical standard error and bias. IPW had lower error when the trials were similar and higher errors when the trials were different.

In addition to performance measures of the simulation, it is useful to consider the implementations of each of the methods. The easiest method to implement was inverse probability weighting. Propensity score-based power priors required a larger pool of historical controls to make matched sets well-suited to the analysis. This may be a limitation because the size of the historical control pool is typically outside of an investigator's control. The commensurate prior method was the most complex, requiring a careful design of the hierarchical model as well as tuning for the spike and slab priors.

When recommending a methodology out of these three methods based on the simulation, I believe the Bayesian methods are slightly more robust to the presence of covariate differences between the trials than inverse probability weighting at the expense of being conservative when they are similar. In the case where the descriptive statistics of baseline characteristics of both trials show similarity, using inverse probability weighting could increase the power to detect a treatment effect. The Bayesian methods are more computationally intensive, especially for huge datasets. The ease of implementation of inverse probability weighting is a strength of the method. Between propensity score-based power priors and commensurate priors, the context of trial is very important. In the case where there is a large historical control pool and the primary analysis is not complicated, the propensity score-based power prior method may be preferable. For a simple primary analysis, the slightly higher variance in bias is worth gains in type I error and power. However, in the current state of the method, the inability to adjust for covariates in the Bayesian model may preclude certain types of primary analyses (for a hypothetical example, consider a desire for a sex-adjusted effect estimate). Also, the method requires a large historical control pool for good matches that may not be available.

In summary, commensurate priors may be the most broadly applicable of the three considered methods. The approach always had good type I error and power but exhibited larger, conservative credible intervals which led to slightly higher error as compared to the propensity score-based power prior method. The complexity of the modeling is challenging but also flexible in that it can handle many types of generalized linear models and covariates. Additionally, although not presented in this simulation, estimates of the commensurability parameters, τ_s, which describe the similarities and differences between the active and historical trial's statistical

models, may be of great interest to researchers. The specifics of how the trials differ in associations between variables and the outcome could be useful for better understanding similarities and differences between the trials and by extension help with justifying the use of an external comparator arm.

Chapter 3. A FIXED-BORROWING ADAPTIVE STUDY DESIGN FOR INCORPORATING HISTORICAL CONTROLS

3.1 STUDY DESIGN METHODOLOGY

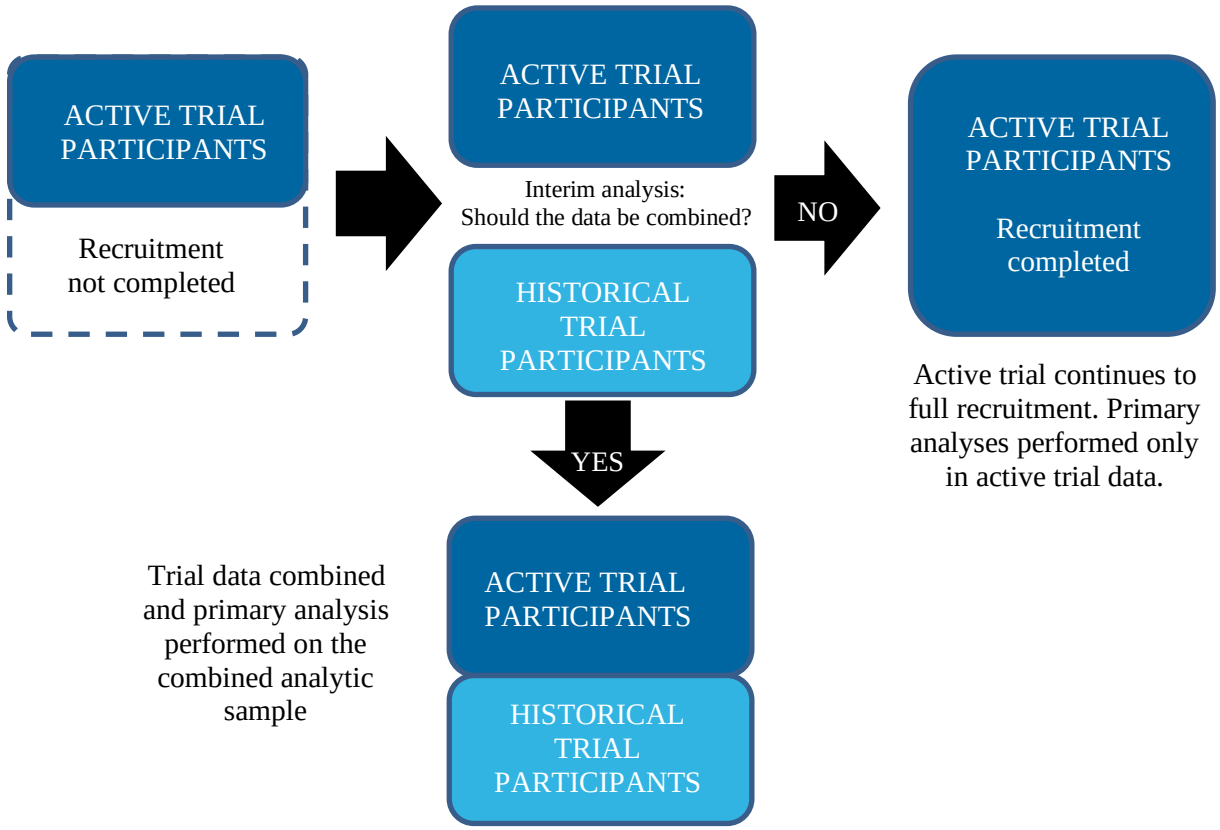
In the previous chapter, the idea was to use an analytic method that planned to incorporate the historical trial data and recruit fewer control arm participants in the active trial from trial onset. A natural extension would be to instead take an early look at the active trial (during participant accrual) and assess whether the historical data is appropriate to incorporate at all in the analysis. This approach focuses on the study design rather than the analytic approach and falls under the classification of an adaptive trial. Adaptive trial designs have planned changes to the study design in response to incoming information at prespecified timepoints. We selected an adaptive trial design proposed by Psioda et al.[98] that has a single preplanned interim analysis that assesses whether the historical control data is appropriate to use to improve the active trial. If combining the historical and active trial data is appropriate, then the active trial is stopped early, and the primary analysis is conducted with a combined analytic dataset. Otherwise, the active trial continues to full recruitment and the primary analysis is conducted only in the active trial data (**Figure 16**).

The adaptive design requires a few key pieces: First is the statistical model used for the primary analysis- whether it includes the historical trial data or not. Second, the exact timing of preplanned interim analysis to determine if the active trial can be stopped early. Third, the protocol for conducting the interim analyses and making the decision whether to stop and conduct the primary analyses with the historical data or continue the active trial to full recruitment. Each of these pieces will be explained in detail.

The statistical model:

The theoretical framework for the statistical model originated as an extension of the power prior[66] described in previous sections. The statistical model is a Poisson regression as also used in earlier chapters. The likelihood of the historical trial data can be discounted using the power parameter, α_0, as demonstrated in section 1.5 and 2.2. Then, the Bayesian posterior is derived using the active trial data's likelihood and the power prior. However, the authors take advantage

Figure 16. Adaptive trial design illustrated



An early look at the data is conducted at the interim analysis. If a stopping rule is reached, then the trial stops early, combines the data, and conducts the primary analysis. Otherwise, the active trial continues to completion.

Of the Bayesian central limit theorem[99] to avoid the complex calculations needed to estimate the posterior, $\pi(\theta|D, D_0, \alpha_0)$, directly:

$$\pi(\theta|D, D_0, \alpha_0) \propto Normal\left(\theta\middle|\hat{\theta}, \sigma_{\hat{\theta}}^2\right) \tag{3.1.1}$$

The expression above states that at reasonably large sample sizes the posterior of a power prior analysis is proportional to a normal distribution centered at the weighted maximum likelihood of an analysis of both the active and historical trials with variance associated with the respective inverse of the observed information matrix[98]. Psioda et al. further define this key connection by

explicitly writing out the equation defining the log of the posterior in terms of log likelihoods:

$$\log\big(\pi(\theta|D, D_0, \alpha_0)\big) \approx \sum_{i=1}^{n} 1 * \ell(\gamma, \psi|D_i) + \sum_{j=1}^{n_0}\big[\alpha_0 * \ \ell\big(\psi\big|D_{0,j}\big)\big] + \log\big(\pi_0(\theta)\big), \qquad (3.1.2)$$

where the model parameter, θ, is composed of the treatment effect, γ, and the remaining nuisance parameters generated by the covariate effects, ψ. The sample sizes of the active and historical trials are written as n and n_0, respectively. It should be noted that "≈" means asymptotically equal, $\ell()$ is the log likelihood function given either the active trial data, D, or the historical trial data, D_0, and that $\pi_0(\theta)$ is the non-informative component of the power prior (see equation 1.3.5 in section 1.5). The right-hand side of 3.1.2 is a standard weighted log-likelihood with the active trial participant data receiving a weight of 1 and the historical trial data receiving a weight of α_0, the power parameter that controls discounting. Also, the historical trial data likelihood is only a function of the nuisance parameters, ψ, because that data cannot provide information about the experimental arm and by extension the treatment effect, γ. This weighted log-likelihood leads to the maximum likelihood estimate and variance seen in expression 3.1.1 that allows for approximation of the Bayesian posterior distribution. A researcher can derive probabilities from a normal distribution that are asymptotically analogous probabilities from the posterior. Posterior probabilities of the treatment effect taking a value of 0 or greater on the log scale can be thought of as similar in usefulness to a one-sided frequentist p-value. Doubling the tail with lesser density (in our case density for values greater than 0) is similar to a two-sided frequentist p-value. Shi et al. discuss and justify this comparison between Bayesian posterior probabilities and frequentist p-values.[100]

The theoretical framework requires a value for α_0, the parameter that controls the discounting in the case where the historical data is used. By the author's recommendation, the value of α_0 can be explicitly defined in terms of the timing of the interim analysis and this target sample size, ensuring that the *pooled* active and historical trial data has the same effective number of outcomes as if the trial had fully recruited to the target sample size.

An examination of the weights of the weighted maximum likelihood expression in 3.1.2 lead to the following insight: if the active trial's likelihood is weighted by a value of 1 and the historical data is weighted by a value of α_0 (recall that $0 \leq \alpha_0 \leq 1$), then the historical data can be viewed as having informational value equal to a proportion (α_0) of the active trial's data. For example, if $\alpha_0 = 0.75$ then data from a historical trial with 100 participants with about ~50 events would be discounted to have the informational value of 75 active trial participants with about ~37 events. The exact calculation of α_0 depends on the number of observed events which corresponds with the sample sizes. Consider the target number of events of the active trial, N, number of events in the active trial participants at interim analysis, $n_{interim}$, and the number events in the historical trial, n_{hist}. Then α_0 can be calculated by solving the equation 3.1.3 in terms of α_0:

$$N = n_{interim} + \alpha_0 * n_{hist} \tag{3.1.3}$$

$$\alpha_0 = \frac{(N - n_{interim})}{n_{hist}} \tag{3.1.4}$$

This formulation allows researchers to reach the informational equivalent of the target sample size for any selected timing of the interim analysis and explicitly defines α_0. An earlier interim analysis timing will correspond to larger values of α_0 and by extension more dependence on the historical information to reach an amount of information equal to the target sample size. Later timing will correspond to smaller values of α_0, meaning less influence and reliance on historical trial data. If the interim analysis determines that the active trial will not stop early and continue to recruit to a full trial size, then 3.1.2 simplifies to only consider the active trial data:

$$\log\big(\pi(\theta|D)\big) \approx \sum_{i=1}^{n} 1 * \ell(\gamma, \psi|D_i) \tag{3.1.5}$$

The timing of the preplanned interim analysis:

This timing of the interim analysis is left up to the expert knowledge of the investigators, potential power and sample size reduction determined in the trials planning simulation (discussed in the next subsection), and the unique demands of the active trial. In the original paper, the authors define it in terms of what percentage of the active trial's target sample size has been

recruited. The minimum percentage the authors considered was 50%, corresponding to an interim analysis halfway through participant recruitment. Although an early interim analysis could be conducted, having 50% or more of the analysis data coming from the active trial is scientifically conservative because there must be enough active trial data to determine whether the historical data is similar enough to justify stopping early. For the current study, 87.5%, 75%, 62.5%, and 50% of the recruited participants were assessed as potential interim analysis timings. An additional consideration with regards to the interim analysis timing is the active trial's randomization scheme. For example, if the trial stops early, it could be beneficial to use a 2:1 randomization scheme so that the pooled historical controls make the treatment group and control groups roughly equal. However, if the trial continues to completion a 1:1 randomization scheme maximizes power for the given sample size. Fortunately, Psioda et al. determined that the active trial's randomization scheme had almost no effect on the design's performance[98]. Therefore, to make the approach more comparable to what was done in section 1.6, we used a 2:1 randomization scheme for the active trial, intervention versus comparator arm.

The protocol for the interim analysis and the stopping decision:

When the active trial has recruited enough participants and obtained enough events to conduct the interim analysis, what should constitute the stopping rule of the adaptive trial? The goal of the authors was to borrow information from "a historical trial using subject-level control data while assuring a reasonable upper bound on the maximum type I error and lower bound on the minimum power."[98] Therefore, the stopping rule was formulated as statistical test, but this test is not the primary analysis. The stopping rule test consists of a likelihood ratio statistic and a constant, critical value determined *a priori* to guarantee prespecified type I error and power for the entire design. If the likelihood ratio statistic is less than or equal to the critical value, then the study is stopped early (participant accrual halts), and the historical trial data is immediately incorporated to augment the active trial to conduct the primary outcome analysis. If it is greater than the critical value, then the historical trial data is discarded, and the active trial continues to full enrollment (**Figure 16**). The calculation of the critical value, labeled as w_0, for the likelihood ratio used at the interim analysis to determine if the study should be stopped early is done using simulation.

Before describing this calculation, let us first define the special log-likelihood ratio statistic, W, used at the interim analysis, exactly as is done in Psioda[98]:

$$W = \log\left(\mathcal{L}\left(\hat{\theta}_1 \middle| D_1\right)\mathcal{L}\left(\hat{\psi}_0 \middle| D_0\right)^{\alpha_0}\right) - \log\left(\mathcal{L}\left(\hat{\theta} \middle| D_1\right)\mathcal{L}\left(\hat{\psi} \middle| D_0\right)^{\alpha_0}\right) \quad (3.1.6)$$

where $\hat{\theta} = \left(\hat{\gamma}, \hat{\psi}\right) = argmax(\mathcal{L}(\theta|D_1)\mathcal{L}(\psi|D_0)^{\alpha_0})$, $\hat{\theta}_1 = \left(\hat{\gamma}_1, \hat{\psi}_1\right) = argmax(\mathcal{L}(\theta|D_1))$, and $\hat{\psi}_0 = argmax(\mathcal{L}(\psi|D_0)^{\alpha_0})$. This likelihood ratio, expressed as a difference in log likelihoods, is a statistic that compares the model fit of a model where the power prior analysis is carried out exactly as described above, estimating one, common value for the parameters among both trials (the log-likelihood corresponding to $\log\left(\mathcal{L}\left(\hat{\theta} \middle| D_1\right)\mathcal{L}\left(\hat{\psi} \middle| D_0\right)^{\alpha_0}\right)$), versus a model where the active trial is presumed to have different nuisance parameters in comparison with the historical trial (corresponding to $\log\left(\mathcal{L}\left(\hat{\theta}_1 \middle| D_1\right)\mathcal{L}\left(\hat{\psi}_0 \middle| D_0\right)^{\alpha_0}\right)$).

This likelihood ratio has a straightforward interpretation. When $W > 0$ (and the likelihood ratio > 1), the data are more consistent with the first model in the subtraction operation (when the active trial and historical trials are modeled with different parameter values). So, when W is large, this indicates that the baseline event rates and covariate parameters between the two studies are likely too different and therefore the trial should continue to full patient recruitment and not incorporate the historical data. In contrast, when $W < 0$ (and the likelihood ratio < 1), the data is more consistent with the second model (the trials are modeled with the same, common parameter values). When W is negative, the baseline event rate and covariate parameters are similar enough to justify stopping the trial early and immediately incorporating the historical data into the analytic dataset for the primary analysis.

A threshold for W, w_0, is computed to meet prespecified type I error and power requirements set by the investigators. When $W < w_0$, the active trial is stopped early, and the historical data is incorporated for the primary analysis. Sometimes w_0 is larger than 0, meaning that the procedure will still incorporate the historical data even if there is evidence of differences between the trials. The fact that w_0 can be a small positive number shows that under certain type I error and power requirements, the procedure tolerates differences between the active and historical trial to

combine them for the primary analysis. This flexibility is the strength of the adaptive trial design but depends on proper calculation of this critical value, w_0.

The critical value is calculated in a way to ensure that the entire procedure meets predetermined levels of type I error and power. The procedure consists of two primary analysis opportunities, an interim analysis that may stop the trial early, and a single primary analysis. The authors propose that this critical value be calibrated through an appropriate simulation that considers an array of possible nuisance parameter values for the new trial (note that the nuisance parameters for the historical data can be estimated before the first active trial participant is recruited). Then, across this full array, data is simulated, the interim analysis is conducted with its respective likelihood ratio statistic and the primary analysis is conducted twice at both of the two opportunities. Then, indicator values for whether or not the null hypothesis should be rejected based on the data is calculated for both primary analysis per parameter set in the array.[98] These indicator values then can be run through an algorithm that tests all possible critical values, w_0, until the desired type I error rate and power are achieved.[98] The authors suggest using slightly more flexible minimum levels of type I error and power than the traditional 0.05 and 0.80 because these are the most extreme seen across the parameter array. Therefore, for this dissertation, we set the maximum type I error to be 0.075 and minimum power to be 0.75.

In the original paper, Psioda et al. examine their study design using a large simulation study as well as real trial data. We will also assess the method with the simulation study described in section 1.6, as well as in real trial data. However, there are important differences between the current work and what was done in the original paper. In the current work, we consider all combinations of the six key characteristics we wished to examine. This contrasts with the Psioda et al.'s paper that contained a simulation that only considered differences in the baseline hazard rate and not the covariate distributions. That paper only perturbed the baseline hazard rate for the sake of clarity and simplicity and argued in an appendix that perturbing the baseline hazard alone is sufficient for the procedure rather than considering the high dimensional space of perturbing all nuisance covariates. This is because an extreme enough perturbation (they used up to 45%) of the baseline hazard rate will dwarf any effect caused by likely differences in the nuisance parameters. In the current work, we consider the full combination of the six characteristics,

which changes how we present the results and answers a slightly different question more comparable to the answers in 1.6. Additionally, in Chapter 4, we look at the study's performance in real trial data.

3.2 SIMULATION PERFORMANCE MEASURES AND SUMMARY

The fixed-adaptive borrowing design was assessed with the same study design and parameter inputs used in section 1.6 with one exception. The methods in chapter 2 used analytic datasets of size determined by traditional formulas in which half of the concurrent controls were discarded, simulating a smaller trial. In this chapter, the same simulation framework is used but the full active trial control arm is kept. This is because this adaptive borrowing approach utilizes an interim analysis and thus reduces the needed sample size in a different way. We simulated datasets from all 64 scenarios. Therefore, we could analyze the results in a similar fashion as was done earlier, but since findings were similar over much of the parameter space, we decided to restrict the results to the circumstances as presented in **Table 2.1**. Specifically, we present the percent bias and type I error or power for influential parameter inputs only: different treatment effect sizes ($H_0\ vs. H_a$), different sizes of the available historical controls (the same size as the active trials treatment arm, indicated with a 1, vs. double the size of the treatment arm, indicated with a 2), and the difference in prevalence between the active and historical trials ($\delta = (0,0)\ vs. \delta = (-0.4, +0.2)$). The remaining simulation characteristics were held constant and set to the following representative values: $(\mu_1, \mu_2) = (0.75, 0.20)$, $Cov_{x_1,x_2} = 0.6$, $\beta = \log(HR) = log\begin{Bmatrix}1.20\\0.65\end{Bmatrix}$. These performance measures are presented for three cases: 1) if the trial always stopped at the interim analysis, 2) if the trial always continued to the full, final analysis, and 3) a hybrid estimate of these performance measures, representative of the whole procedure's performance if conducted many times (i.e. a weighted average of 1 and 2). Additionally, we present the (rounded) stopping probability for the interim analysis as well as the average reduction in trial size, which incorporates the proportion of times when the study stopped at the interim analysis. The full simulation results can be downloaded from GitHub. See appendix B for more details.

Table 3.1 Performance of the fixed, adaptive borrowing study design

A) Interim analysis opportunity

H	Hx size	δ		Percent bias, %			Type I error or power, %		
				50%[a]	75% [a]	87.5% [a]	50%[a]	75% [a]	87.5% [a]
H_0	1	$(0,0)$		0.4	0.3	0.3	0.05	0.03	0.03
H_0	1	$(-0.4,+0.2)$		0.6	0.6	0.5	0.03	0.03	0.04
H_0	2	$(0,0)$		0.2	0.2	0.3	0.09	0.09	0.07
H_0	2	$(-0.4,+0.2)$		0.1	0.2	0.2	0.05	0.04	0.04
H_A	1	$(0,0)$		-1.4	-1.4	-0.8	0.71	0.81	0.82
H_A	1	$(0.4,+0.2)$		-0.7	-0.3	0	0.66	0.80	0.81
H_A	2	$(0,0)$		-1.6	-1.9	-2	0.77	0.90	0.94
H_A	2	$(-0.4,+0.2)$		-1.1	-0.8	-0.7	0.75	0.88	0.93

B) Final analysis opportunity

H	Hx size	δ	Percent bias, %			Type I error or power, %		
			50%[a]	75% [a]	87.5% [a]	50%[a]	75% [a]	87.5% [a]
H_0	1	$(0,0)$	0.1	0.1	0.1	0.05	0.05	0.05
H_0	1	$(-0.4,+0.2)$	0.4	0.4	0.4	0.05	0.05	0.05
H_0	2	$(0,0)$	0.1	0.1	0.1	0.05	0.05	0.05
H_0	2	$(-0.4,+0.2)$	-0.2	-0.2	-0.2	0.04	0.04	0.04
H_A	1	$(0,0)$	0.3	0.3	0.3	0.79	0.79	0.79
H_A	1	$(0.4,+0.2)$	0.6	0.6	0.6	0.80	0.80	0.80
H_A	2	$(0,0)$	0.1	0.1	0.1	0.79	0.79	0.79
H_A	2	$(-0.4,+0.2)$	1.1	1.1	1.1	0.81	0.81	0.81

C) Performance of the whole procedure (Hybrid)

H	Hx size	δ	Percent bias, %			Type I error or power, %		
			50%[a]	75% [a]	87.5% [a]	50%[a]	75% [a]	87.5% [a]
H_0	1	$(0,0)$	0.4	0.3	0.3	0.05	0.03	0.04
H_0	1	$(-0.4,+0.2)$	0.6	0.6	0.5	0.03	0.03	0.04
H_0	2	$(0,0)$	0.0	0	0.1	0.07	0.07	0.07
H_0	2	$(-0.4,+0.2)$	0.1	0.2	0.2	0.05	0.04	0.04
H_A	1	$(0,0)$	0.1	-1.4	-0.8	0.75	0.81	0.82
H_A	1	$(0.4,+0.2)$	1	-0.3	0	0.75	0.80	0.81
H_A	2	$(0,0)$	-1.6	-1.9	-2	0.77	0.90	0.94
H_A	2	$(-0.4,+0.2)$	-0.4	-0.8	-0.7	0.75	0.88	0.93

D) Stopping percentage and average trial size reduction of the whole procedure (Hybrid)

H	Hx size	δ	Stopping probability, %			Average trial size reduction, %		
			50%[a]	75% [a]	87.5% [a]	50%[a]	75% [a]	87.5% [a]
H_0	1	$(0,0)$	1	1	1	50	25	12
H_0	1	$(-0.4,+0.2)$	1	1	1	50	25	12
H_0	2	$(0,0)$	0.52	0.38	0.45	26	9	6
H_0	2	$(-0.4,+0.2)$	1	1	1	50	25	12
H_A	1	$(0,0)$	0.46	1	1	23	24	12
H_A	1	$(0.4,+0.2)$	0.42	1	1	21	24	12
H_A	2	$(0,0)$	1	1	1	50	24	12
H_A	2	$(-0.4,+0.2)$	0.89	1	1	44	24	12

H= Presence of a treatment effect (H_A) or no treatment effect (H_0);

Hx size = number of historical controls used relative to the size of the treatment arm of the active trial. 1 indicates that the same size was used and 2 means double the number were used;

δ = the difference in covariate prevalence between the active and historical trials covariates, X_1,X_2.

[a]Represents the timing of the interim analysis in the procedure. 50% represents it occurs when the active trial reaching half of its target participant recruitment. 75% and 87.5% represent reaching those percentages in recruitment, respectively.

The presented performances measures were assessed with target type I error of 7.5% and target power of 75% as recommended by Psioda et al.

The study design achieved the prescribed type I error and power and inconsequential bias across the simulation. In our specific simulation, the trial almost always stopped early, regardless of the similarities or differences between the historical and active trials. This makes sense for multiple reasons. The simulation parameters that varied were covariate prevalence, prevalence in the active trial, correlation between covariates, the association between the covariates and outcome (the β_s), and the number of available historical controls relative to the active trial's treatment arms. Key components that were not varied in the data generation were the baseline hazard rate (which was the same in both trials) and the fact that the association (β_s) between the covariates and outcome were consistent in both trials (even if the parameter values varied, the same β_s were used in both the active and historical trials in each scenario). These parameters (β_s) are the nuisance parameters in the model. The log likelihood difference shown in (3.1.6) is sensitive to differences in the nuisance parameters comparing two cases: $\hat{\psi}$ when common values for the parameters is shared by both trials, $\hat{\psi}_1$ and $\hat{\psi}_0$ when two different values of the nuisance parameters are used in the active and historical trials. In the simulation, the data generating mechanism ensures that we are in the first case (i.e. common values for the parameters is shared by both trials, $\hat{\psi}$). Therefore, it makes sense that negative values of the interim analysis log likelihood difference statistic, W, are obtained and therefore trial is stopped early and the historical data is incorporated for the primary analysis.

An additional consideration is that in the procedure to determine the critical threshold value, w_0, the baseline hazard is estimated from the historical data and in the planning simulation. Possible concurrent data is simulated with a baseline hazard centered around this historical estimate (within 2 standard errors). Therefore, more lenient critical threshold values, e.g. w_0 is a small but positive value larger than 0, could be more likely due to the threshold calculation procedure producing data that is on average the same as the known data generating mechanism. Pooling the active trial data with the historical data seemed to not significantly reduce power or Type I error. More lenient critical values may have been allowed such that the procedure would always stop at the interim analysis and combine the historical and active trial samples.

When the active trial stopped at 50% enrollment, type I error and power would sometimes fall outside the traditional ranges (table 3.1.A). Less type I error and greater power could be achieved

in this case if the interim analysis occurred, later at 75% or 87.5%. It can also be noted that when more historical control participants were available (Hx size = 2), the power improved in table 3.1.A and 3.1.C. This number of historical control participants (Hx size) does not influence 3.1.B because the historical data is not used in this table.

Another observation is that when the trial continued to the final analysis opportunity, type I error and power exactly met their targets (3.1.B). This is reassuring because it confirms that in the case when the historical data is ignored, the active trial with complete recruitment acts as expected as essentially a traditional, randomized controlled trial.

When examining the adaptive trial design as a whole, the most appropriate measure is a weighted average of the performance measures at the interim analysis opportunity and at the final analysis opportunity. The weight of the interim analysis opportunity would be the stopping probability and the weight of the final analysis opportunity would be one minus the stopping probability. This gives performance measures on average for the entire procedure presented in 3.1.C. Even at its worst, percent bias was controlled to be no more than 2%. Type 1 error is almost always well controlled, occasionally going up to 7% in the null hypothesis with no differences (row 3 of 3.1.C). Power also met the target of 80% except it sometimes dipped down to 75% when the interim analysis occurred at half requirement. These slight decreases in power and small violation in type 1 error are well worth the trial reduction as seen in table 3.1.D. When the interim analysis was stopped halfway through recruitment, the average trial sample size reduction ranged from 21%-50%. This is a large reduction in sample size for a modest cost in power (3-5%) and possible small increase in type 1 error of up to 2%. Even when the trial stopped later, there were meaningful reductions in the required sample sizes. The adaptive trial method is able to accomplish its goal. It reduces the needed sample sizes without sacrificing important power and type I error requirements.

In summary, the simulation provides evidence that the procedure works as intended and is comparable to the methods used in 1.6. This adaptive trial design is a contender for use in a hypothetical study design that allows for external controls.

Chapter 4. APPLICATION IN CLINICAL TRIALS OF PULMONARY EXACERBATIONS IN CF PATIENTS AFTER *PSEUDOMONAS AERUGINOSA* INFECTION

4.1 BACKGROUND: THE EPIC AND OPTIMIZE TRIALS

Informed by the previous chapters, we sought to assess these analytic methods and adaptive study design in real CF clinical trial data. Therefore, we decided to harmonize the clinical trial data from two completed randomized controlled trials that focused on antibiotic therapy regimen aimed at reducing pulmonary exacerbations (PEx), the OPTIMIZE[18] and EPIC[17] trials. We would create a hypothetical situation where OPTIMIZE was a historically controlled trial with historical controls taken from the EPIC trial, using the approaches assessed in section 1.6 and Chapter 3.

The work of this dissertation that included the EPIC and OPTIMIZE trial data was approved by the institution review board at the University of Washington in Seattle and was classified as non-human subject research due to the fact that both trials were completed.

The Optimizing Treatment for Early *Pseudomonas aeruginosa* Infection in Cystic Fibrosis[18] clinical trial (OPTIMIZE) was a follow-on randomized, placebo-controlled trial of azithromycin, testing the efficacy and safety of 18 months of azithromycin vs. placebo added to the standard of care culture-based tobramycin treatment regimen evaluated in EPIC. The study was a multicenter, randomized, double-blind, placebo-controlled, trial in children ages 6 months to 18 years of age with a diagnosis of CF and a positive culture of *Pa* within the last 40 days of the baseline visit. Eligible participants were recruited from over 45 Cystic Fibrosis Foundation-accredited centers in the United States. A Cox proportional hazards regression was used to access the treatment effect of azithromycin on time to first PEx. At the planned interim analysis, azithromycin was associated with a statistically significant and substantial reduction in the risk of pulmonary exacerbation. Additionally, azithromycin was noted to improve weight gain but did not change the rate of acquisition of *Pa*. Due to these strong findings, the trial was stopped early

and led to improvements in treatment protocols for PEx in children as young as 6 months of age. In this dissertation, OPTIIMIZE is treated as the active trial, since it occurred later than EPIC.

The Early *Pseudomonas aeruginosa (Pa)* Infection Control trial (EPIC)[17] was a multifactorial trial design that assessed the key question: Does more aggressive cycled tobramycin inhalation therapy versus less aggressive culture-based tobramycin inhalation therapy achieve lower rates of PEx? Participants were children aged 1 year to 12 years with a diagnosis of CF and a documented *Pa* positive culture within 6 months prior to randomization. In addition to the primary outcome, secondary outcomes were assessed, including anthropometric measures (linear growth and weight gain), pulmonary function as measured by forced expiratory volume in 1 second, FEV_1, and safety. The study was designed to have at least 80% power to detect a clinically significant reduction of 40% in risk of PEx. During the study, no significant adverse events occurred, however no statistically significant nor meaningful difference in the rate of PEx or incidence of *Pa* was observed between the cycled and culture-based therapy arms. The EPIC trial participants can serve as historical controls because all participants received the standard of care, tobramycin, exactly like both the treatment and control arms of OPTIMIZE. The cycled and culture-based therapies were not statistically significantly different. Therefore, all arms in EPIC should have the same risk of PEx as the control arm in OPTIIMIZE.

Both trials were conducted by the CF Therapeutics Development Network (TDN) coordinating center and were conducted from 2004 to 2009 and 2014 to 2017 for EPIC and OPTIMIZE respectively. They had almost identical eligibility criteria[18,101]. These two clinical trials had the same primary endpoint of pulmonary exacerbations (PEx) requiring antibiotic therapy. Importantly, these trials are additionally similar enough to permit combining historic and concurrent controls, suggesting OPTIMIZE could have been a historically controlled trial. Specifically, the two studies largely meet Pocock's criteria[65], or a set of six guidelines proposed in 1976 for acceptability of historical controls as an external comparison group: 1) Both studies featured standardized treatment regimens with identical protocols, 2) trials were completed within five years with no substantial changes in treatment, 3) the same outcome definition was used by the researchers, 4) major eligibility criteria matched with the exception of OPTIMIZE participants recruited up to half a year younger and up to 6 years older, 5) both trials were

implemented by the same organization and researchers, and 6) there is no reason to suspect systematic confounding with the exception of changes in prevalence of CF therapeutic usage (**Table 4.1**). Additionally, the EPIC trial did not obtain statistically significant results when comparing their treatment arms. Therefore, all EPIC participants can be treated as a potential source of historical controls for the OPTIMIZE trial.

Table 4.1 Evaluation of the general conditions needed before combining historical and concurrent controls as recommended by Pocock.

Condition	**Comparison of EPIC and OPTIMIZE Trials**
Standardized treatment	Standard of care anti-PA antibiotic therapy for new onset *Pa* among all participants over the duration of follow-up included inhaled tobramycin. Differing intensities of treatment regimen with inhaled tobramycin with or without ciprofloxacin among historical controls were shown to have no impact on outcomes, enabling use of all EPIC trial participants and comparability of early *Pa* treatment between OPTIMIZE and EPIC studies.
Recent time interval	Trials were completed within the same decade with no major changes in treatment for early *Pa* infection over this time
Common method of evaluation	The same definition was used to define pulmonary exacerbation across trials.
Common patient characteristics	Major eligibility criteria match, requiring: • Diagnosis of CF • Age 1 to 12 years old • A first lifetime positive respiratory tract culture for *Pa* or a new positive culture for *Pa* with 2 years documentation of negative cultures • Clinically stable • Limited recent anti-*Pa* antibiotic therapy
Performed by same organization	Both trials were implemented at participating sites in the Cystic Fibrosis Therapeutics Development Network (CF TDN)
No systematic confounding	Standard of care was similar between cohorts except: • Use of chronic hypertonic saline became recommended • Modulator therapy became available for some CF patients

Reprinted from "A new path for CF clinical trials through the use of historical controls," by A Magaret, M Warden, N Simon, S Heltshe, G Retsch-Bogart, B Ramsey, N Mayer-Hamblett[39], 2022, Journal of Cystic Fibrosis, vol 21, pgs. 293-299, © 2021 European Cystic Fibrosis Society. Published by Elsevier B.V. All rights reserved. Adapted with permission. *Pa = Pseudomonas aeruginosa*.

Small differences did exist between the studies. For example, OPTIMIZE had participants that were a little younger or older than EPIC's eligibility criteria. In these cases, we simply restricted the OPTIMIZE or EPIC group such that the more restrictive eligibility criteria was met.

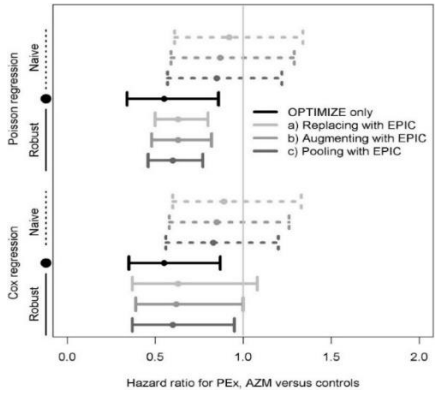
Therefore, the two trials are well suited for this dissertation.

In our preliminary work, we examined three possible strategies for incorporating historical control data to determine the focus of our future research[39]:

> "(1) Pooling: Combining all available participants from both studies to increase the size of the OPTIMIZE control group, (2) Augmenting: omitting half the OPTIMIZE participants randomized to placebo and augmenting the OPTIMIZE control group with all EPIC historical controls, (3) Replacing: omitting all OPTIMIZE participants randomized to placebo and replacing with all EPIC historical controls." (pg. 294)

We determined that all three approaches produced less biased treatment effect estimates and had tighter, more precise confidence intervals relative to their respective naïve approaches that did not consider differences between the trials. Due to the goals of making more logistically feasible trials, the pooling approach was eliminated from consideration in this dissertation because it does not reduce the burden of the active trial. Similarly, regulators likely will still require some form of concurrent controls, therefore we decided to focus both the simulation and the implementation in real trial data on the augmenting approach. We demonstrated that inverse probability weighting utilized historical controls and worked well in this hypothetical setup where OPTIMIZE is the active trial and EPIC is the source of historical controls (**Figure 17**). Also, Poisson regression worked just as well as the Cox proportional hazard models used in the original trials in terms of bias and precision while allowing a fully parametric approach.

Figure 17. Pooling, augmenting, and replacing in the EPIC and OPTIMIZE trials

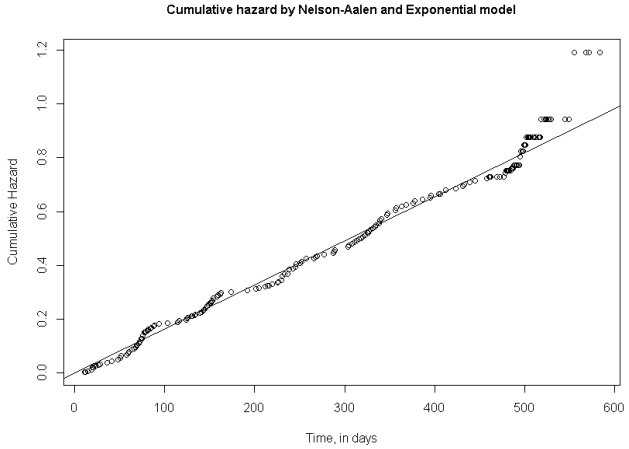


Reprinted from "A new path for CF clinical trials through the use of historical controls" by Amalia Magaret, Mark Warden, Noah Simon, Sonya Heltshe, George Retsch-Bogart, Bonnie

Ramsey, and Nicole Mayer-Hamblett, 2022, Journal of Cystic Fibrosis (21) Copyright © 2021 European Cystic Fibrosis Society. Published by Elsevier. DOI: 10.1016/j.jcf.2021.11.007 Reprinted with permission. Treatment effect estimates for the effect of azithromycin on time to first PEx. Naive approaches, shown with dashed lines, combine EPIC and OPTIMIZE participants into a single dataset without any adjustment. Inverse probability weighting was used and indicated as "robust." The solid black line is the treatment effect estimated in the full OPTIMIZE trial.

It was especially important in our preliminary work to confirm that the use of a generalized linear model (GLM) with log link function and Poisson likelihood was appropriate for these trial data. The original publications delineating the approaches of Chapter 2 and Chapter 3 all use the generalized linear model framework (with plans to expand to semi-parametric methods, such as Cox proportional hazards model, not yet fully implemented). Furthermore, when planning the simulation study, an exponential model was selected as the data generating mechanism for the time to event variable. An exponential model implies a constant hazard rate (λ) and by extension a linear cumulative hazard function. We graphed both the Nelson-Aalen, non-parametric estimate of the cumulative hazard function for first pulmonary exacerbations, as well as the expected cumulative hazard function using an estimated baseline hazard rate of a Poisson GLM regression in the OPTIMIZE trial data. The trial data supported the assumption of a constant baseline hazard rate central to this dissertation and removed the need to adapt the methods to a Cox proportional hazards model. We also replicated this finding in EPIC (not shown).

Figure 18. Empirically observed constant hazard rate (slope) in OPTIMIZE



Modeling time to event using a Poisson generalized linear moel (GLM) is equivalent to assuming a constant base hazard rate, which implies a linear cumulative hazard function. The plot above graphs both the empirical (Nelson-Aalen) estimate of the cumulative hazard

function for time to first PEx in OPTIMIZE and the implied linear cumulative hazard rate when modeling the data using a Poisson GLM.

These conclusions were foundational in the design of this dissertation that expands on this prior work. We decided to focus exclusively on the augmenting case, explored the additional Bayesian methods discussed in previous chapters, and embraced a generalized linear model with Poisson link to analyze this outcome data. Augmenting is the natural choice because regulators will likely require at least some concurrent controls and pooled analyses do not address the logistical issues we are trying to avoid.

4.2 HARMONIZATION OF TRIAL DATA AND THE DEVELOPMENT OF THE PROPENSITY SCORE MODEL

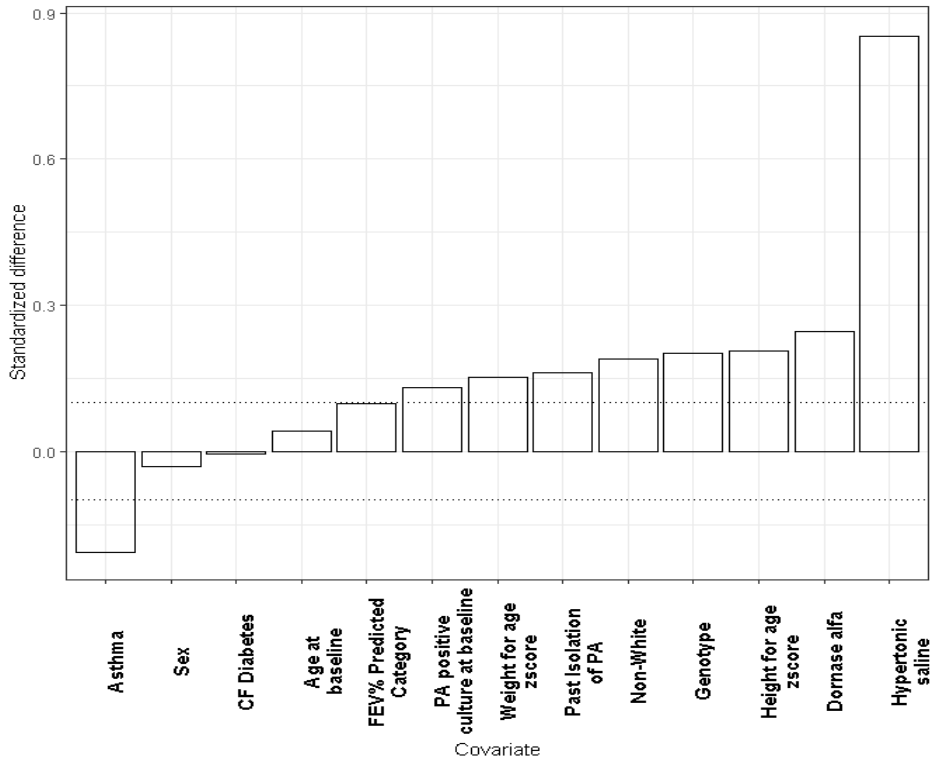
Although the EPIC and OPTIMIZE trials were conducted by the same organization with standardized definitions and protocols, there were still many differences when comparing the case reporting forms, data dictionaries, and raw datasets of the two trials. Correctly combining the trial data into a single analytic dataset required careful review of the trials' protocols and variable definitions. In situations where important variables did not have the same definition in both trials or when certain variables were never explicitly calculated in one trial, data harmonization was required to prepare for the historically controlled trial methods.

In the OPTIMIZE trial, two key airway treatments including dornase alfa and hypertonic saline were explicitly recorded because they had become routine for many CF patients. However, in the EPIC trial occurring five years earlier, these treatments existed but hypertonic saline use had not become widespread and dornase alpha was not explicitly recorded. For this reason, indicator variables for these treatments in the EPIC trial did not exist and had to be manually derived from text fields. For both trials, extensive documentation of concomitant medications and treatments was standard protocol. Therefore, in EPIC, a careful, time-intensive review of the text entries of trial participants' medical histories and current concomitant medications and treatments allowed for the creation of these needed variables for the harmonization. The same process was applied for the creation of CF-related diabetes and asthma variables in both trials because expert

knowledge available at the CF therapeutics development network coordinating center determined that these variables could be important factors relevant to pulmonary exacerbations that could be different between the two trials. Additionally, definitions of *Pseudomonas aeruginosa* history and status at the start of the trials varied between EPIC and OPTIMIZE. Therefore, recalculation of these variables was required within both trial datasets. Finally, there was a subset of OPTIMIZE participants that had ages too young or old relative to the exclusion criteria in the EPIC trial. Therefore, we restricted the OPTIMIZE participants to those only within the age criteria set by the EPIC trial.

Once the trial data was harmonized and combined into a single analytic dataset, we examined differences between the trials using descriptive statistics as well as visually using standardized differences, a graphical approach proposed by Austin et al.[81] (**Figure 19**).

Figure 19. Standardized differences of variables in the EPIC and OPTIMIZE trials



Differences between the EPIC and OPTIMIZE trials on the above covariates were represented on a comparable scale. Specifically, the differences in mean values of the covariates between the studies are appropriately scaled by their pooled estimated standard deviations. Higher standardized differences correspond to larger means in the OPTIMIZE

trial. As a rule of thumb, differences above 0.1 or below -0.1 are considered unbalanced as indicated in dotted lines. Details are available in Austin (2009).

Table 4.2 Demographics and relevant variables in EPIC and OPTIMIZE

	Stratified by study		
Characteristics	**Epic (N=304)**	**Optimize (N=163)**	**P-value[1]**
Female (%)	154 (50.7)	80 (49.1)	0.82
Race or Ethnicity (%)			**0.01**
White	285 (93.8)	144 (88.3)	
Hispanic	4 (1.3)	11 (6.7)	
Black	7 (2.3)	6 (3.7)	
Asian or Pacific Islander	1 (0.3)	1 (0.6)	
Other	7 (2.3)	1 (0.6)	
Age at baseline, years (mean, sd)	5.70 (3.51)	5.84 (3.45)	0.67
Age Category[2] (%)			0.51
1-3	71 (23.4)	45 (27.6)	
>3 to 6	109 (35.9)	59 (36.2)	
>6 to 12	124 (40.8)	59 (36.2)	
Genotype			0.65
Delta F508 Heterozygous	149 (49.0)	86 (52.8)	
Delta F508 Homozygous	116 (38.2)	62 (38.0)	
Other	33 (10.9)	12 (7.4)	
Unknown	6 (2.0)	3 (1.8)	
Height for age z-score at baseline, CDC equations[3] (mean, sd)	-0.50 (0.93)	-0.30 (1.02)	**0.03**
Weight for age z-score at baseline, CDC equations[3] (mean, sd)	-0.33 (0.95)	-0.18 (1.04)	0.11
FEV_1 percent predicted at baseline, Wang-Hanson equations[3,4] (mean, sd)	96.18 (16.68)	94.13 (16.59)	0.34
FEV_1 category[3,4]			0.50
< 6 years of age	173 (56.9)	91 (55.8)	
<90% predicted	40 (13.2)	24 (14.7)	
≥ 90% predicted	87 (28.6)	48 (29.4)	
Chronic Dornase Alfa use at baseline[5] (%)	163 (53.6)	107 (65.6)	**0.02**
Chronic Hypertonic Saline use at baseline[5] (%)	13 (4.3)	58 (35.6)	**<0.001**
Modulator use at before or during the study (%)	0 (0)	17 (10.5)	**<0.001**
Chronic Azithromycin use at baseline[6] (%)	2 (0.7)	N/A[5]	N/A[5]
CF related diabetes at baseline (%)	2 (0.7)	1 (0.6)	1
Asthma at baseline (%)	51 (16.8)	16 (9.8)	**0.05**
PA positive culture at baseline (%)	118 (38.8)	72 (44.2)	**0.01**
Past Isolation of PA[7]	96 (31.6)	64 (39.3)	0.12

[1] A t-test assuming unequal variance was used for continuous variables and a chi-squared test with a continuity correction for binary and categorical variables.
[2] Age categories are defined the traditional way: For example, 3.5 years old would be considered 3 in this categorization.
[3] There were at most 4 missing values.
[4] FEV1was not measured for participants 0-5 years of age, and are therefore not included in the continuous variable.
[5] Chronic use is defined as starting the treatment ≥31 prior to baseline with the intent of continuing it during the study.
[6] Optimize's central question was regarding Azithromycin use and therefore chronic Azithromycin use is not reportable.
[7] The positive Pa culture used for eligibility into the trial was newly acquired after a 2 year history of negative Pa cultures, and there was at least one Pa positive culture prior to this window since birth.
Bolded terms were statistically significantly different between the two trials.

Descriptive statistics were calculated for all participants, stratified by source study (EPIC,

OPTIMIZE). All of EPIC was grouped together because there was no evidence of a difference between the study's arms. Statistically significant differences are bolded.

Hypertonic saline use dramatically increased during the time between the trials with rates of 35.6% in OPTIMIZE and only 4.3% in EPIC. Dornase alfa also had a slightly higher use rate of 65.6% as compared to 56.9% respectively. There were more non-white participants (11.7% vs. 4.9%), higher rates of PA positive cultures at baseline (44.2% vs. 38.8%) and higher height for age z-scores on average (-0.3 vs -0.5) in OPTIMIZE. CFTR modulators were not yet available for participants in EPIC but 17 (10.5%) of participants reported use in OPTIMIZE. Therefore, we determined that we would conduct analysis excluding those who reported modulator use as a planned sensitivity analysis. Finally, rates of asthma at baseline (manually derived for both trials in the harmonization process), were slightly higher in the EPIC study (16.8% vs. 9.8%).

When building the propensity score model needed for inverse probability weighting and the propensity-score based power prior method, we followed the example set by Brookhart et al. The authors conclude that a model with all variables thought to be related to the outcome regardless of their relationship with the exposure of interest is the best approach.[102] This contrasts with building the best predictive model (perhaps by stepwise regression). Their conclusions align with the best practices outlined by Austin et. al. who suggest that "statistical hypothesis testing in the analytic sample not be used to identify the requisite variables [for the propensity score].[81]" The reasoning for this recommendation is that analytic sample may not be powered to detect all important prognostic or confounding covariates because the study is powered for the outcome. Therefore, expert opinion and disease etiology should be the primary determinant of variable inclusion in a propensity score model. For the EPIC/OPTIMIZE trial data, we derived and harmonized variables identified as important by primary investigators and MDs with experience in pulmonary exacerbations.

We fit a propensity score model using logistic regression on the binary outcome of participation in the active trial, OPTIMIZE, vs the historical trial, EPIC (**Table 4.2**). We included all variables in **Figure 19** with a few small exceptions. We did not include CF related diabetes at baseline because only 3 people (2 in EPIC and 1 in OPTIMIZE) had this condition. Additionally, we only adjusted for FEV_1 as a categorical variable because of complexities with defining FEV_1 in

children under the age of 6. We only adjusted for height for age z-score instead of both height for age z-score and weight for age z-score at baseline because these values were highly correlated and measuring a similar phenomenon. Finally, we did not adjust for CFTR modulator use before or during the study because only 10% had any use in the OPTIMIZE study and none did in the EPIC study. However, the influence of modulator use was considered in a preplanned sensitivity analysis.

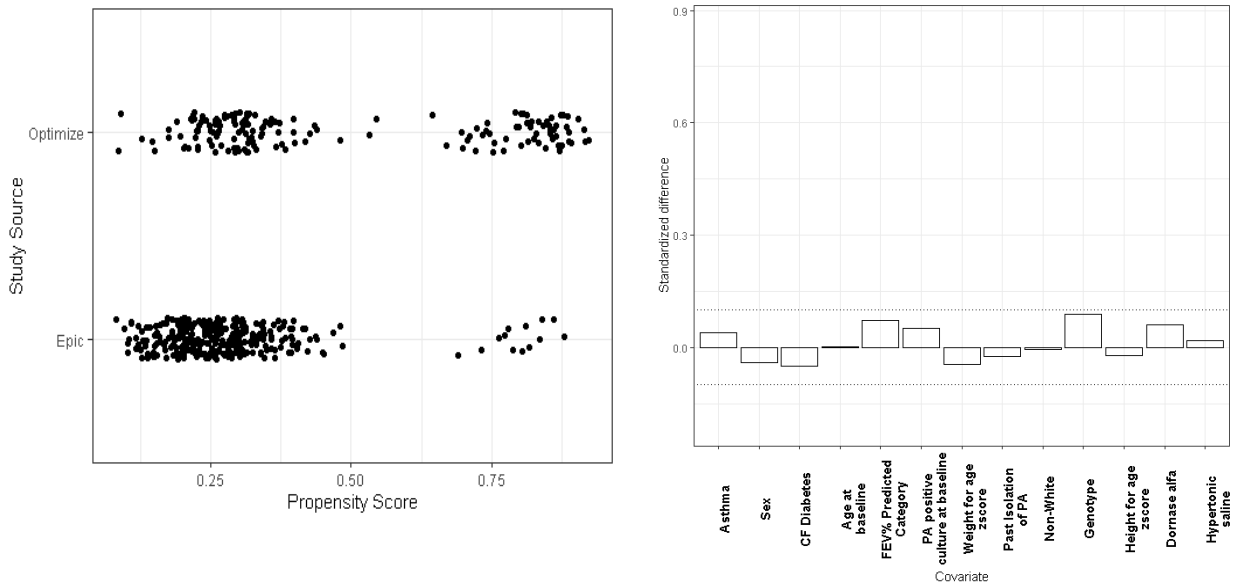
Table 4.3. Final propensity score model covariate estimates

Model Parameters	**Estimate (OR, 95% CI)**	**P-value**
Age >3 to 6[1]	Reference	Reference
Age 0-3[1]	1.12 (0.59,2.14)	0.74
Age >6 to 12[1]	0.99 (0.41, 2.41)	0.99
Female	1.30 (0.84, 2.02)	0.24
Height for age z-score, CDC equations	1.25 (0.93, 1.67)	0.14
Non-White	2.03 (0.94, 4.4)	0.07
F508del Homozygous	Reference	Reference
F508del Heterozygous	0.76 (0.47, 1.23)	0.26
Other Genotype	0.57 (0.24, 1.32)	0.19
Genotype Unknown	0.55 (0.11, 2.83)	0.47
FEV not available (<6 years of age)	Reference	Reference
FEV ≥ 90 percent predicted	0.92 (0.34, 2.52)	0.88
FEV <90 percent predicted	0.79 (0.27, 2.35)	0.67
Chronic hypertonic saline use at baseline[2]	11.6 (5.87, 22.91)	<0.0001
Chronic dornase alfa use at baseline[2]	1.26 (0.79, 1.99)	0.33
Asthma at baseline	0.56 (0.28, 1.11)	0.10
PA positive culture at baseline	1.39 (0.90, 2.16)	0.14
Past Isolation of PA[3]	1.14 (0.68, 1.93)	0.62

[1] Age categories are defined the traditional way: For example, 3.5 years old would be considered 3 in this categorization.
[2] Chronic use is defined as starting the treatment ≥31 prior to baseline with the intent of continuing it during the study.
[3] The positive Pa culture used for eligibility into the trial was newly acquired after a 2 year history of negative Pa cultures, and there was at least one Pa positive culture prior to this window since birth.

We examined standardized differences between covariate values in the two trials at baseline and after inverse probability weighting by the final propensity score. Standardized differences compare key variables between the trials by setting all the variables regardless of their type (binary, categorical, continuous) to their respective values on a single metric allowing for comparison (see Austin et al. for details). The purpose of examining the variables in this manner is to assess whether the propensity score well balances the combined trials' data in terms of the observed covariates. This is a fundamental property of a well-modeled propensity score and an easy criterion to assess (Figure 19 and the right panel of Figure 20).

Figure 20. Graphical diagnostics of the propensity score



The propensity score did not suffer from violations of positivity. Specifically, high and low propensity scores were observable in both source studies (left). Additionally, the propensity score succeeded in balancing the covariates as seen on a second standardized differences graph (right).

Another important consideration when using propensity score-based methods is to ensure that there are no positivity violations. The positivity assumption is that at all levels of confounders participants exist from both exposure groups (active trial and historical trial participants). This is true for all variables used in the propensity score model. However, it is also important that this property is reflected in the distribution of propensity scores as well. There cannot be a subset of very high or low propensity scores in either arm. This would represent a certain covariate profile seen in one group but not the other and therefore would violate the assumption. The consequence of this violation is that the treatment effect estimate is biased by a covariate profile group that exclusively exists in one of either the treatment or control arm. For example, suppose there was a subset of asthmatic, F508del genotyped individuals who did not use hypertonic saline (a specific covariate profile) in the treatment arm. This imbalance would bias the treatment effect estimate because there is no way to accurately estimate the counterfactual outcome of this subset as if they were in the control (or historical control) group. Fortunately, as can be seen in the left panel of **Figure 20**, a good range of high and low propensity scores exist in both trials and there are no problematic high or low clusters. There is no evidence of a violation of positivity, justifying a

correctly estimated treatment effect. Additionally, the propensity score plot shows that there are no propensity scores with values near 0 or 1, meaning that inverse probability weights will not have any strange behavior such as having extremely large or small weights.

After the trials' datasets were harmonized and combined into a single analytic dataset and a well-modeled propensity score demonstrating characteristics of a good fit was developed, then it was time to implement all the historically controlled trial methods in real data.

4.3 ANALYTIC METHODS' PERFORMANCE IN TRIAL DATA

Preliminaries

For implementation of the three analytic methods discussed in section 1.6, there are a few important caveats unique to each of the methods. One issue that affected all the methods was the presence of missing covariate data. For the propensity score-based methods, missing covariate data meant that the desired propensity score model could not be used. In the EPIC trial, there were 4 participants for whom FEV_1 data was not recorded and 2 participants who did not have height nor weight data. In the OPTIMIZE trial, there were 4 participants who did not have a valid *Pa* culture at baseline. The result was a total of 10 participants with missing data in the analytic dataset.

To address the missing data problem in inverse probability weighting, we explored two options: 1) assigning a weight of 1 to the 10 participants thereby neither upweighting or down weighting their influence and 2) Fitting a second propensity score using all the data that did not adjust for FEV_1 category, height for weight z-score, nor PA culture status at baseline and then calculate the inverse probability weight for only these 10 participants using this simpler propensity score model. Interestingly, both options produced the nearly identical treatment effect estimates (0.60 vs. 0.59) and identical confidence intervals for that effect.

The outcome model for the inverse probability weighting approach was the same as was done in the original OPTIMIZE trial. It contained only age, adjusted for as a categorical variable as

presented in table 4.3, and the treatment effect, weighted by the inverse probability weights.

For the propensity score-base power prior method, we used option 2 discussed above so that all participants had a propensity score. However, the approach had another issue to address. The approach has two key steps, a matching step that creates the analytic sample and then the Bayesian analyses that uses the propensity score-based power prior in this sample. In the simulation, we expanded the eligible historical control participant pool such that it was four times as large but had the final analytic sample in the Bayesian analysis step be the same size as the other methods for a fair comparison. Therefore, for this application in real trial data, we decided to let the EPIC trial (n=304) represent the whole historical control pool and that we would select approximately 231 (~75%) of this pool corresponding to exactly 3 times the treatment arm in OPTIMIZE (n=77). This would mean that the matched sets would avoid possible bad matches that could occur if the entire pool was used but at the same time this may run the risk of underutilization of the EPIC data. This is simply a limitation of the propensity score-based power prior method.

The commensurate prior method requires more specifications than inverse probability weighting and the propensity score-based power prior method. Specifically, it requires specification of the spike and slab priors for the τ parameters discussed in **section 2.3**. Fortunately, in the case of real trial data, the calibration of the spike and slab can be directly informed and determined by the historical data. Recall that the spike exists to help the model converge in the case that the parameters between the historical and active trials are commensurate while the slab covers the case in which they are not. Therefore, a straightforward approach is to fit the model in the historical data (EPIC) and let the standard error of the parameters help determine the value of the spike. Then, for the slab, let the upper part of the slab correspond to twice the standard error of the EPIC model's respective parameters. Regarding the missing data, we simply excluded the 6 EPIC participants and 4 OPTIMIZE participants that had missing covariate data because this was simple and straightforward. Bayesian hierarchical modeling has the option of treating missing covariate values as parameters to be estimated in the model, which is a fascinating approach but something we did not wish to explore in this dissertation.

In the table below, for the augmenting approach, it represents the hypothetical case where OPTIMIZE was able to recruit only half of its controls. To accomplish this, we randomly sampled half of the controls to use in the analytic dataset 20 times, and for each dataset we conducted the methods. We were able to combine the 20 results using multiple imputation methods developed by Rubin.[103] A key assumption of the imputation methods to combine the results is that the results are asymptotically normality. Because of the Bayesian central limit theorem[99], we can use the posterior's mean and standard deviation in the imputation formulas. Another approach for future research would be to use bootstrap estimates for confidence intervals and probabilities; however, we opted for the simplicity of Rubin's rules. When we fit the models, a visual inspection also confirmed that the posteriors did approach an approximately normal distribution with mild if any skew.

Table 4.4 Analytic method's performance in EPIC/OPTIMIZE hypothetical historical trial

SCENARIOS	OPTIMIZE AZM (n=77) vs.	Treatment Effect		
		Inverse probability weighting HR (95% Confidence Intervals) P-value	Propensity score-based power priors HR (95% Credible Intervals[1]) Posterior probability[2]	Commensurate priors HR (95% Credible Intervals[1]) Posterior probability[2]
OPTIMIZE Only[3]	N=86 OPTIMIZE Placebo	0.55 (0.34, 0.86) P=0.01	0.62 $[0.39,0.99]_{HPDI}$ PoP=0.04	0.52 $[0.31,0.86]_{HPDI}$ PoP=0.01
Pooling OPTIMIZE with EPIC	N=86 OPTIMIZE Placebo with N=304 EPIC	0.60 (0.46,0.77) P<0.001	0.65 $[0.42,0.99]_{HPDI}$ PoP=0.04	0.53 $[0.33, 0.87]_{HPDI}$ PoP<0.001
Augmenting OPTIMIZE with EPIC[4]	N=43 OPTIMIZE Placebo with N=304 EPIC	0.63 (0.48, 0.82) P<0.001	0.65 $[0.41, 1.03]_{HPDI}$ PoP=0.06	0.58 $[0.32, 1.05]_{HPDI}$ PoP=0.07

[1]Credible intervals are highest posterior density intervals (HPDI) and are interpreted differently than confidence intervals.
[2]PoP is defined as the posterior probability of the null hypothesis, and Pr is defined as probability taken from the posterior distribution. Mathematically, $PoP = 2 * [1 - \max(\Pr(HR < 1|D), \Pr(HR > 1|D)]$. This definition, proposed by Shi et al.[100], aligns with two-sided hypothesis testing and by extension two-sided p-values.
[3]For this row, corresponding standard techniques were used to estimate the treatment effect.
[4]Multiple imputation methods suggested by Rubin[103] were used on 20 random samples where half of the active concurrent trial controls were utilized in the analysis to estimate the augmenting scenario.
AZM=Azithromycin, HR=Hazard ratio

A sensitivity analysis was conducted to determine if the presence of the 17 participants with CFTR modulators influenced the results of the historically controlled trial methods. When the analysis was rerun with these participants excluded, no consequential changes were observed in the treatment effect estimates. The sensitivity analysis showed that CFTR modulator use was not

largely impactful for this short-term outcome, and therefore did not lead to a positivity violation due to the presence of these 17 participants in OPTIMIZE.

Summary

For inverse probability weighting (IPW), both pooling ($p<0.001$) and augmenting ($p<0.001$) had statistically significant treatment effect estimates. However, the pooling and augmenting estimates were slightly biased towards the null, suggesting that the weighted sample was unable to account for all the differences between the trials even after balancing. When considered as a whole, IPW performed acceptably in terms of bias and power, and achieved success in reducing the size of the active trial, concurrent control arm participants in half. An important caveat to the augmenting approach is that the variance used in calculation of the confidence interval and p-value may be underestimated[104] because of influence from both the propensity score estimation and the imputation step. Presenting bootstrapped confidence intervals and p-values may avoid this problem in future research.

When looking at the propensity score-based power prior (PSPP) method in the augmenting approach, the treatment effect did not obtain statistical significance (PoP=0.06) despite being close to the threshold. Interestingly, even when the full OPTIMIZE data was considered alone, the treatment effect was biased towards the null in a way not seen in the other two methods: 0.62 [0.39, 0.99] compared to 0.55 (0.34, 0.86) for IPW and 0.52 [0.31, 0.86] for commensurate priors. The propensity score-based power prior method on the OPTIMIZE data alone was a simple model with an intercept and treatment effect that modeled the cumulative event rate with non-informative priors for both parameters. This model may not be well suited to answer the research question even in the original OPTIMIZE randomized controlled trial data, even if it did detect a statistically significant treatment effect in the OPTIMIZE only (PoP=0.04) and pooling scenarios (PoP=0.04). Therefore, the augmenting approach may have only increased the problem of a poor fitting model rather than taking advantage of the availability of more data. The approach may have performed better with a different model formulation. However, lack of covariate adjustment is a critical limitation of the current implementation of propensity score-based power priors and an area of possible future research. This approach failed to achieve the goals of an unbiased and well powered model when the OPTIMIZE control arm has reduced to

half the size of the original trial.

The commensurate prior model performed well when incorporating the EPIC data in both the OPTIMIZE only (PoP=0.01) and pooling scenario (PoP<0.001). The pooling point estimate was also the most accurate of the three methods, maintaining a confidence interval like the gold standard original trial results. However, in the augmenting scenario, it was unable to reach statistical significance (PoP=0.07) despite achieving the least biased point estimate of 0.58 [0.32, 1.05]. The large confidence interval is likely due to the complexity of the model that simply needed more data to reduce the variance. Of the three approaches, the commensurate prior model had 30 nuisance parameters (an intercept and 14 covariates for both trials) in contrast to the 3 nuisance parameters (the intercept and age as a 3-level categorical variable with 1 reference level) in IPW and 1 nuisance parameter (the intercept) in PSPP. A unique feature of the commensurate prior method not shared by the other two methods is that it allowed the historical controls to have different nuisance parameters. Most importantly, the baseline hazard rate (intercept) was allowed to be different between the historical and active trial participants as well as certain covariate parameters such as the age-categories. In certain situations, the model flexibility may have been advantageous, but it does cost additional parameters.

An interesting observation when comparing the OPTIMIZE only scenario to the pool scenario is that the confidence interval greatly decreases for inverse probability weighting but the credible intervals for the Bayesian approaches do not change much. This is also reflected in the augmenting scenario as well. The lack of a narrowing interval estimate is likely due to the difference between a confidence interval and a credible interval. Recall that a confidence interval represents a range of values that could be reasonably observed under repeated sampling while credible intervals represent a 95% probability the true effect estimate lies within the interval. In essence, the Bayesian model is a convolution of two distributions, the difference between the treatment arm's log-hazard rate and the control arm's log-hazard rate. Adding historical information improves the characterization of the control arm's log-hazard rate probability distribution by increasing the density around a value. For the treatment effect, no additional information is added and therefore its probability distribution remains unchanged. When considered together, the convolution of these distributions only slightly improves the credible

interval because to further narrow it the treatment arm's event rate probability distribution is more influential and must be better estimated.

Although in this application the commensurate prior method was slightly underpowered, it produced the least biased point estimate when the number of concurrent controls were halved. In a different trial where the hierarchical model has fewer adjustment variables and may have important differences in the nuisance parameters, this method may be well fit to model the treatment effect.

When considering the three approaches performance in the real trial data, inverse probability weighting is the most appropriate for EPIC/OPTIMIZE. First, the trials are extremely similar. The simulation showed that IPW shines in this circumstance, where it achieved power much greater than 80%. The results of the augmenting approach agree with this conclusion.

Considering the simulation and real trial results, inverse probability weighting likely works better when trials are congenial. As the historical trial data gets less similar and relevant to the active trial data, the commensurate prior may be better suited. The propensity score-based power prior may have potential but needs further developments to complete with the other two methods.

4.4 Adaptive study design's performance in trial data

For the adaptive trial design, we began with the historical trial data (EPIC) and started the planning and simulation procedures outlined in Chapter 3. The goal of the planning phase is to identify an appropriate log likelihood difference threshold that is justified based on prespecified type I error and power. This threshold is used at the interim analysis to determine if stopping the trial early and pooling the historical controls with the active trial is appropriate for the primary analysis. With the EPIC data, this phase involves simulating possible versions of active trial data using the distributions of variables observed in EPIC.

For every variable in the EPIC data, we first transformed non-continuous variables into integers

corresponding to each value of a category. With the inclusion of continuous variables, all variables were now numeric and therefore we could calculate means, variance, and pairwise covariances for the entire dataset in traditional ways. The purpose of this was to parameterize one large multivariate normal distribution from which to simulate possible active trial datasets. Then after generating data using this random multivariate normal distribution, we transformed the simulated variable values that were originally categorical or binary back to their original forms. We used a categorizer function that takes numeric values and re-categorizes them back into distributions like the original data. This simulation procedure as well as the categorizer function were developed by Tannenbaum et al.[105] and are convenient for simulating complex and large datasets with a mix of continuous, binary, and categorical variables.

The outcome variable, simulated time to first pulmonary exacerbation, was generated from an exponential distribution with a rate modeled by a linear combinations of the covariates and a baseline hazard rate (intercept). The coefficients (β_s) for the covariates were the point estimates when a model was fit to the EPIC data and these same values were always used in all simulated datasets. However, the baseline hazard rate, β_0, was randomly generated for each dataset based on the distribution of the baseline hazard estimated in the EPIC data. Specifically, it was taken from a normal distribution centered at the baseline hazard estimate for the EPIC data with a standard deviation equal to this baseline hazard's parameter's estimated standard error. Therefore, all simulated active datasets were slightly different in terms of their baseline hazard rates but the covariates and covariates association with the outcome was always the same.

We simulated 4,000 possible active datasets as described above and measured operating characteristics of the interim and final analysis under a set of interim analysis timings (at 50%, 75%, and 87.5%) and treatment effects (HR=1 vs. HR=0.54). Using the procedure outlined in Psioda et al.[98] and discussed in Chapter 3, the critical log likelihood difference threshold value, w_0, was calculated at a value of -0.1, meaning we needed at least some evidence of similarity between the trial parameters to stop the trial early, incorporate the historical controls, and do the primary analysis. Recall that as w_0 goes from zero to negative infinity, the similarity requirement (both trials can be modeled with a common set of parameters as measured by likelihood) becomes more stringent.

Once the critical threshold value was calculated, we subsetted the OPTIMIZE data to represent the interim analysis at 50% recruitment. We calculated the two necessary log likelihoods and took their difference (discussed in detail in Chapter 3) and compared the result to the threshold value, w_0, of -0.1. The difference in log likelihoods (a difference of 9.03) was much larger than the threshold, w_0. Recall that a large log-likelihood difference means there was strong evidence of differences between the nuisance parameters in the historical data and the active trial data. The model that allowed two sets of nuisance parameters performed much better than a model with a shared set. Therefore, we concluded that the study could not be stopped early and needed to continue to full recruitment. In fact, we calculated the estimated treatment effect anyway and found a hazard ratio of 0.96 with a p-value of 0.89, which we know is very different from the complete OPTIMIZE results (**Table 4.5**).

Table 4.5 Adaptive trial analysis models at interim analysis

Model Parameters	**Common set of Parameters**		**Different sets of nuisance parameters**			
			Active Trial at interim analysis (OPTIMIZE)		**Historical Trial at interim analysis (EPIC)**	
	Estimate	**P-value**	**Estimate**	**P-value**	**Estimate**	**P-value**
Intercept[1]	**-0.53**	0.21	**-0.31**	0.63	**-0.57**	0.32
	OR	**P-value**	**OR**	**P-value**	**OR**	**P-value**
Azithromycin (treatment)	0.96	0.89	0.88	0.72	-	-
Age >3 to 6[1]	Reference	Reference	Reference	Reference	Reference	Reference
Age 0-3[1]	**1.41**	0.35	**2.20**	0.15	**0.86**	0.77
Age >6 to 12[1]	**2.73**	0.11	**6.81**	0.07	**1.68**	0.52
Male sex	1.22	0.40	1.44	0.29	0.96	0.92
Height for age z-score, CDC equations	0.93	0.56	0.89	0.46	0.99	0.95
Non-White	1.29	0.53	1.55	0.38	0.71	0.66
F508del Heterozygous	N/A	Reference	Reference	Reference	Reference	Reference
F508del Homozygous	0.79	0.37	0.55	0.12	0.95	0.88
Other Genotype/Unknown	0.57	0.20	0.30	0.13	0.88	0.81
FEV not available (<6 years of age)	N/A	Reference	Reference	Reference	Reference	Reference
FEV ≥ 90 percent predicted	0.93	0.32	0.38	0.39	0.46	0.38
FEV <90 percent predicted	0.53	0.40	0.23	0.22	0.82	0.83
Chronic hypertonic saline use at baseline[4]	**2.22**	0.01*	**3.06**	0.003*	**0.75**	0.77
Chronic dornase alfa use at baseline[4]	**0.81**	0.41	**0.55**	0.11	**1.18**	0.64
Asthma at baseline	**1.33**	0.41	**0.67**	0.60	**1.91**	0.11
PA positive culture at baseline	0.84	0.46	0.70	0.29	0.97	0.93
Past Isolation of PA[6]	**0.93**	0.79	**0.69**	0.36	**1.27**	0.56

[1]The intercept is presented on the log-scale because it represents a participants' starting log odds of a pulmonary exacerbation and does not make sense to present as an odds ratio.
*p-value<0.05

At the interim analysis, models are presented to compare the trials in terms of nuisance parameters. One model has a common set of parameters shared by both trials. The other allows the trials to have different parameters. Parameter estimates are presented to provide a sense of why the trial did not stop at the interim analysis. Bolded parameters are ones that may have contributed to the algorithm's decision to continue recruitment in the active trial.

When comparing the active trial (OPTIMIZE) to the historical trial (EPIC), there are a few notable observations. First, the parameter estimate for chronic hypertonic saline dramatically differs between the trials. Other nuisance parameter estimates also vary in magnitude and/or direction between the trials, including for the two categorical age variables, chronic dornase alfa use, asthma, past isolation of PA, and the intercept. These differences, although not statistically significant, do contribute greatly to the log-likelihood difference estimates used in the stopping rule. There appears to be evidence that without analytic adjustment (like in Chapter 2), it is inappropriate to combine these two trials' data. Adjustment parameters do not appear to be consistent suggesting that directly combining the trials could lead to confounding.

Therefore, the study continued to completion and the EPIC data was ignored. The full OPTIMIZE study achieved a treatment effect of 0.53 (0.32, 0.87) with a p-value of 0.01, achieving statistical significance in the same way as the original trial.

The application of the adaptive trial in the EPIC/OPTIMIZE data suggests possible considerations for use of the method in future work. There is no adjustment (other than discounting by α) when assessing the active and historical trial data at the interim analysis. We know from section 4.2 that there are key differences between the trials including rates of chronic hypertonic saline use and asthma that could lead to different model estimates for these parameters in each of the trials. For this reason, it might make sense to only consider a matched set of historical controls to be used at the interim analysis rather than the full historical trial (all of EPIC). A matched set may have nuisance parameter estimates that are more aligned and increase the likelihood of early stoppage and correct inference. In this dissertation, we recreated

what was done in Psioda et al. exactly, but in future work we may consider this possible improvement.

Chapter 5. CONCLUSIONS

5.1 SUMMARY

The work of this dissertation is the first of the authors' knowledge to directly compare methods to conduct historically controlled trials in a rigorous setting. As the literature continues to develop, analyses like those conducted in this work are important for the development of guidelines for future historically controlled trials and to provide direction for future improvement and development of methods, especially in cystic fibrosis.

The simulation study provided a reproducible and rigorous way to compare and explore the methods. It also confirmed that all the methods could perform acceptably in the right conditions. It reinforced some intuitive but critical observations about these methods. Both inverse probability weighting and propensity score-based power priors rely on a well-modeled propensity score to perform well. In the case where the two trials are different and the propensity score or matching technique used to create exchangeable comparisons has problems, then the whole primary analysis will be affected. The commensurate prior method seemed the most conservative in the simulation study, and, in the application in real data, this also seemed to be true. It is suspected that the performance of the method is connected to the primary outcome model. When the primary analysis is well modeled, including all key variables, and has a good fit in both the active and historical trials, then the commensurate prior method can perform very well. However, its complexity can lead to greater uncertainty and higher variance.

Considering the performance of the methods in both the simulation study and in real data as well as their complexity, the conclusions at the end of the simulation study are reinforced. Among the analytic approaches, the commensurate prior method is most flexible in handling many types of generalized linear models and performs well under many conditions. However, due to its complexity it may have higher variance in estimates and not obtain statistical significance even if it is more appropriately modeling the uncertainty. Inverse probability weighting maintained good statistical power even if it exhibited a bit more bias than the commensurate prior method. The ease of implementation with standard software as well as being a standard technique known to

many researchers means that it remains a useful tool in historically controlled trials. As for propensity-score based power priors, the simulation study showed that the method exhibited a wider range of bias but had potential to be the most powerful. In the application in real data, the approach suffered most likely because the simple model and matching was unable to fit the trial data well enough to estimate the true treatment effect. Although it did not perform as well as hoped, it did not completely fail as a method. Instead, the results may be an indication that the approach should be expanded upon. Perhaps this approach could be altered to include covariates with non-informative priors. Then, the propensity score-based power prior would only apply to the intercept term. Therefore, differences in unmeasured confounding between the active and historical trials might be able to be controlled by discounting information about the model intercept while retaining the ability to adjust for covariates.

The adaptive trial design is an approach that likely would be acceptable to regulators. In the application with real trial data, we were unable to reduce the logistical burden by stopping early and would have had to continue to the final trial. The simulation showed that in certain situations, there is a high chance of early stoppage and by extension great potential for reducing the trial's size and duration. When contrasting with the analytic approaches, the interim analysis test is in some ways restrictive. It is an all or nothing approach that does not allow for a discounted version of the historical trial information to inform the analysis. In essence, the adaptive trial approach is a tradeoff. A researcher gains the possibility of reducing the trial at the expense of a level of conservatism that could lead to ignoring information that might still be useful in some form. Overall, it performed as intended in both the simulation and the real trial data. Other adaptive trial designs should be explored that may have even better performance.

This dissertation has generated thought on the greater philosophical approach when considering historically controlled trials. What kind of beliefs about the connection between historical data and active trial data are appropriate? The statistical theory behind Bayesian power priors[66] and hierarchical models measuring commensurability[94] is well developed and there are interesting interconnections between the approaches. Yet, questions remain regarding how to select the right method to implement in practice and do the assumptions of these models hold in real data? What are best practices for modeling selection, fitting, and interpretation for historically controlled

trials? As the literature continues to develop and new methods come to the surface, I hope that more practical application guides are put into place for the epidemiologists and biostatisticians to best utilize existing information and design better trials.

5.2 LIMITATIONS

The simulation study's goal was to compare methods in a simple, representative setting. However, its simplicity leaves room for additional questions such as whether if more covariates are used does the performance and comparison of the methods change. Also, if the differences between the historical and active trials are more extreme, then conclusions could favor one method over the other in a different manner. Fortunately, the implementation in real data gives at least one example of the method's performance in more complex settings. Additionally, we focused on changing the distribution of the covariates but did not consider a situation in which the association between the covariates and the outcome changes between the trials (the β_s in the outcome model). We also did not deeply explore the possible effect of unmeasured confounding between the trials. These additional questions were beyond the scope of this work but offer interesting avenues to explore in the future.

In the application in real data, the EPIC and OPTIMIZE trials are unusual in that they are extremely well suited for the research presented here. In practice, researchers may be faced with less similar studies that meet fewer of Pocock's criteria[65] and the conclusions drawn from this work may not be as applicable. OPTIMIZE also identified a large treatment effect, and these methods may not work as well in cases where there is a treatment effect of lower magnitude. However, well powered studies should empower the methods to perform well. Simulation studies in the planning stages of historically controlled trials are one way to help ensure appropriate performance in generating regulatory grade evidence.

5.3 FUTURE WORK

New methods continue to be published as possible ways to rigorously incorporate historical, external information. Future research will be required to explore and evaluate these new methods. Additionally, these methods and guidelines on how to use them have not become widespread, meaning researchers are not familiar with their potential. Hopefully, we can conduct real historically controlled trials that detect the presence of a treatment effect while also enjoying the benefit of smaller control arms and speedier trials to demonstrate the effectiveness and utility of these approaches.

As mentioned in **Chapter 1**, these approaches can also be applied in real world evidence trials that do not have as high quality data as seen in the real world trial application. It could be valuable to explore these types of situations where data is not as congenial between sources because such situations often arise in specific applications.

BIBLIOGRAPHY

1. FDA. Complex Issues in Developing Drugs and Biological Products for Rare Diseases and Accelerating the Development of Therapies for Pediatric Rare Diseases Including Strategic Plan: Accelerating the Development of Therapies for Pediatric Rare Diseases. FDA. 2014;(July).
2. Austin CP, Cutillo CM, Lau LPL, Jonker AH, Rath A, Julkowska D, Thomson D, Terry SF, de Montleau B, Ardigò D, Hivert V, Boycott KM, Baynam G, Kaufmann P, Taruscio D, Lochmüller H, Suematsu M, Incerti C, Draghia-Akli R, Norstedt I, Wang L, Dawkins HJS. Future of Rare Diseases Research 2017–2027: An IRDiRC Perspective [Internet]. Clinical and Translational Science. 2018 [cited 2022 Oct 29]. p. 21–27. Available from: www.cts-journal.com PMID: 28796445
3. Rees CA, Pica N, Monuteaux MC, Bourgeois FT. Noncompletion and nonpublication of trials studying rare diseases: A cross-sectional analysis. PLoS Med [Internet]. PLOS; 2019 [cited 2022 Oct 29];16(11):1–16. Available from: /pmc/articles/PMC6871779/ PMID: 31751330
4. Downing NS, Aminawung JA, Shah ND, Krumholz HM, Ross JS. Clinical trial evidence supporting FDA approval of novel therapeutic agents, 2005-2012. JAMA - J Am Med Assoc [Internet]. American Medical Association; 2014 Jan 22 [cited 2019 Jun 20];311(4):368–377. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2013.282034 PMID: 24449315
5. 21st Century Cures Act | FDA [Internet]. 2020 [cited 2020 Apr 6]. Available from: https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/21st-century-cures-act
6. Cder FDA. Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products Guidance for Industry Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products Guidance for Indust. (February 2023).
7. Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clin Pharmacol Ther. 2017;102(6):924–933. PMID:

28836267

8. U.S. Food and Drug Administration. Framework for FDA's Real-World Evidence Program [Internet]. U.S. Food and Drug Administration. 2018 [cited 2020 Jan 18]. Available from: www.fda.gov
9. Klonoff DC. The New FDA Real-World Evidence Program to Support Development of Drugs and Biologics. J Diabetes Sci Technol. 2020;14(2):345–349. PMID: 30862182
10. U.S. Department of Health and Administration F and DA. Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics Guidance for Industry. Off Commun Div Drug Inf Cent Drug Eval Res Food Drug Adm [Internet]. 2019 [cited 2020 Jan 18];(May 2019):1–8. Available from: https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm
11. Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the Use of Nonrandomized Real-World Data Analyses for Regulatory Decision Making. Clin Pharmacol Ther. Nature Publishing Group; 2019;105(4).
12. RCT DUPLICATE - Home [Internet]. [cited 2022 Oct 29]. Available from: https://www.rctduplicate.org/
13. Rockhold FW, Enas GG. 10 Years with ICH E10: Choice of Control Groups. Pharm Stat J Pharm Ind. Chichester, UK: John Wiley & Sons, Ltd; 2011;10(5):407–409.
14. Gibson RL, Burns JL, Ramsey BW. Pathophysiology and Management of Pulmonary Infections in Cystic Fibrosis. Am J Respir Crit Care Med. 2003;168(8):918–951. PMID: 14555458
15. Veit G, Avramescu RG, Chiang AN, Houck SA, Cai Z, Peters KW, Hong JS, Pollard HB, Guggino WB, Balch WE, Skach WR, Cutting GR, Frizzell RA, Sheppard DN, Cyr DM, Sorscher EJ, Brodsky JL, Lukacs GL. From CFTR biology toward combinatorial pharmacotherapy: Expanded classification of cystic fibrosis mutations. Mol Biol Cell [Internet]. 2016 [cited 2020 Jan 17];27(3):424–433. Available from: www.genet.sickkids.on.ca/home.html; PMID: 26823392
16. Rowe SM, Miller S, Sorscher EJ. Mechanisms of Disease: Cystic Fibrosis [Internet]. N Engl J Med. 2005. Available from: www.nejm.org
17. Treggiari MM, Retsch-Bogart G, Mayer-Hamblett N, Khan U, Kulich M, Kronmal R,

Williams J, Hiatt P, Gibson RL, Spencer T, Orenstein D, Chatfield BA, Froh DK, Burns JL, Rosenfeld M, Ramsey BW. Comparative efficacy and safety of 4 randomized regimens to treat early Pseudomonas aeruginosa infection in children with cystic fibrosis. Arch Pediatr Adolesc Med. 2011 Sep;165(9):847–856.

18. Mayer-Hamblett N, Retsch-Bogart G, Kloster M, Accurso F, Rosenfeld M, Albers G, Black P, Brown P, Cairns AM, Davis SD, Graff GR, Kerby GS, Orenstein D, Buckingham R, Ramsey BW. Azithromycin for early pseudomonas infection in cystic fibrosis the OPTIMIZE randomized trial. Am J Respir Crit Care Med [Internet]. 2018 [cited 2020 Jan 17];198(9):1177–1187. Available from: www.atsjournals.org. PMID: 29890086
19. FDA. KALYDECO FDA Label [Internet]. Available from: www.fda.gov/medwatch.
20. FDA. Orkambi FDA Label [Internet]. Available from: www.fda.gov/medwatch.
21. FDA. SYMDEKO™ (tezacaftor/ivacaftor) [Internet]. [cited 2020 Jan 18]. Available from: www.fda.gov/medwatch.
22. FDA. Trikafta FDA Label [Internet]. Available from: www.fda.gov/medwatch.
23. Walter E, Bass JL. The Effect of Elexacaftor/Tezacaftor/ Ivacaftor on Hospitalizations and Intravenous Antibiotic Use. Perm J. 2022;26(1):73–79. PMID: 35609157
24. Nichols DP, Paynter AC, Heltshe SL, Donaldson SH, Frederick CA, Freedman SD, Gelfond D, Hoffman LR, Kelly A, Narkewicz MR, Pittman JE, Ratjen F, Rosenfeld M, Sagel SD, Schwarzenberg SJ, Singh PK, Solomon GM, Stalvey MS, Clancy JP, Kirby S, Van Dalfsen JM, Kloster MH, Rowe SM. Clinical Effectiveness of Elexacaftor/Tezacaftor/Ivacaftor in People with Cystic Fibrosis A Clinical Trial. Am J Respir Crit Care Med. 2022;205(5):529–539. PMID: 34784492
25. VanDevanter DR, Mayer-Hamblett N, Boyle M. Feasibility of placebo-controlled trial designs for new CFTR modulator evaluation. J Cyst Fibros [Internet]. The Authors; 2017;16(4):496–498. Available from: http://dx.doi.org/10.1016/j.jcf.2017.02.012 PMID: 28284527
26. Mall MA, Mayer-Hamblett N, Rowe SM. Cystic fibrosis: Emergence of highly effective targeted therapeutics and potential clinical implications. Am J Respir Crit Care Med. 2020;201(10):1193–1208. PMID: 31860331
27. Simplify Study Indicates Potential to Reduce Medication Burden for People With CF Taking Trikafta | Cystic Fibrosis Foundation [Internet]. [cited 2023 Feb 9]. Available

from: https://www.cff.org/news/2022-11/simplify-study-indicates-potential-reduce-medication-burden-people-cf-taking-trikafta

28. Mayer-Hamblett N, Nichols DP, Odem-Davis K, Riekert KA, Sawicki GS, Donaldson SH, Ratjen F, Konstan MW, Simon N, Rosenbluth DB, Retsch-Bogart G, Clancy JP, Van Dalfsen JM, Buckingham R, Gifford AH. Evaluating the impact of stopping chronic therapies after modulator drug therapy in cystic fibrosis: The SIMPLIFY clinical trial study design. Ann Am Thorac Soc. 2021;18(8):1397–1405. PMID: 33465316
29. Rowe SM, Darrah R, Hastings M. Plenary 1: Hope for All: Addressing the Needs of Those With Untreated CF Mutations. North Am Cyst Fibros Conf. Philadelphia, PA: Cystic Fibrosis Foundation; 2022.
30. McGarry ME, McColley SA. Cystic fibrosis patients of minority race and ethnicity less likely eligible for CFTR modulators based on CFTR genotype. Pediatric Pulmonology. 2021. PMID: 33470563
31. Mayer-Hamblett N, van Koningsbruggen-Rietschel S, Nichols DP, VanDevanter DR, Davies JC, Lee T, Durmowicz AG, Ratjen F, Konstan MW, Pearson K, Bell SC, Clancy JP, Taylor-Cousar JL, De Boeck K, Donaldson SH, Downey DG, Flume PA, Drevinek P, Goss CH, Fajac I, Magaret AS, Quon BS, Singleton SM, VanDalfsen JM, Retsch-Bogart GZ. Building global development strategies for cf therapeutics during a transitional cftr modulator era. J Cyst Fibros [Internet]. Elsevier B.V.; 2020 Sep 1 [cited 2021 Mar 3];19(5):677–687. Available from: https://doi.org/10.1016/j.jcf.2020.05.011 PMID: 32522463
32. Ghadessi M, Tang R, Zhou J, Liu R, Wang C, Toyoizumi K, Mei C, Zhang L, Deng CQ, Beckman RA. A roadmap to using historical controls in clinical trials - By Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). Orphanet J Rare Dis. Orphanet Journal of Rare Diseases; 2020;15(1):1–19. PMID: 32164754
33. Magaret A, Warden M, Simon N, Heltshe S, Mayer-Hamblett N. Real-world evidence in cystic fibrosis modulator development: Establishing a path forward. Journal of Cystic Fibrosis. Elsevier B.V.; 2020. p. e11–e12. PMID: 32291159
34. De Boeck K, Lee T, Amaral M, Drevinek P, Elborn JS, Fajac I, Kerem E, Davies JC. Cystic fibrosis drug trial design in the era of CFTR modulators associated with substantial

clinical benefit: stakeholders' consensus view. J Cyst Fibros [Internet]. 2020 [cited 2021 Mar 30];19:688–695. Available from: https://doi.org/10.1016/j.jcf.2020.05.012

35. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med [Internet]. John Wiley and Sons Ltd; 2019 May 20 [cited 2021 Jan 27];38(11):2074–2102. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8086 PMID: 30652356
36. Friedman LM, Furberg C, DeMets DL, Reboussin D, Granger CB. Fundamentals of clinical trials . Fifth edit. Cham: Springer; 2015.
37. Aldrich JH, Nelson FD. Linear probability, logit, and probit models . Beverly Hills, California: Sage Publications; 1984.
38. McCullagh, P., & Nelder JA. Generalized Linear Models. 2nd ed. Chapman & Hall/CRC.; 1989.
39. Magaret AS, Warden M, Simon N, Heltshe S, Retsch-Bogart GZ, Ramsey BW, Mayer-Hamblett N. A new path for CF clinical trials through the use of historical controls. J Cyst Fibros [Internet]. 2022 [cited 2022 Nov 9];21(2):293–299. Available from: https://doi.org/10.1016/j.jcf.2021.11.007 PMID: 34879997
40. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2020. Available from: https://www.r-project.org/
41. Robinson D, Hayes A. broom: Convert Statistical Analysis Objects into Tidy Tibbles [Internet]. 2020. Available from: https://cran.r-project.org/package=broom
42. Wolodzko T. extraDistr: Additional Univariate and Multivariate Distributions [Internet]. 2020. Available from: https://cran.r-project.org/package=extraDistr
43. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org
44. Meredith M, Kruschke J. HDInterval: Highest (Posterior) Density Intervals [Internet]. 2020. Available from: https://cran.r-project.org/package=HDInterval
45. Grolemund G, Wickham H. Dates and Times Made Easy with {lubridate}. J Stat Softw [Internet]. 2011;40(3):1–25. Available from: http://www.jstatsoft.org/v40/i03/
46. Bache SM, Wickham H. magrittr: A Forward-Pipe Operator for R [Internet]. 2014. Available from: https://cran.r-project.org/package=magrittr
47. Venables WN, Ripley BD. Modern Applied Statistics with S [Internet]. Fourth. New

York: Springer; 2002. Available from: http://www.stats.ox.ac.uk/pub/MASS4/

48. Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T. {mvtnorm}: Multivariate Normal and t Distributions [Internet]. 2020. Available from: https://cran.r-project.org/package=mvtnorm
49. Genz A, Bretz F. Computation of Multivariate Normal and t Probabilities. Heidelberg: Springer-Verlag; 2009.
50. Hansen BB, Klopfer SO. Optimal full matching and related designs via network flows. J Comput Graph Stat. 2006;15(3):609–627.
51. Su Y-S, Masanao Yajima. R2jags: Using R to Run "JAGS" [Internet]. 2021. Available from: https://cran.r-project.org/package=R2jags
52. Hester J, Csárdi G, Wickham H, Chang W, Morgan M, Tenenbaum D. remotes: R Package Installation from Remote Repositories, Including "GitHub" [Internet]. 2021. Available from: https://cran.r-project.org/package=remotes
53. Hansen BB, Bowers J. Covariate balance in simple, stratified and clustered comparative studies. Stat Sci. 2008;23(2):219–236.
54. Bowers J, Fredrickson M, Hansen B. RItools: Randomization Inference Tools [Internet]. 2022. Available from: http://www.jakebowers.org/RItools.html
55. Plummer M. rjags: Bayesian Graphical Models using MCMC [Internet]. 2019. Available from: https://cran.r-project.org/package=rjags
56. Leydold J. rstream: Streams of Random Numbers [Internet]. 2020. Available from: https://cran.r-project.org/package=rstream
57. Zeileis A, Köll S, Graham N. Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in {R}. J Stat Softw. 2020;95(1):1–36.
58. Zeileis A. Econometric Computing with {HC} and {HAC} Covariance Matrix Estimators. J Stat Softw. 2004;11(10):1–17.
59. Zeileis A. Object-Oriented Computation of Sandwich Estimators. J Stat Softw. 2006;16(9):1–16.
60. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. Welcome to the {tidyverse}. J Open Source Softw. 2019;4(43):1686.

61. L'Ecuyer P. Good Parameters and Implementations for Combined Multiple Recursive Random Number Generators. Oper Res. Linthicum, MD: INFORMS; 1999;47(1):159–164.
62. L'Ecuyer P, Simard R, Chen EJ, Kelton WD. An Object-Oriented Random-Number Package with Many Long Streams and Substreams. Oper Res. Linthicum: INFORMS; 2002;50(6):1073–1075.
63. Matsumoto M, Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans Model Comput Simul. ACM; 1998;8(1):3–30.
64. Theussl S, Hornik K. Rglpk: R/GNU Linear Programming Kit Interface [Internet]. 2019. Available from: https://cran.r-project.org/package=Rglpk
65. Pocock SJ. The combination of randomized and historical controls in clinical trials. J Chronic Dis. Pergamon Press; 1976;29(3):175–188. PMID: 770493
66. Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: Theory and applications. Stat Med. John Wiley and Sons Ltd; 2015 Dec 10;34(28):3724–3749.
67. Morita S, Thall PF, Müller P, Thall F. Consultant ' s Forum : Determining the Effective Sample Size of a Parametric Prior Published by : International Biometric Society Linked references are available on JSTOR for this article : You may need to log in to JSTOR to access the linked references . 2015;64(2):595–602.
68. Baeten D, Baraliakos X, Braun J, Sieper J, Emery P, Van Der Heijde D, McInnes I, Van Laar JM, Landewé R, Wordsworth P, Wollenhaupt J, Kellner H, Paramarta J, Wei J, Brachat A, Bek S, Laurent D, Li Y, Wang YA, Bertolino AP, Gsteiger S, Wright AM, Hueber W. Anti-interleukin-17A monoclonal antibody secukinumab in treatment of ankylosing spondylitis: A randomised, double-blind, placebo-controlled trial. Lancet. 2013;382(9906):1705–1713. PMID: 24035250
69. Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter DJ. Summarizing historical information on controls in clinical trials. Clin Trials. 2010;7(1):5–18. PMID: 20156954
70. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics. Blackwell Publishing Inc.; 2014 Dec 1;70(4):1023–1032. PMID:

25355546

71. Abadie A, Diamond A, Hainmueller AJ. Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco control program. J Am Stat Assoc. 2010;105(490):493–505.
72. Hernán MA, Robins JM. Causal Inference: What If. 2020 [cited 2022 Oct 11]; Available from: https://www.hsph.harvard.edu/miguel-
73. Chen WC, Wang C, Li H, Lu N, Tiwari R, Xu Y, Yue LQ. Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data. J Biopharm Stat. 2020;30(3):508–520. PMID: 32370640
74. Cochran WG. The Planning of Observational Studies of Human Populations. J R Stat Soc Ser A Gen. London: Royal Statistical Society; 1965;128(2):234–266.
75. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Matched Sampl Causal Eff [Internet]. 2006 [cited 2020 Jan 18]. p. 170–184. Available from: https://academic.oup.com/biomet/article/70/1/41/240879
76. Rosenbaum PR. Design of observational studies . New York: Springer; 2010.
77. Propensity Score Analysis | Columbia Public Health [Internet]. [cited 2022 Oct 11]. Available from: https://www.publichealth.columbia.edu/research/population-health-methods/propensity-score-analysis
78. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med [Internet]. John Wiley & Sons; 1998 [cited 2022 Oct 11];17(19):2265–2281. Available from: https://onlinelibrary.wiley.com/terms-and-conditions PMID: 9802183
79. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res [Internet]. Taylor & Francis; 2011 May [cited 2022 Oct 10];46(3):399. Available from: /pmc/articles/PMC3144483/ PMID: 21818162
80. Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. J Am Stat Assoc. Alexandria, VA: Taylor & Francis Group; 1984;79(387):516–524.
81. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of

treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med [Internet]. John Wiley and Sons Ltd; 2015 Dec 10 [cited 2020 May 29];34(28):3661–3679. Available from: http://doi.wiley.com/10.1002/sim.6607 PMID: 26238958

82. Lefebvre G, Delaney JAC, Platt RW. Impact of mis-specification of the treatment model on estimates from a marginal structural model. Stat Med. Chichester, UK: John Wiley & Sons, Ltd; 2008;27(18):3629–3642.
83. Cribari-Neto F. Asymptotic inference under heteroskedasticity of unknown form. Comput Stat Data Anal. Elsevier; 2004 Mar 1;45(2):215–233.
84. Sheather SJ, Jones MC. A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. J R Stat Soc Ser B. 1991;53(3):683–690.
85. Christian B, Griffiths T. Algorithms to live by : the computer science of human decisions . First inte. Computer science of human decisions. New York: Henry Holt and Company; 2016.
86. Little RJ. Calibrated Bayes: A Bayes/frequentist roadmap. Am Stat. 2006;60(3):213–223.
87. Samaniego FJ. A Comparison of the Bayesian and Frequentist Approaches to Estimation. 1. Aufl. New York, NY: New York, NY: Springer Science + Business Media; 2010.
88. Salsburg D. The lady tasting tea : how statistics revolutionized science in the twentieth century. First Holt. New York: New York : Henry Holt; 2002.
89. Neath AA, Langenfeld N. A note on the comparison of the bayesian and frequentist approaches to estimation. Adv Decis Sci. 2012;2012.
90. Lin J, Gamalo-Siebers M, Tiwari R. Propensity-score-based priors for Bayesian augmented control design. Pharm Stat. 2019;18(2):223–238. PMID: 30537087
91. Klein JP, Moeschberger ML. Survival analysis : techniques for censored and truncated data . New York: Springer; 1997.
92. Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. Hierarchical Commensurate and Power Prior Models for Adaptive Incorporation of Historical Information in Clinical Trials. Biometrics. 2011;67(3):1047–1056. PMID: 21361892
93. Ibrahim JG, Chen MH. Power prior distributions for regression models. Stat Sci. 2000;15(1):46–60.
94. Hobbs BP, Sargent DJ, Carlin BP. Commensurate priors for incorporating historical

information in clinical trials using general and generalized linear models. Bayesian Anal [Internet]. Bayesian Anal; 2012 [cited 2020 Dec 7];7(3):639–674. Available from: https://pubmed.ncbi.nlm.nih.gov/24795786/

95. Berger JO, Berger JO. Statistical decision theory and Bayesian analysis . 2nd ed. New York: Springer-Verlag; 1985.
96. Mitchell TJ, Beauchamp JJ. Bayesian Variable Selection in Linear Regression. Source J Am Stat Assoc. 1988;83(404):1023–1032.
97. Robert CP, Casella G. Introducing Monte Carlo methods with R . New York ; Springer; 2010.
98. Psioda MA, Soukup M, Ibrahim JG. A practical Bayesian adaptive design incorporating data from historical controls. Stat Med [Internet]. John Wiley and Sons Ltd; 2018 Nov 30 [cited 2020 Dec 7];37(27):4054–4070. Available from: /pmc/articles/PMC6327964/?report=abstract PMID: 30033617
99. Ghosal S, Ghosh JK, Samanta T. On Convergence of Posterior Distributions. Ann Stat. HAYWARD: Institute of Mathematical Statistics; 1995;23(6):2145–2152.
100. Shi H, Yin G. Reconnecting p-Value and Posterior Probability Under One-and Two-Sided Tests. Am Stat [Internet]. General; 2020 [cited 2022 Sep 22];2021(3):265–275. Available from: https://www.tandfonline.com/action/journalInformation?journalCode=utas20
101. Treggiari MM, Rosenfeld M, Mayer-Hamblett N, Retsch-Bogart G, Gibson RL, Williams J, Emerson J, Kronmal RA, Ramsey BW. Early anti-pseudomonal acquisition in young patients with cystic fibrosis: Rationale and design of the EPIC clinical trial and observational study. Contemp Clin Trials [Internet]. Elsevier Inc.; 2009;30(3):256–268. Available from: http://dx.doi.org/10.1016/j.cct.2009.01.003 PMID: 19470318
102. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163(12):1149–1156. PMID: 16624967
103. RUBIN DB, SCHENKER N. MULTIPLE IMPUTATION IN HEALTH-CARE DATABASES - AN OVERVIEW AND SOME APPLICATIONS. Stat Med. W SUSSEX: Wiley; 1991;10(4):585–598.
104. Qu Y, Lipkovich I. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. Stat Med. Chichester, UK: John Wiley

& Sons, Ltd; 2009;28(9):1402–1414.

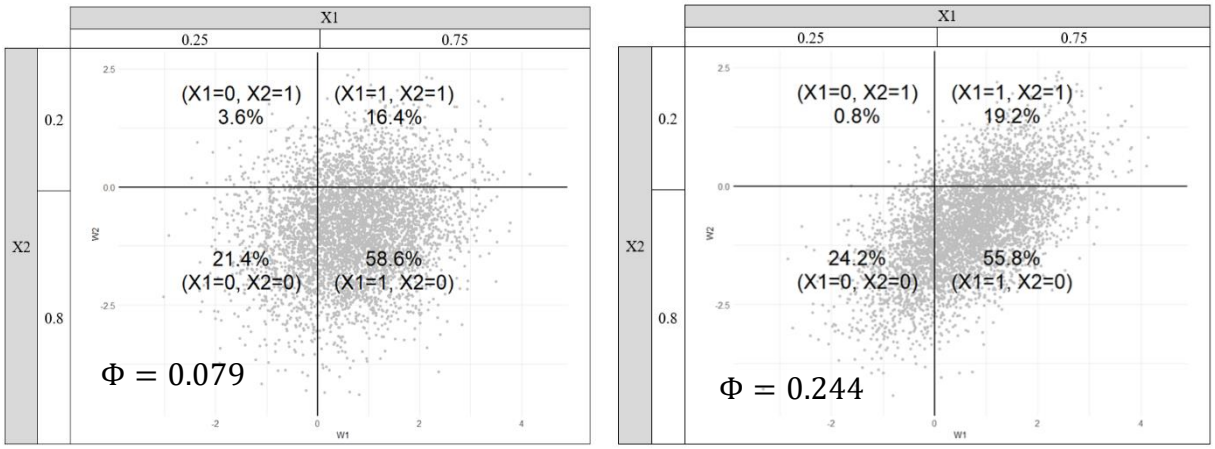
105. Tannenbaum SJ, Holford NHG, Lee H, Peck CC, Mould DR. Simulation of correlated continuous and categorical variables using a single multivariate distribution. J Pharmacokinet Pharmacodyn. 2006;33(6):773–794. PMID: 17053984

106. David FN, Cramer H. Mathematical Methods of Statistics. Biometrika. Princeton: Princeton University Press; 1947.

APPENDIX A: COVARIATE SIMULATION DETAILS

$$\boldsymbol{\delta = (0, 0)}$$

$\Sigma = 0.2$ $\Sigma = 0.6$



$$\boldsymbol{\delta = (-0.4, 0.2)}$$

$\Sigma = 0.2$ $\Sigma = 0.6$

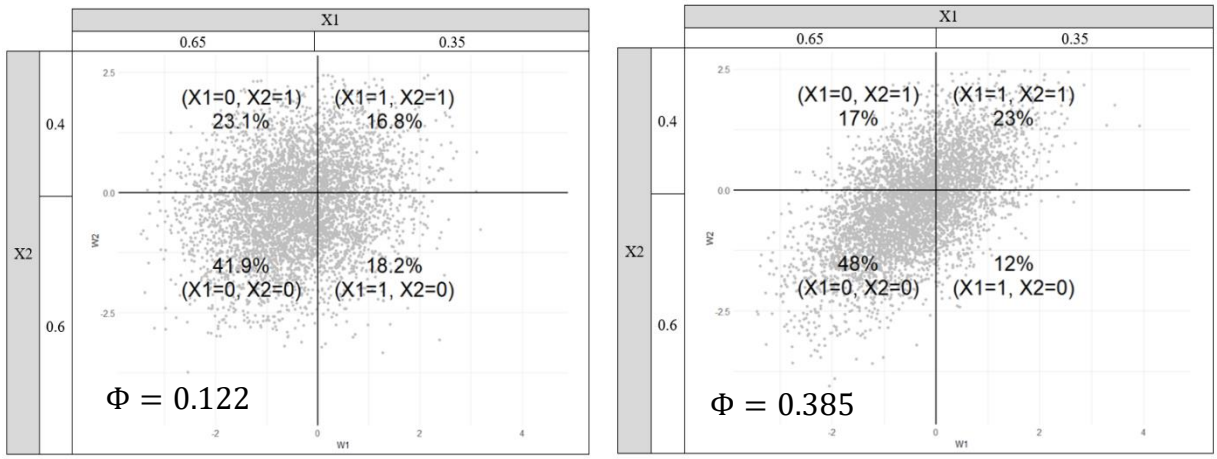


Figure 21. Covariate construction illustrated A) When X_1 and X_2 have prevalence of 0.75 and 0.2 respectively as seen in the marginals. The latent random variables (W_1, W_2) are shown as grey dots and have covariance and correlation equal to Σ. Each quadrant corresponds to the four possible pairs of the binary variables, X_1 and X_2, and their frequencies. In the lower left-hand corner of each plot, a ϕ coefficient[106] is calculated to measure the level of association between the binary variables. It is interpreted the same way as a Pearson correlation coefficient. ϕ ranges between -1 and 1 with values further from 0 indicating greater association. For $\boldsymbol{\delta = (0, 0)}$, this represents the active trial and the historical trial covariate distributions when there is no difference between the trials. For $\boldsymbol{\delta = (-0.4, 0.2)}$, this represents the historical trial covariate distribution when the prevalence differs between the trials.

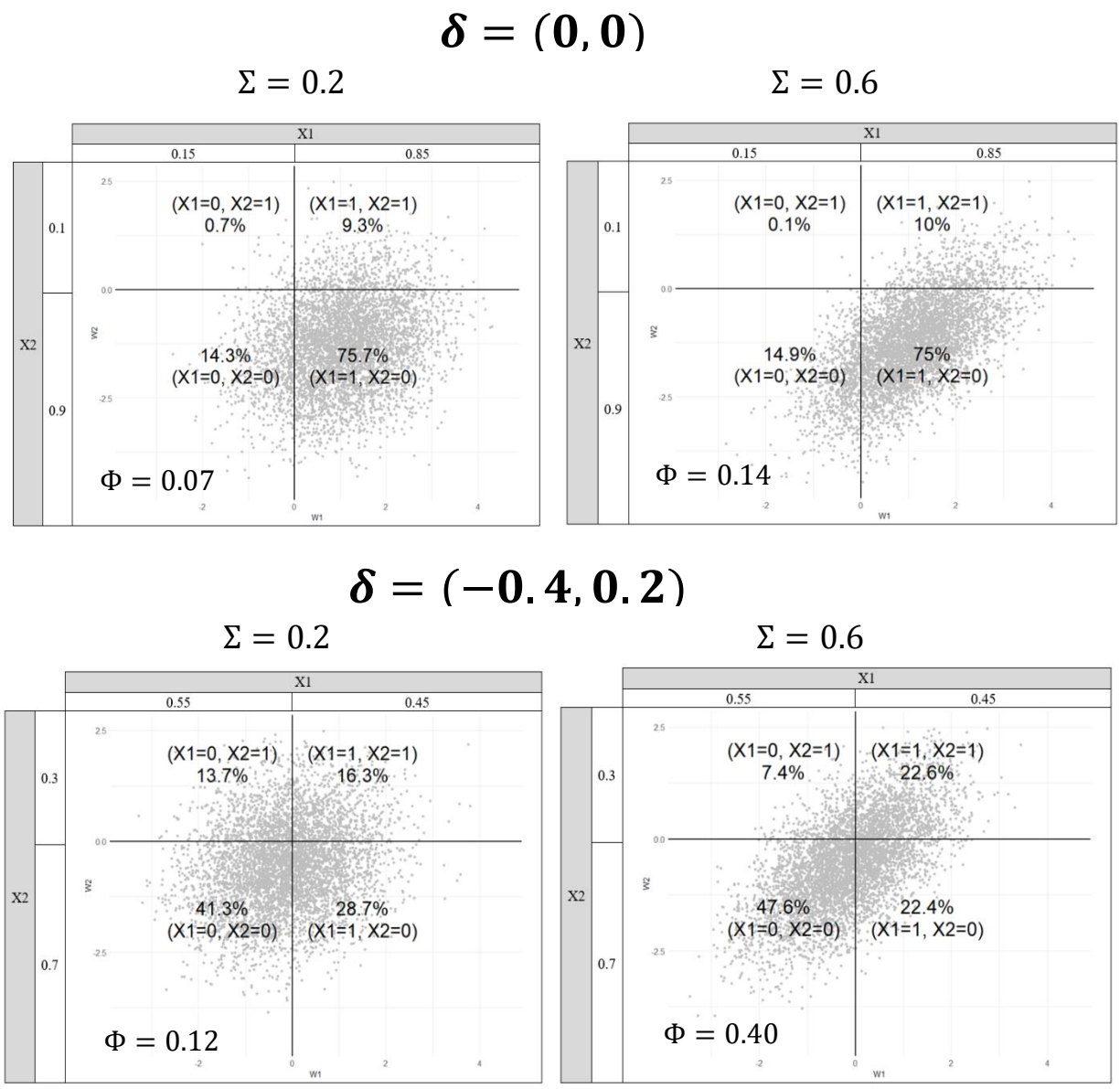


Figure 21. Covariate construction illustrated B) When $X_1$and X_2 have prevalence of 0.85 and 0.1 respectively as seen in the marginals. The latent random variables (W_1, W_2) are shown as grey dots and have covariance and correlation equal to Σ. Each quadrant corresponds to the four possible pairs of the binary variables, $X_1$and X_2, and their frequencies. In the lower left-hand corner of each plot, a ϕ coefficient[106] is calculated to measure the level of association between the binary variables. It is interpreted the same way as a Pearson correlation coefficient. ϕ ranges between -1 and 1 with values further from 0 indicating greater association. For $\delta = (0, 0)$, this represents the active trial and the historical trial covariate distributions when there is no difference between the trials. For $\delta = (-0.4, 0.2)$, this represents the historical trial covariate distribution when the prevalence differs between the trials.

For binary variables, the level of association between them is proportional to the difference between the frequency with which they agree, i.e. $(X_1, X_2) \in \{(1,1); (0,0)\}$ and the frequency with which they disagree, i.e. $(X_1, X_2) \in \{(1,0); (0,1)\}$. A measure of association often used is called the Φ coefficient discussed on pages 282-283 by Cramer[106], which is defined as follows for two binary variables:

		X1	
		0	1
X2	1	C	D
	0	A	B

$$\Phi = \frac{(A*D)-(B*C)}{\sqrt{(A+C)*(B+D)*(A+B)*(C+D)}}$$

It is interpreted in the same manner as a Pearson correlation in the case of two binary variables, with an absolute value of 0-0.3 as none to weak association, 0.3-0.7 as a moderate association.

However, a key concept regarding Φ is that the marginal distributions of $X_1$and X_2 restrict the level of agreement. For example, if $\Phi = 1$, then all variables agree and are contained in quadrant A and D, meaning the marginal distributions of $X_1$and X_2 must be identical. In other words, the marginal distributions are connected to the quadrants' frequencies and thus restrict values of ϕ.

For this data generating mechanism, the marginal probabilities are predetermined, meaning that there are limits placed on the values of Φ. Although the theoretical maximum value of Φ is calculatable, we can also approximate it quickly through simulation and report the maximum.

μ_{X_1}, μ_{X_2}	δ	$\mu_{X_1}, \mu_{X_2} + \delta$	~max (Φ)[a]
$(0.75, 0.2)$	$(0, 0)$	$(0.75, 0.2)$	0.31
$(0.75, 0.2)$	$(-0.4, 0.2)$	$(0.35, 0.4)$	0.92
$(0.85, 0.1)$	$(0, 0)$	$(0.85, 0.1)$	0.16
$(0.85, 0.1)$	$(-0.4, 0.2)$	$(0.45, 0.3)$	0.75

[a]max (Φ) is calculated from the conditions in column 3 across many possible correlation values for W_1, W_2. The maximum possible Φ value is approximated through simulation

Associations can be stronger when the marginal densities of the binary variables $X_1$and X_2 have marginal probabilities of taking the value of 1 are close to 50%. When $\mu_{X_1}, \mu_{X_2} = (0.75, 0.2)$ and

$\delta = (-0.4, 0.2)$, this implies historical means of $(0.35, 0.4)$ and thereby allows for the possibility of a stronger association. This is also true to a slightly lesser extent for $\mu_{X_1}, \mu_{X_2} = (0.85, 0.1)$ and $\delta = (-0.4, 0.2)$. The maximum possible value of ϕ is lowest for $\mu_{X_1}, \mu_{X_2} = (0.85, 0.1)$ and $\delta = (0, 0)$ because the marginal frequencies are very different.

In this simulation, when the $\delta = (0, 0)$, changing Σ does not dramatically affect the association of $X_1$and X_2. However, when $\delta = (-0.4, 0.2)$, higher values of Σ will change the association from none to moderate as seen in **Figure 21**. This fact is important to keep in mind for the interpretation of Σ's role in the simulation results. Additionally, it suggests that larger values of Σ should perhaps be considered in a post-hoc sensitivity analysis.

APPENDIX B: SIMULATION CODE AND ADDITIONAL DETAILS

All code used in the simulation, the detailed simulation results, and additional items are available on Github: https://github.com/marknwarden. For the figures presented in **Chapter 2,** the ordering used in the line plot portions of the figures was determined such that it was ordered based on a hierarchy {(3),(2),(4),(5), and (6)}:

- $\delta = (0,0); \delta = (-0.4, +0.2)$
- $(\mu_1, \mu_2) = (0.75,0.20); (\mu_1, \mu_2) = (0.85,0.10)$
- $Cov_{x_1,x_2} = 0.2; Cov_{x_1,x_2} = 0.6$
- $\beta = \log(HR) = log\begin{Bmatrix}1.10\\1.30\end{Bmatrix}; \beta = \log(HR) = log\begin{Bmatrix}1.20\\0.65\end{Bmatrix}$
- Number of historical controls is the same size as the expected treatment arm; It is double the size of the expected treatment arm

For example, in a row that varied (5) strength of association between covariates and outcome (β), the remaining scenarios would be ordered first in terms of δ, then (μ_1, μ_2), Cov_{x_1,x_2}, β, and number of historical controls, respectively. This leads to the following ordering for the *red* line with a treatment HR=0.5:

Varying (5) strength of association between covariates and outcome (β)				
Characteristic:	**(3)**	**(2)**	**(4)**	**(6)**
Id	$\boldsymbol{\delta}$	$\boldsymbol{(\mu_1, \mu_2)}$	$\boldsymbol{Cov_{x_1,x_2}}$	**Hx Size**
1	$(-0.4, +0.2)$	(0.75,0.20)	0.2	1
2	$(-0.4, +0.2)$	(0.75,0.20)	0.2	2
3	$(-0.4, +0.2)$	(0.75,0.20)	0.6	1
4	$(-0.4, +0.2)$	(0.75,0.20)	0.6	2
5	$(-0.4, +0.2)$	(0.85,0.10)	0.2	1
6	$(-0.4, +0.2)$	(0.85,0.10)	0.2	2
7	$(-0.4, +0.2)$	(0.85,0.10)	0.6	1
8	$(-0.4, +0.2)$	(0.85,0.10)	0.6	2
9	(0,0)	(0.75,0.20)	0.2	1
10	(0,0)	(0.75,0.20)	0.2	2
11	(0,0)	(0.75,0.20)	0.6	1
12	(0,0)	(0.75,0.20)	0.6	2
13	(0,0)	(0.85,0.10)	0.2	1
14	(0,0)	(0.85,0.10)	0.2	2
15	(0,0)	(0.85,0.10)	0.6	1
16	(0,0)	(0.85,0.10)	0.6	2
Periodicity	**8**	**4**	**2**	**1**

The line plots offer a visual aid in comparing the two values of the examined simulation characteristic being examined. The 32 scenarios (16 red and 16 blue) are ordered following the hierarchy so that the points along the blue and red lines correspond to simulations which are identical except for the red/blue characteristic. Using the example table given above, if we were comparing the **Red:** $\boldsymbol{\beta = \log(HR) = log\begin{Bmatrix}1.20\\0.65\end{Bmatrix}}$ and $\boldsymbol{\beta = \log(HR) = log\begin{Bmatrix}1.10\\1.30\end{Bmatrix}}$ lines, then the 8th point on both lines corresponds to scenarios with the following simulation parameters:

Id	$\boldsymbol{\delta}$	$(\boldsymbol{\mu_1, \mu_2})$	$\boldsymbol{Cov_{x_1,x_2}}$	**Hx Size**
8	$(-0.4, +0.2)$	(0.85,0.10)	0.6	2

This lets scenarios be compared directly on certain simulation characteristics. Additionally, each of the line plots has certain *periodicities* as shown in the table above. This allows for identification of patterns in the plots that can arise.

Tables for all performance methods for the methods are available on Github.

VITA

Mark Niklas Warden is the son of Dennis and Kendall Warden and the brother of Nathan Warden. He was raised in Durham, North Carolina, and graduated from Durham School of the Arts (DSA) in 2007. He earned a Bachelor of Arts in mathematics from the University of North Carolina at Chapel Hill in 2011. He went on to earn a master's degree in teaching from Duke University in 2012 and taught middle school and high school mathematics at both Northern high school and DSA from 2011-2015. He received a North Carolina teaching fellowship and a Durham teaching fellowship. He then returned to school and earned a master's degree in biostatistics from Washington University in St. Louis, Missouri, in 2016. He worked as a biostatistician I in Washington University's medical school's department of neurology from 2016-2018 under Dr. Brad Racette and Dr. Susan Searles Nielsen, an epidemiology PhD from the University of Washington. Under Susan's guidance, he decided to pursue a Doctor of Philosophy in epidemiology from the University of Washington in Seattle (2018-2023).