BAYESIAN METHODS FOR RESPONSE-ADAPTIVE RANDOMIZATION AND DRUG REPURPOSING

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

JENNIFER L. PROPER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

ADVISED BY DR. THOMAS A. MURRAY

December, 2022

Acknowledgements

The completion of this dissertation would not have been possible without the support of my friends, family, and colleagues. Firstly, I would like to express my sincere gratitude to my research advisor, Dr. Thomas Murray. Thank you for your guidance, mentorship, and encouragement over the past four years and for developing me into the independent researcher that I am today. When I started this program, I had zero research experience and so much to learn. Now, I feel confident and ready to begin my career as a biostatistician and for that I am truly grateful. Next, I would like to thank Ashley Peterson for going above and beyond as my academic advisor. Getting your Ph.D. is far from easy, and I would not be here today without our monthly check-ins and of course your Trader Joe's peanut butter cups. I would also like to thank Dr. James Neaton and the National Heart, Lung, and Blood Institute (T32HL129956) for funding my work during my first three years. Working on clinical trials is what inspired me to go to graduate school, and I am so thankful to have had the opportunity to receive additional training in clinical trial methodology. I would also like to thank Dr. Haitao Chu, Dr. Erika Helgeson, and Dr. Michael Puskarich for serving on my dissertation committee; thank you for your insightful feedback and for encouraging me to think critically by asking stimulating questions. Thank you as well to Dr. Michael Sonksen and Dr. Purvi Prajapati for giving me the opportunity to intern at Eli Lilly and Company last summer. This experience solidified my passion to pursue a career in the pharmaceutical industry and also served as the inspiration for Chapter 4 of this dissertation. Finally, thank you to my incredibly supportive and loving family. Your unrelenting belief in me has meant more to me than you will ever know.

Dedication

I dedicate this dissertation to my parents, Beth and Gil Proper, whose endless support of my personal and professional goals has always allowed me to pursue my dreams. Thank you for your unconditional love and teaching me to never give up, no matter how difficult the journey ahead may seem.

I also dedicate this dissertation to Sharon Ling, a fellow Ph.D. student and cherished friend, whose life was tragically cut short last year. Thank you for believing in me even when I didn't believe in myself. Your infectious spirit and profound impact on my life will never be forgotten.

Abstract

Phase II clinical trials are an expensive and risky component of the drug development process used to study the safety and efficacy of new treatments. Recent studies suggest that the average cost of a phase II trial ranges from 7.0 to 19.6 million dollars, depending on therapeutic area, and that about one-third of all phase II trials fail. Failures in phase II generally occur when a new toxic side effect arises or the observed treatment effect is smaller than anticipated. To minimize harm to trial participants, drug developers require statistical methods that can be used to reduce patient exposure to ineffective or harmful treatments. This thesis focuses on innovations in early phase development aimed at reducing exposure to ineffective treatments.

Chapters 2 and 3 concern response-adaptive randomization (RAR), which alters the allocation ratio based on accruing data in favor of the empirically superior treatment. In contrast to fixed 1:1 allocation, RAR gives participants a greater chance of receiving the treatment during the trial, which tends to reduce the number of participants assigned to the inferior arm. Yet, existing RAR approaches are commonly criticized for reducing power relative to 1:1 allocation, inflating type I error rate when a time-trend is present, and engendering nontrivial probabilities of allocating more participants to the inferior arm. To temper these problematic behaviors, Chapter 2 proposes a new probability model and randomization strategy for implementing Bayesian RAR in a binary outcomes setting. Simulation studies show that the proposed methods engender smaller average sample sizes with similar power, better control over type I error rate, and a negligible chance of a sample size imbalance in the wrong direction compared to the traditional design. Chapter 3 proposes a new metric for comparing group sequential designs that measures the expected number of failures in the fixed group of individuals who are directly impacted by the design choice. In contrast to within-trial metrics, this approach considers designs with equal type I and

II error rates and assesses their impact with respect to relevant, equal-sized populations. Simulation studies show that various implementations of group sequential Bayesian RAR offer modest improvements with respect to the proposed metric relative to conventional group sequential monitoring alone.

Chapter 4 concerns drug repurposing, which is the process of discovering new therapeutic uses for existing treatments. Drug repurposing involves studying treatments with well-established safety profiles, which can dramatically shorten the drug development timeline and reduce the occurrence of toxic side effects in patients. However, existing approaches for drug repurposing involve complex, computationally-intensive analytical methods that are not widely used in practice. This chapter proposes a novel Bayesian network meta-analysis (NMA) framework that can predict the efficacy of an approved treatment in a new indication and thereby identify candidate treatments for repurposing. We obtain predictions using two main steps: first, we use standard NMA modeling to estimate average relative effects from a network comprised of treatments studied in both indications in addition to one treatment studied in only one indication. Then, we model the correlation between relative effects using various strategies that differ in how they model treatments across indications and within the same drug class. Simulation studies find that the model minimizing root mean squared error of the posterior median for the candidate treatment depends on the amount of available data, the level of correlation between indications, and whether treatment effects differ, on average, by drug class.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Phase II clinical trials are an expensive and risky component of the drug development process used to study the safety and efficacy of new treatments. Recent studies suggest that the average cost of a phase II trial ranges from 7.0 to 19.6 million dollars, depending on therapeutic area, and that about one-third of all phase II trials fail. [1] Failures in phase II generally occur when a new toxic side effect arises or the observed treatment effect is smaller than anticipated. To minimize harm to trial participants, drug developers require statistical methods that can be used to reduce patient exposure to ineffective treatments. One such method that has garnered considerable attention in phase II clinical trials is response-adaptive randomization (RAR), which alters the allocation ratio based on accruing data in favor of the empirically superior treatment. [2–4] In contrast to 1:1 allocation, RAR gives participants a greater chance of receiving the treatment during the trial, which tends to reduce the number of participants assigned to the inferior treatment. Yet, RAR remains contentious and has seen limited use in practice. It is commonly criticized for reducing power relative to 1:1 allocation, inflating type I error rate when a time-trend is present, and engendering nontrivial probabilities of allocating more participants to the inferior arm. [3, 5–10]

Drug repurposing, the process of discovering new therapeutic uses for existing treatments, has also received interest for its potential to minimize in-trial patient harm. This process occurs when drug developers hypothesize that an additional clinical indication will respond similarly to a licensed drug due to biological similarities or comparable mechanisms of action, and can advantageously be used to satisfy unmet clinical needs or maximize a treatment's therapeutic ability. It also involves studying treatments with well-established safety profiles, which can dramatically shorten the drug development timeline and reduce the occurrence of toxic side effects in patients. [11] However, existing approaches for drug repurposing involve complex, computationally-intensive analytical methods and are not widely used in practice. Instead, repurposing decisions are often based on subjective judgements and limited empirical evidence, which increases the likelihood of conducting a futile trial, wastes valuable resources, and exposes many individuals to an ineffective treatment. [12] This thesis focuses on Bayesian methods for response-adaptive randomization and drug repurposing in an effort to reduce patient exposure to ineffective treatments during early phase drug development.

The second chapter of this thesis proposes an alternative probability model and randomization strategy for implementing Bayesian RAR in a binary outcomes setting. This work was motivated by a recently conducted phase II trial called the Advanced R^2Eperfusion STrategies for Refractory Cardiac Arrest (ARREST) trial (ClinicalTrials.gov, NCT03880565), which assessed the impact of extracorporeal membrane oxygenation (ECMO) facilitated resuscitation versus advanced cardiac life support on survival to hospital discharge in adults who experienced an out-of-hospital cardiac arrest and refractory ventricular fibrillation. [13, 14] Bayesian RAR was performed to minimize patient exposure to the inferior treatment, as the consequences of providing an inferior treatment were grave. To implement RAR, investigators used independent beta-binomial probability models for the treatment arm response rates in conjunction with complete randomization, which is the only way in which RAR has been traditionally applied in two-arm trials with a binary outcome. [13]

Although simple to implement, this conventional design may engender substantial type I error rate inflation when the underlying response rate deviates from its hypothesized value. It may also result in arbitrarily large deviations from the target allocation in favor of the inferior treatment. To temper these problematic behaviors, Chapter 2 proposes the use of a logistic regression probability model coupled with one of two alternative randomization methods that limit deviations from the targeted allocation ratio. The relative merits of the proposed Bayesian RAR design are assessed using simulations.

Chapter 3 proposes a new metric for comparing group sequential designs based on the cohort most acutely impacted by the choice of the design. Group sequential designs are often used to facilitate early stopping for efficacy, harm, or futility at pre-specified interim sample sizes throughout the trial, which protects participants from unnecessary exposure to ineffective or harmful treatments and allows limited resources to be allocated to other trials. [15] Response-adaptive randomization may be considered when investigators wish to maximize in-trial patient benefit, or equivalently minimize in-trial patient harm. To inform this decision, investigators have traditionally compared Bayesian RAR and classical frequentist group sequential trials using within-trial metrics such as the percentage or number of patients assigned to the inferior arm or the expected number of trial failures. [16–18] However, these evaluations are often limited by failures to hold type I and II error rates constant or to account for the different sample size requirements of the designs under consideration, which arguably leads to unfair comparisons. The metric introduced in this chapter improves on existing patient benefit metrics by considering a set of feasible group sequential designs with equal type I and II error rates and measuring the expected number of failures in the fixed group of individuals who are directly impacted by the design choice. Namely, those who would participate in the trial if enrollment were open when they become eligible. This chapter concludes by illustrating how this metric may be applied to select a design in the motivating ARREST trial context.

The fourth chapter of this thesis develops a novel Bayesian network meta-analysis framework that can assist drug developers in answering the commercial decision of "what to study next" in the context of drug repurposing. Network meta-analysis (NMA) is a statistical method that combines data from multiple studies assessing multiple treatments to reduce bias and improve efficiency. [19–21] Although traditionally applied to data from late-stage studies for comparative effectiveness research or regulatory submissions, NMA is useful in different stages of drug development. Mawdsley et al. [22] developed model-based NMA (MBNMA) by incorporating dose-response models into an NMA framework. Their method uses plausible physiological dose-response models to predict the effect of multiple treatments across a range of doses. Pedder et al. [23] expanded MBNMA to model the time-course relationship of multiple treatments using a continuous function. This method can similarly predict treatment efficacy at unstudied time points and enhance the understanding of pharmacodynamic profiles of new compounds. Importantly, predictions from both methods can be used to inform the design of future clinical trials, evaluate the competitive landscape, and accelerate compound development by mitigating trial failures. [24] Despite these applications, NMA has not been used to predict the efficacy of an approved treatment in a new indication and thereby identify repurposable treatments. This chapter develops an NMA framework that can predict whether a treatment is likely to succeed in a new indication based on data from other treatments studied in both the new and approved indications, and concludes by discussing an illustrative example in psoriasis and psoriatic arthritis.

Chapter 5 summarizes the findings in previous chapters and describes future work related to these topics.

Chapter 2

Alternative Models and Randomization Techniques for Bayesian Response-Adaptive Randomization with Binary Outcomes

2.1 Introduction

Fixed, equal randomization is a well-established approach to obtain unbiased treatment effect estimates and high power in clinical trials. However, response-adaptive designs continue to receive interest for phase II clinical trials due in part to their ability to introduce randomization while considering ethical constraints, contrary to single-arm studies. While fixed randomization allocates patients among treatments with fixed probabilities throughout the trial, response-adaptive randomization commonly alters the allocation ratio based on interim analyses, favoring treatments with higher empirical success rates, which tends to reduce the number of patients assigned to an inferior treatment. [25]

The proposal to use response-adaptive randomization in binary outcome trials dates back to at least Thompson [26] and has recently been discussed extensively by Berry et al. [2] In this paper, we consider Thompson sampling methods where participants are randomized to the treatment arm in proportion to the posterior probability that the treatment is superior to the control, based on the available data at the time of randomization. [26] The Advanced R^2Eperfusion STrategies for Refractory Cardiac Arrest (ARREST) trial, a National Heart, Lung, and Blood Institute funded, phase II clinical trial, used a Bayesian response-adaptive design and is the motivation for this research. [13,14] The ARREST trial aimed to evaluate the efficacy of extracorporeal membrane oxygenation facilitated resuscitation versus standard advanced cardiac life support resuscitation on adults who experienced a refractory ventricular fibrillation or pulseless ventricular tachycardia out-of-hospital cardiac arrest. For logistical reasons, adaptations to the randomization schedule were designed to occur in a group sequential manner. The primary outcome of the ARREST trial was survival to hospital discharge. The principal investigators wanted to use response-adaptive randomization to minimize patient exposure to the inferior treatment, as the consequences of providing an inferior treatment were grave. [13]

Two-arm Bayesian response-adaptive trials with a binary outcome, such as the ARREST trial, have been conventionally implemented using independent beta-binomial probability models for the treatment arm response rates in conjunction with complete randomization, where every subject is independently randomized among the arms with probabilities that reflect the current target allocation (i.e. coin flipping). While evaluating designs for the ARREST trial with this conventional implementation, we noticed sensitivity in the type I error rate, defined as the probability of claiming either efficacy or harm when the response rates are equal, to the underlying response probability and prior specification. Because the control response rate is uncertain when designing any trial, the actual type I error rate of this design may differ considerably from what was intended.

Type I error inflation for response-adaptive designs may also be exacerbated in the presence of time trends including patient drift, a well-documented phenomenon characterized by evolving patient prognoses. [6] To protect against time trends and maintain type I error at the nominal level, Simon and Simon have developed a general class of randomization tests and Villar et al. recommend the use of several correction methods. [9, 10] However, a tailored solution may result in better trade-offs when one is particularly worried about incorrectly hypothesized response rates in response-adaptive designs. Other limitations also explain why these designs remain widely unused in clinical practice. [3] Hey and Kimmelman, for instance, argue that Bayesian response-adaptive randomization is undesirable in the two-armed setting because it notably increases trial sample size, [8] while Thall et al. relatedly contend that it reduces power, defined as the probability of claiming efficacy when the response rate is higher in the treatment arm, and leads to a nontrivial probability of a subject imbalance toward the inferior treatment. [5]

We hypothesize that an alternative probability model and randomization method will temper some of the problematic behaviors engendered by the conventional two-arm Bayesian response-adaptive design with binary outcomes. Following the work of Ghosh and Gelman for prior specification in logistic regression modeling, [27, 28] we surmise that using logistic models with t-distribution priors might engender a design with reduced prior sensitivity and improved operating characteristics. We further postulate that urn and permuted block randomization methods, which to our knowledge have yet to be studied in this context, may reduce the risk of assigning more patients to the inferior arm. Contrary to complete randomization, these methods restrict deviations from the target allocation throughout the trial.

In this paper, we report a simulation study evaluating the operating characteristics of Bayesian group sequential response-adaptive designs using Thompson sampling methods based on various probability models coupled with various randomization methods, including an existing mass-weighted urn strategy and a new weighted permuted block strategy. Below,

we outline each model and randomization scheme, report the results of our simulation study, and recommend using a probability model and randomization technique that ameliorate previously noted limitations of the conventional Bayesian response-adaptive design for the motivating ARREST trial context.

2.2 Methods

2.2.1 Probability Models

Continuing in the context of the ARREST trial, let π_z reflect the true, but unknown, probability of survival to hospital discharge in arm z = E or C, where E denotes extracorporeal membrane oxygenation (treatment) and C denotes standard advanced cardiac life support (control). The posterior probability that the treatment is superior to the control is then given by $P_{E>C} = \Pr(\pi_E > \pi_C|\boldsymbol{y})$, where $\boldsymbol{y}$ is the currently available outcome data, i.e. the numbers assigned to each arm and who survived to hospital discharge. [2] We consider two probability models for calculating $P_{E>C}$ below.

Independent Beta-Binomial Model

The independent beta-binomial model is often the basis for two-arm Bayesian response-adaptive trials with binary outcomes. Assuming n_z subjects are assigned to treatment z and Y_z subjects survive, the independent beta-binomial model assumes:

$$\begin{aligned} Y_z|\pi_z &\overset{ind}{\sim} \text{Binomial}\left(n_z,\ \pi_z\right) \text{ and} \\ \pi_z &\overset{ind}{\sim} \text{Beta}\left(\pi^* n_0,\ (1-\pi^*)n_0\right),\ \text{z} = \text{E}, \text{C} \end{aligned} \tag{2.1}$$

such that the posterior distribution arises as:

$$\pi_z | Y_z \overset{ind}{\sim} \text{Beta}\left(\pi^* n_0 + Y_z,\ (1-\pi^*)n_0 + n_z - Y_z\right) \tag{2.2}$$

where π^* is the prior mean and n_0 is the prior effective sample size presented by Morita et al. [29] Additional details for this model are provided in the Appendix. The non-informative, uniform Beta(1,1) prior arises with $n_0 = 2$ and $\pi^* = 0.50$. Because the hypothesized survival probability for the control was 12%, we also consider using a weakly informative Beta(0.24,1.76) prior with $n_0 = 2$ and $\pi^* = 0.12$.

Logistic Regression Model

We next consider a logistic regression model resulting from a logit transformation of π_z:

$$\text{logit}(\pi_z) = \log\left(\frac{\pi_z}{1-\pi_z}\right) = \beta_0 + \beta_1 \cdot (\text{I}\{\text{z} = \text{E}\} - 0.5) \tag{2.3}$$

Gelman et al. first recommended using Cauchy prior distributions for logistic regression coefficients. [28] However, because the posterior mean of the regression coefficients may not exist when assuming Cauchy priors in the presence of complete or quasicomplete separation, Ghosh et al. later recommended using a lighter tailed t-distribution with between 5 to 7 degrees of freedom. [27] Let $t_\nu(\mu, \sigma)$ denote a generalized t-distribution with ν, μ, and σ respectively signifying the degrees of freedom, location, and scale parameters. The density function for this distribution is provided in Section 1 of Appendix A. We consider specifying independent $t_7(0, 2.5)$ priors for β_0 and β_1, as well as a $t_7(\log(\frac{0.12}{0.88}), 2.5)$ prior distribution for β_0. The latter location hyperparameter value arises from solving for β_0 when $\pi_z = 0.12$ and $\beta_1 = 0$ in equation (3.3), and thus reflects the hypothesized survival probability under the null hypothesis of no treatment effect. The logistic regression model is formally defined as follows:

$$Y_z|\beta_0,\beta_1 \sim \text{Binomial}\left(n_z, \frac{\exp(\beta_0+\beta_1\cdot(\text{I}\{\text{Z}=\text{E}\}-0.5))}{1+\exp(\beta_0+\beta_1\cdot(\text{I}\{\text{Z}=\text{E}\}-0.5))}\right) \tag{2.4}$$
$$\text{and } \beta_0,\beta_1 \overset{ind}{\sim} t_7(0,2.5),\ \text{Z}=\text{E},\text{C}$$

The resulting posterior distribution for the model in equation (3.4) is not recognizable, which makes computation of $P_{E>C}$ slightly more challenging. For posterior calculations, we use a conditionally conjugate Polya-Gamma Gibbs sampler analogous to Ghosh et al. [27] Additional details are provided in the Appendix.

2.2.2 Randomization Methods

In a Bayesian group sequential response-adaptive design, the target allocation for each group is altered based on interim analyses in favor of the treatment that is currently performing better. Let p_m be the target randomization probability to the treatment arm for group $m = 1, \ldots, M$ of size b. During its conception, the ARREST trial considered group sizes of b = 15, 30 or 50. Due to the high variability of $P_{E>C}$ with small sample sizes, using Thompson sampling by setting $p_m = P_{E>C}$ can reduce power and engender a considerable risk of allocating more subjects to the interior treatment. [4] Following Thall and Wathen, [25] we apply the following modifications in an effort to stabilize the randomization probability: the first group uses equal randomization, i.e. $p_1 = 0.5$, and for subsequent groups p_m is restricted between 0.25 and 0.75 (e.g., if $P_{E>C} = 0.82$ then $p_m = 0.75$ for the next group). The latter may preclude power loss due to extreme allocation to one arm and limit selection bias that may arise from clinicians determining the randomization scheme.

One complication with response-adaptive randomization is that it may not be possible to achieve the target allocation within each group. For example, when group size $b = 15$ and $p_m = 0.5$, it not possible to allocate $bp_m = 7.5$ participants to each arm; however, it is possible to use a randomization method that allocates 7.5 participants to each arm

in expectation. We consider three randomization methods that achieve this property: the weighted coin design, the mass-weighted urn design, and a new modified permuted block design. Other methods would have also been suitable for consideration, including the block urn design, [30] drop-the-loser urn design, [31] and wide brick tunnel randomization, [32] among others. [33] To our knowledge, the weighted coin design is the only randomization method studied for Bayesian response-adaptive trials. While the mass-weighted urn design may be applied in Bayesian response-adaptive randomization trials, its impact on the operating characteristics remains unclear. We evaluate and compare these randomization methods in terms of power, type I error, allocation accuracy, and the risk of a treatment imbalance in the wrong direction under the target alternative scenario.

Weighted Coin Design

Let p_{mi} denote the probability that the i^{th} subject in the m^{th} group is assigned to the treatment given the prior data. The weighted coin method uses $p_{mi} = p_m$ for each $i = 1, \ldots, b$, and thus may heuristically be described as repeated flipping of a weighted coin. An important aspect of any randomization method is the trade-off between allocation randomness and treatment balance; although the weighted coin method maximizes randomness, its lack of imbalance control may result in arbitrarily large deviations from the target allocation. For example, in the first group with $b = 30$ and $p_1 = 0.5$, the weighted coin approach may assign 16 and 14, 17 and 13, or even 18 and 12 subjects to the treatment and control respectively, despite a targeted allocation of 15 to each arm. There is also a nontrivial probability (0.20) of observing allocations as or more extreme than 11 and 19, or 19 and 11.

Mass Weighted Urn Design

Zhao recently developed a randomization method for unequal allocations called the mass weighted urn design. [34] This method uses an urn randomization scheme with one ball per arm. The probability of selecting a ball is proportional to its mass, which changes after

every draw so that any allocation ratio may be targeted exactly. Let $n^m_{i-1,z}$ denote the number of participants allocated to treatment z among the initial $(i-1)$ participants in group m. The expression for p_{mi} arises as:

$$p_{mi} = \frac{\max[\alpha p_m - n^m_{i-1,E} + (i-1)p_m, 0]}{\max[\alpha p_m - n^m_{i-1,E} + (i-1)p_m, 0] + \max[\alpha(1-p_m) - n^m_{i-1,C} + (i-1)(1-p_m), 0]} \tag{2.5}$$

The initial mass of the balls corresponding to the treatment and control are αp_m and $\alpha(1-p_m)$, respectively, where α is the total mass of the balls in the urn. [34] If $n^m_{i-1,E}$ exceeds the desired allocation of $(i-1)p_m$ by more than αp_m, the numerator of equation (3.12) becomes 0 and the next subject is assigned to the control. Similarly, if $n^m_{i-1,C}$ exceeds the desired allocation of $(i-1)(1-p_m)$ by more than $\alpha(1-p_m)$, the second term in the denominator of equation (3.12) becomes 0 and the next subject is assigned to the treatment. In this way, α controls how far the realized allocation can deviate from the targeted allocation. For example, consider the first group of size $b = 30$ with $\alpha = 3$. The mass-weighted urn design ensures allocations of 14:16, 15:15, or 16:14 to the treatment and control, respectively. While the maximum tolerated imbalance property engendered by the α parameter is desirable, it limits allocation randomness and thereby may be susceptible to selection bias. [33] We set $\alpha = 3$ as this strikes a reasonable trade-off between randomness and balance.

Modified Permuted Block Design

The permuted block design is a common unequal allocation method used in clinical practice that provides an effective imbalance control when block sizes are small. However, it cannot achieve every target allocation that may arise in a response-adaptive trial. While the mass-weighted urn design has shown promise in improving the precision of unequal allocation, it sacrifices some control over deviation from the target allocation for greater unpredictability. We propose a weighted permuted block strategy, herein referred to as the modified permuted block design, that precisely targets general allocation ratios. For example, with $b = 30$ and $p_1 = 0.5$, this strategy guarantees 15:15 allocation.

The targeted allocation within each block cannot be achieved when bp_m is not an integer. However, there often exist two allocations that most closely achieve the target allocation, $\lceil bp_m \rceil$ and $\lfloor bp_m \rfloor$, where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ denote the ceiling and floor functions, respectively. Letting E_m denote the number of subjects to be assigned to the treatment in block m, E_m will be sampled as follows:

$$E_m|U_m = \begin{cases} \lceil bp_m \rceil, & \text{if } U_m = 1 \\ \lfloor bp_m \rfloor, & \text{if } U_m = 0 \end{cases}$$
$$\text{where } U_m = \begin{cases} 1, & \text{with probability } bp_m - \lfloor bp_m \rfloor \\ 0, & \text{with probability } \lceil bp_m \rceil - bp_m \end{cases} \tag{2.6}$$

We define $\lfloor bp_m \rfloor = bp_m - 1$ when bp_m is an integer. If $bp_m = 7.5$, where $b = 15$ and $p_m = 0.5$, this design will assign either 7 or 8 participants to the treatment with equal probability. Suppose instead $p_m = 0.68$ such that $bp_m = 10.2$; this design will assign 11 patients to the treatment with probability 0.2 and 10 with probability 0.8. Letting a_i be the block assignment for subject i and p_u be the probability that $U_m = 1$, in Section 2 of Appendix A we show that the expected value of $E_m = bp_m$ when $p_u = bp_m - \lfloor bp_m \rfloor$, suggesting that the modified permuted block design targets the group allocation ratio exactly. We also show that each subject is assigned to the experimental treatment with marginal probability p_m, given their block assignment. This design achieves the allocation ratio preserving property described by Kuznetsova and Tymofyeyev, which advantageously protects against evaluation, selection, and accidental biases that may arise even in double-blind trials. [35] The mass-weighted urn design is non-allocation ratio preserving, especially with $\alpha < 6$. [33]

2.2.3 Simulation Study Design

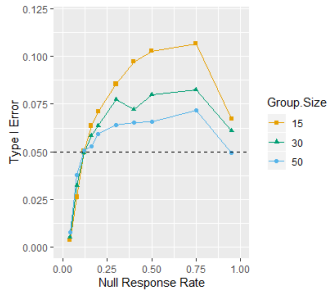
We used computer simulation to evaluate the relative merits of the probability models and randomization methods for the ARREST trial. We considered designs with up to

$n = 150$ participants and group sizes of $b = 15$, 30, or 50. The investigators specified a targeted treatment effect of 25%, where $\pi_C = 12\%$ and $\pi_E = 37\%$, which we define as the hypothesized alternative scenario. We similarly define the hypothesized null scenario to have $\pi_C = \pi_E = 12\%$. The type I error rate in each null scenario was defined as the proportion of simulated trials where either $P_{E>C}$ or 1-$P_{E>C}$ exceeded the stopping boundary value p_{stop}, and power in the hypothesized alternative scenario was defined as the proportion of simulated trials where $P_{E>C}$ exceeded p_{stop}. For each implementation, we simulated 10,000 trials under the hypothesized alternative scenario and a range of null scenarios with a survival probability between 0.04 and 0.95, and we calibrated p_{stop} so that the type I error rate under the hypothesized null scenario was 0.05. We computed $P_{E>C}$ after the primary outcome was obtained for a new group of subjects, and assessed each trial using a group sequential test with symmetric p_{stop} stopping boundaries. A trial was terminated early for efficacy or harm when $P_{E>C}$ exceeded either boundary. We ran all simulations with RStudio version 3.5.0 and the software for reproducing our simulation study is available at `https://github.com/prope012/Adaptive-Design-Evaluation`.

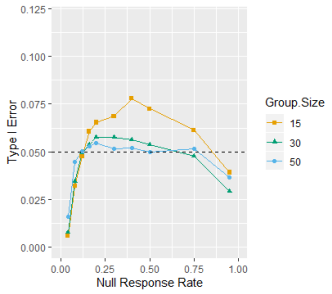
2.3 Results

Figure 2.1 depicts the type I error rate curves for each probability model paired with the weighted coin randomization method at sequential group sizes of 15, 30, and 50. The corresponding stopping boundary p_{stop} is reported in Table 2.1. Analogous figures for the mass-weighted urn design and modified permuted block design are provided in Sections 4 and 5 of Appendix A and show similar patterns. The independent beta-binomial model with $\pi^* = 0.50$ performed the worst in terms of type I error rate inflation, while the independent beta-binomial model with $\pi^* = 0.12$ performed the best. Type I error increased substantially for null response rates above 12% for all 3 group sizes in the former model, peaking at approximately 0.11. Conversely, the latter model produced type I error rates below the nominal level of 0.05. This better type I error control came with a substantial

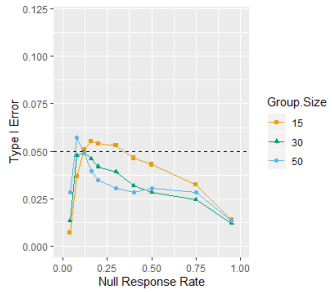
power loss, however. While power was about 90% when $\pi^* = 0.5$, it decreased as low as 80% with $\pi^* = 0.12$ (Table 2.2). Figure 2.1 reveals that the logistic regression model also improved type I error control with all values below 0.075 when the intercept prior location equaled $\log(\frac{0.12}{0.88})$.



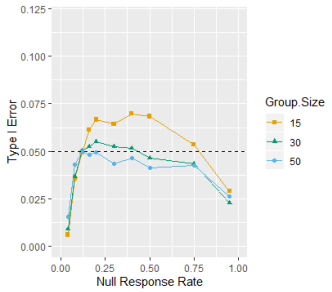
(a) Independent Beta-Binomial Model with Prior Mean 0.50



(b) Logistic Regression Model with Location = 0[1]



(c) Independent Beta-Binomial Model with Prior Mean 0.12



(d) Logistic Regression Model with Location = $\log(0.12/0.88)$[1]

[1] Refers to the Student-t prior location of the intercept parameter

Figure 2.1: Type I Error at Various Null Response Rates and Sequential Group Sizes by Probability Model (Weighted Coin Randomization Method)

Table 2.1: Posterior Probability Stopping Boundary p_{stop} by Sequential Group Size, Probability Model, and Prior Mean Response Rate (Weighted Coin Randomization Method)

Prior Mean		**Probability Model**	
Response Rate	Group Size	Independent Beta-Binomial	Logistic Regression
0.50[1]	15	0.9872	0.9925
	30	0.9860	0.9920
	50	0.9842	0.9890
0.12[1]	15	0.9972	0.9941
	30	0.9963	0.9931
	50	0.9944	0.9911

[1] Refers to a Beta(1,1) prior distribution

[2] Refers to a Beta(0.24,1.76) prior distribution

We also evaluated each model in terms of average sample size and power under the hypothesized alternative scenario. When $b = 30$, power notably dropped from 0.90 to 0.82 and sample size increased from 82.6 to 95.3 between the independent beta-binomial models with $\pi^* = 0.50$ versus 0.12 (Table 2.2). The logistic regression models also outperformed the independent beta-binomial model with prior mean 0.12 insofar as they produced higher power (0.87 when $b = 30$) and smaller average sample sizes (89 when b = 30).

Table 2.2: Power (Average Sample Size) under the Alternative Hypothesis by Sequential Group Size, Probability Model, and Prior Mean Response Rate (Weighted Coin Randomization Method)

Prior Mean		**Probability Model**	
Response Rate	Group Size	Independent Beta-Binomial Model	Logistic Regression
0.50[1]	15	0.89 (79.1)	0.86 (84.0)
	30	0.90 (82.6)	0.87 (88.8)
	50	0.92 (89.3)	0.90 (92.3)
0.12[2]	15	0.80 (89.9)	0.86 (84.0)
	30	0.82 (95.3)	0.87 (88.6)
	50	0.86 (98.3)	0.89 (94.2)

[1] Refers to a Beta(1,1) prior distribution

[2] Refers to a Beta(0.24,1.76) prior distribution

The main selling point of response-adaptive trials is they reduce the number assigned to the inferior treatment. In Figure 2.2 we plotted the empirical cumulative distribution functions for the difference in the number of subjects assigned to the treatment and control, $N_E - N_C$, under the hypothesized alternative scenario with $b = 30$. Analogous figures for $b = 15$ and $b = 50$ exhibited similar patterns. Because larger, positive values of $N_E - N_C$ indicate more subjects were assigned to the superior treatment, curves with more mass shifted to the right indicate better model performance. The independent beta-binomial model with $\pi^* = 0.5$ performed least desirably in this regard, whereas $\pi^* = 0.12$ performed most desirably; the latter provided the worst power, however.

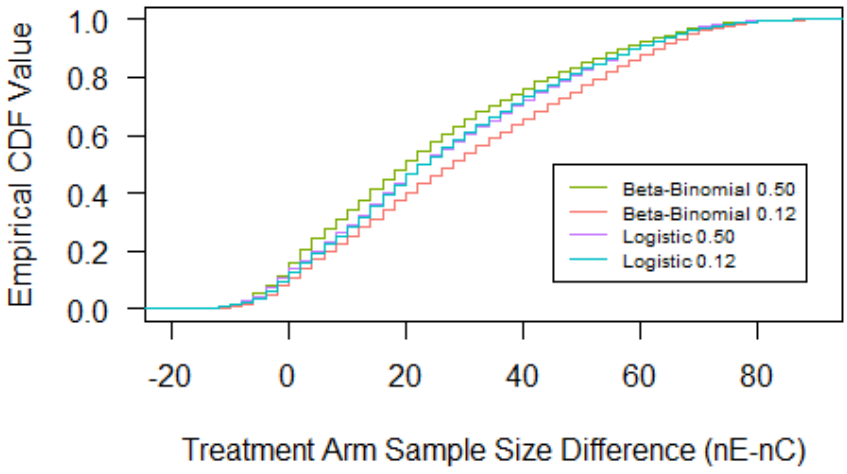


Figure 2.2: Empirical Cumulative Distribution Functions for the Difference in Treatment Arm Sample Sizes, $N_E - N_C$, under the Alternative Hypothesis for each Probability Model using a Group Size of 30 (Weighted Coin Randomization Method)

Table 2.3 contains summary measures for the three randomization methods, which performed comparably in terms of power, average sample size, and $N_E - N_C$ under the hypothesized alternative scenario. Each method also controlled type I error rate in a similar manner, as depicted by Figure A.1 in Appendix A. To compare these methods for achieving the targeted allocation across the trial, we calculated the root mean square error between the realized and intended allocations as follows:

(i) The actual number of people assigned to the treatment, $A_{E(\ell)}$, for each trial ℓ was determined.

(ii) The expected number of people assigned to the treatment, $E_{E(\ell)}$, was found for each sequential group m such that $E_{E(\ell)} = \sum_{j=1}^{m} bp_j^{(\ell)}$.

(iii) Root mean square error was calculated as the square root of the mean squared difference across all 10,000 trials and formally arises as:

$$\sqrt{\frac{1}{n}\sum_{\ell=1}^{n}\cdot(A_{E(\ell)}-E_{E(\ell)})^2} \tag{2.7}$$

To evaluate the tendency to allocate more subjects to the inferior arm under the hypothesized alternative scenario, we also computed the proportion of trials in which $N_E < N_C$, i.e. when there was a sample size imbalance in favor of the inferior treatment.

The mass-weighted urn and modified permuted block designs were consistently more accurate than the weighted coin method, as shown in Table 2.3. The modified permuted block design also has lower root mean square error values and appears to be more accurate than the mass-weighted urn design. The proportion of trials in which $N_E < N_C$ varies substantially across methods. While about 18% of trials using the weighted coin design with $b = 50$ had more participants assigned to the control than the treatment, this number decreased to 7.5% for the mass-weighted urn design and even further to 0.4% for the modified permuted block design. For comparison purposes, the response-adaptive designs evaluated in Table 2.3 were repeated using fixed, equal randomization and their analogous summary measures are provided in Table 2.4. As expected, this allocation was associated with higher power and slightly smaller average sample sizes, but more subjects on the control arm.

Table 2.3: Summary Measures Comparing Three Randomization Methods **using Response-Adaptive Allocation** for the Logistic Regression Model with Prior Intercept Location = log(0.12/0.88)

Summary Measure	**Group Size**[1]	**Weighted Coin**	**Mass-Weighted Urn**	**Modified Permuted Block**
Power (Average Sample Size) under the Alternative	15	0.87 (84.0)	0.86 (84.4)	0.86 (84.4)
	30	0.87 (88.6)	0.87 (88.3)	0.87 (87.6)
	50	0.89 (94.2)	0.89 (94.3)	0.89 (94.3)
Root Mean Square Error under the Null (Alternative)	15	5.58 (4.11)	1.77 (1.30)	1.34 (1.04)
	30	5.65 (4.32)	1.28 (0.98)	0.91 (0.68)
	50	5.71 (4.56)	0.99 (0.81)	0.64 (0.47)
Average $N_E - N_C$ (Proportion of $N_E < N_C$)	15	30.2 (0.06)	30.8 (0.04)	30.7 (0.03)
	30	26.6 (0.10)	26.6 (0.04)	26.3 (0.00)
	50	21.0 (0.18)	20.8 (0.08)	20.8 (0.00)

[1] The block size after which adaptations are made. This number differs from the maximum sample size of 150.

Table 2.4: Summary Measures Comparing Three Randomization Methods **using 1:1 Allocation** for the Logistic Regression Model with Prior Intercept Location = log(0.12/0.88).

Summary Measure	**Group Size**[1]	**Weighted Coin**	**Mass-Weighted Urn**	**Modified Permuted Block**
Power (Average Sample Size) under the Alternative	15	0.91 (79.1)	0.91 (77.3)	0.90 (80.5)
	30	0.91 (85.1)	0.93 (80.6)	0.92 (81.8)
	50	0.92 (91.9)	0.92 (92.1)	0.92 (92.0)
Root Mean Square Error under the Null (Alternative)	15	6.04 (4.49)	1.89 (1.38)	1.55 (1.14)
	30	6.05 (4.61)	1.36 (1.00)	0.00 (0.00)
	50	6.07 (4.83)	1.06 (0.84)	0.00 (0.00)
Average $N_E - N_C$ (Proportion of $N_E < N_C$)	15	0.21 (0.47)	0.02 (0.41)	-0.01 (0.40)
	30	0.03 (0.45)	-0.04 (0.29)	0.00 (0.00)
	50	-0.05 (0.46)	-0.01 (0.25)	0.00 (0.00)

[1] The block size after which adaptations are made. This number differs from the maximum sample size of 150.

2.3.1 Implementation of the Proposed Methodology in Practice

Based on the above simulations, in phase II trials considering a response-adaptive strategy, we recommend using the logistic regression probability model with either the proposed permuted block design or the mass-weighted urn design. [34] In general, we suggest using weakly informative t-distribution priors with an intercept location of $\log\left(\frac{\pi^*}{1-\pi^*}\right)$, where π^* reflects the hypothesized response rate in the control arm. The proposed methods are not robust to patient drift, a well-documented phenomenon characterized by evolving patient prognoses that can exacerbate type I error rate inflation and introduce bias in response-adaptive designs, [6] and thus may not be appropriate for definitive phase III trials. However, when patient drift is of concern, a block-stratified analysis and randomization test could be

implemented alongside the proposed methods. [6, 9, 10] Finally, while we have shown that these methods reduce type I error rate sensitivity to the underlying response probability, and thus have practical merit, they may not address this issue in a universal manner. When designing any trial, the operating characteristics engendered by the chosen design should be thoroughly evaluated. With these caveats in mind, our proposed methodology may be used to design and implement a Bayesian response-adaptive phase II clinical trial using the following step-by-step approach. R code available on the first author's GitHub page (`https://github.com/prope012/Adaptive-Design-Evaluation`) may be leveraged to complete steps ending with an asterisk ($*$).

Design Phase:

1. Elicit the hypothesized control and treatment arm response rates from the investigator.

2. Determine the sample size required for the two-sample binomial proportion test to achieve the targeted power of the trial. This number will be used as an initial value in an iterative procedure to find the sample size of the recommended Bayesian response-adaptive design.

3. Determine an appropriate sequential group number and size after taking trial feasibility and the sample size estimate from step 2 into consideration.

4. Simulate 10,000 trials under the hypothesized null scenario using the logistic regression probability model with weakly informative t-distribution priors and the chosen randomization procedure (mass-weighted urn or modified permuted block design). Use the sample size estimate from step 2 in the first iteration.*

5. Determine the stopping boundary value p_{stop} that controls type I error rate at 0.05, i.e. for which 5% of the simulated trials had either $P_{E>C}$ or $1 - P_{E>C}$ exceed p_{stop}.*

6. Simulate 10,000 trials under the hypothesized alternative scenario using the same probability model and randomization procedure as in step 4.*

7. Determine the power, defined as the proportion of simulated trials where $P_{E>C}$ exceeds p_{stop} determined in step 5.*

8. If the targeted power has not been met, repeat steps 4-7 using an updated sample size estimate. Letting $N_{current}$ and N_{new} be the current and updated sample sizes, respectively, and Φ be the quantile function of the standard normal distribution, we recommend using the following identity to update the sample size estimate:

   - $N_{new} = \text{ceiling}\left(N_{current} \cdot \frac{(\Phi(0.975)+\Phi(\text{Targeted Power}))^2}{(\Phi(0.975)+\Phi(\text{Current Power}))^2}\right)$

Implementation Phase:

1. Follow the steps in the Design Phase to determine the sample size of the recommended Bayesian response-adaptive design required to achieve the targeted power.

2. Set $m = 1$ and $p_1 = 0.5$.*

3. Generate a randomization schedule for group m using the chosen randomization procedure such that the probability of being allocated to the treatment arm is p_m.*

4. After obtaining the primary outcomes of group m, compute $P_{E>C}$ and update $p_{m+1} = \max\{0.25, \min\{0.75, P_{E>C}\}\}$ (restricted to be between 0.25 and 0.75). Set $m = m+1$.*

5. Repeat steps 3-4 until the trial has finished.

2.4 Discussion

Bayesian response-adaptive clinical trials using the independent beta-binomial probability model with a weighted coin randomization method have been criticized for their high type I error rates, insufficient power, and inability to prevent more patients from being allocated to the inferior treatment arm by chance. Our simulations demonstrate that the logistic regression probability model engenders high power while providing better type I error rate

control, whereas the alternative randomization methods (the proposed modified permuted block design and the mass-weighted urn design) [34] substantially reduce the risk of a subject imbalance in the wrong direction. They also more effectively target the evolving allocation ratio throughout the trial.

This research has several limitations. We focused on a large difference in treatment efficacy (37% vs. 12%), though these methods may also be considered for phase II trials targeting smaller effect sizes, in which case the simulation results may indicate larger average sample sizes and smaller differences in treatment arm sample sizes. It is important to keep in mind, however, that Bayesian response-adaptive designs exhibit the most benefit for large treatment differences. [36] We also considered group sizes of 15, 30, and 50. Smaller group sizes could have been used to, on average, further reduce the number assigned to the inferior arm. However, smaller group sizes would increase the logistical burden of the trial, as randomization schedules would need to be modified with greater frequency, and increase variability early in the trial, potentially exacerbating type I error inflation and the probability of assigning more participants to the inferior arm.

Further research is needed to evaluate our recommendations in a multi-arm trial setting. As discussed by Trippa et al., another way to improve the performance of multi-arm response-adaptive design is by incorporating tuning parameters into the randomization algorithm that regulate the exploration versus exploitation trade-off. [37] A possible extension of our work would be to assess the performance of a multi-arm response-adaptive design using our recommended probability model and a tuning randomization algorithm. In the context of multi-arm bandit models, Villar et al. have also demonstrated that non-myopic randomization procedures may influence the operating characteristics of response-adaptive designs. [38, 39] Defining our randomization procedure in terms of the patient horizon, or the total patient population, is another possible direction for future work. [39, 40]

2.5 Appendix

2.5.1 Additional Details for the Independent Beta-Binomial Probability Model:

Note from equation (2.1) that π_E and π_C follow the same prior distribution. This assumption suggests that the prior treatment difference is null and that posterior evidence for a difference between π_E and π_C will reflect differences in the observed data for each arm. For a Beta(α_z, β_z) prior distribution, $n_0 = \alpha_z + \beta_z$ where α_z and β_z may be interpreted as the prior number of responders and non-responders, respectively. [41] Because the variance of the Beta distribution decreases monotonically in $\alpha_z + \beta_z$, larger values of n_0 characterize a more informative prior. The hyperparameter π^* similarly equals $\frac{\alpha_z}{\alpha_z+\beta_z}$ and reflects the prior expected response rate in each arm. The conjugacy property of the beta-binomial model also facilitates efficient calculation of $P_{E>C}$ using adaptive quadrature rather than a Monte Carlo approach.

2.5.2 Additional Details for the Logistic Regression Probability Model:

The logistic regression probability model may be appealing for several reasons. In contrast with the beta distribution, the t-distribution has three model parameters: location, scale, and degrees of freedom (df), which can alter the position, spread, and kurtosis of the distribution, respectively. The ability to control kurtosis, or thickness of the tails, makes the t-distribution appealing when a weakly informative prior is desirable.

The conditionally conjugate Polya-Gamma Gibbs sampler used to facilitate joint sampling of β_0 and β_1 from the posterior distribution uses the key fact leveraged by Polson et al. that the binomial likelihood used in logistic regression can be characterized as a Gaussian mixture with respect to a Polya-Gamma distribution. [42] The logistic regression model also notably induces a prior correlation of 0.50 between π_E and π_C via the shared intercept parameter, β_0. Since our prediction for π_E depends on the original assumption for π_C, prior correlation between the two probabilities is reasonable. Because response-adaptive

trials make decisions early in the trial with little information, these prior assumptions may substantially affect the performance of the design.

Chapter 3

Alternative Models and Randomization Techniques for Bayesian Response-Adaptive Randomization with Binary Outcomes

3.1 Introduction

Randomized controlled trials (RCTs) are widely regarded as the gold standard for studying the efficacy of new treatments. Randomly assigning participants to either treatment or control precludes selection bias and, on average, balances groups with respect to both known and unknown confounders so that unbiased treatment effect estimates can be obtained. [36, 43] Two-arm RCTs traditionally allocate participants to treatments using 1:1 randomization throughout the course of the trial to obtain unbiased comparisons and high power. [5, 36] However, physicians and potential participants often find it appealing when a trial allows for a higher probability of being randomized to the better performing arm as the trial data

accrue. [44, 45] For this reason, response-adaptive randomization (RAR), which alters the allocation ratio based on accruing data in favor of the empirically superior treatment, has garnered considerable attention in Phase II clinical trials. [2–4]

A recent example of a trial using a Bayesian RAR design is the Advanced R^2Eperfusion STrategies for Refractory Cardiac Arrest (ARREST) trial (ClinicalTrials.gov, NCT03880565), which assessed the impact of extracorporeal membrane oxygenation-facilitated resuscitation versus standard advanced cardiac life support on survival to hospital discharge in adults who experienced an out-of-hospital cardiac arrest and refractory ventricular fibrillation. [13, 14] Another example is the ACCESS trial (ClinicalTrials.gov, NCT03119571) in adults who were resuscitated from an out-of-hospital ventricular fibrillation or ventricular tachycardia cardiac arrest, which evaluated whether being admitted directly to the cardiac catheterization laboratory versus the intensive care unit improved survival to hospital discharge with no more than moderate disability, including the ability to walk on their own (modified Rankin scale score ≤ 3). [46]

Bayesian RAR is most easily conceptualized in the context of a two-arm trial with binary outcomes. Let π_E and π_C respectively denote the true, but unknown, probability of response under the E=treatment and C=control, $\boldsymbol{D}_n$ the available data from n participants, and $P_{E>C} = \Pr(\pi_E > \pi_C|\boldsymbol{D}_n)$ the posterior probability that the probability of a response is greater under the treatment than the control. Thompson [26] first proposed randomizing the $(n+1)$-st participant to the treatment with probability $P_{E>C}$ and to the control with probability $P_{C>E}$, which is equivalent to 1-$P_{E>C}$ in the two-arm setting. When the treatment is efficacious, this sampling method, in addition to several modifications to Thompson sampling that have since been proposed, gives participants a greater chance of receiving the treatment during the trial. [4, 25] Yet, Bayesian RAR remains contentious and has seen limited use in practice. It is commonly criticized for inducing biased treatment effect estimates, type I error inflation in the presence of trial population drift or effect heterogeneity over time, reduced power relative to 1:1 allocation, and non-negligible probabilities of allocating

more participants to the inferior treatment. [5–8]

When planning a Phase II trial, investigators often seek a design with specified type I and II error rates that is ethically, administratively, and economically feasible. For these reasons, a group sequential design is frequently used to facilitate early stopping for efficacy, harm, or futility at pre-specified interim sample sizes throughout the trial, which protects participants from unnecessary exposure to ineffective or harmful treatments and allows limited resources to be allocated to other trials (Jennison and Turnbull, 2000). Response-adaptive randomization may be considered when investigators wish to maximize in-trial patient benefit, or equivalently minimize in-trial patient harm. Over the last several decades, researchers have studied the potential patient benefits of combining group sequential and RAR methodology to limit exposure to an inferior arm. For example, Yao and Wei [47] extended the randomized play-the winner (RPW) rule to a multi-stage setting and found that RAR provides ethical benefits to study participants while maintaining adequate power. Morgan and Coad [17] investigated other group sequential adaptive sampling rules, including the drop-the-loser, randomized Polya urn, and sequential maximum likelihood estimation rules, and observed useful reductions in the number of treatment failures. Coad and Rosenberger [48] further found that combining the RPW rule and triangular test for clinical trials with binary responses is an effective way to reduce the expected number of treatment failures and suggested extending their procedure to a multi-stage adaptive design similar to Yao and Wei. [47]

In this paper, we propose a metric for comparing group sequential designs based on the cohort most acutely impacted by the choice of design and illustrate how this metric may be applied to select a design in the ARREST and ACCESS contexts. RAR designs are commonly compared using inferential and estimation metrics (e.g., type I error, power, and bias) rather than measures of patient benefit, which remain under-reported and have received little attention in the RAR literature. [49] This is in part because existing patient benefit metrics, including the expected number of trial failures, the proportion of patients

assigned to the inferior arm, and the probability of a treatment imbalance in the wrong direction, are often limited by failures to hold type I and II error rates constant or to account for the different sample size requirements of the designs under consideration. [16–18,49] One approach to correct for the latter issue is to compare designs with respect to the expected number of failures within a finite patient horizon. [38,39] However, as far as we are aware, no specific guidance exists for selecting an appropriate horizon and there is a need, as suggested by Robertson et al., [49] for patient benefit metrics that clearly quantify the ethical properties of RAR designs while considering patient benefit both within and outside of a trial. Our proposed metric improves on existing patient benefit metrics by considering a set of feasible group sequential designs with equal type I and II error rates and measuring the expected number of failures in the fixed group of individuals who are directly impacted by the design choice. Namely, those who would participate in the trial if enrollment were open when they become eligible.

This paper is organized as follows. In Section 2, we introduce our proposed metric. In Section 3, we discuss the underpinnings of Bayesian RAR group sequential designs and the probability model and randomization scheme used in our particular implementation. We consider six different variations of Bayesian RAR, including two modifications to Thompson sampling and Bayesian versions of Neyman allocation, the optimal allocation of Rosenberger et al., [50] the doubly adaptive biased coin design (DABCD) of Hu and Zhang, [51] and the efficient randomized adaptive design (ERADE) of Hu et al. [52] In Section 4, we outline our simulation study and provide results in the context of the ARREST and ACCESS trials. Although these results are in terms of a two-arm trial with a binary primary outcome and Bayesian RAR, our metric may be directly applied to contexts comparing multiple treatments or other RAR procedures. We conclude by discussing how our metric may be applied to select a design.

3.2 Design Comparison Metric

Our proposed metric measures the expected number of failures in the cohort that is directly impacted by the design choice for a set of practically feasible designs with equal type I and II error rates. We define the size of this cohort as the maximum sample size among all designs compared and call this group the *potential study sample*, as this cohort comprises all the individuals who will participate in the trial if it is open when they become eligible to enroll.

Let $\mathbf{T_L} = (T_1, ..., T_L)$ denote a set of designs satisfying budgetary, recruitment, or logistical constraints and achieving a desired significance level, α, and power, $1 - \beta$, to detect the hypothesized treatment effect. In Section 4, we will discuss this set in the context of the ARREST and ACCESS trials. Let $n_{max,l}$ denote the maximum sample size for design T_l that provides the specified error rates. Then, the size of the potential study sample may be defined as $n_{max} = \max_{l=1,...,L} n_{max,l}$. For a particular design T_l, suppose that n is the sample size of the trial at the time the trial is stopped, and $\mathbf{Z_n} = (Z_1, ..., Z_n)$ and $\mathbf{Y_n} = (Y_1, ..., Y_n)$ are the randomization assignments and observed outcomes for trial participants $i = 1, ..., n$, respectively, where $Z_i \in \{C, E\}$ and $Y_i \in \{0, 1\}$. Conditional on the observed data $\mathbf{D_n} = (\mathbf{Y_n}, \mathbf{Z_n})$, we define the number of failures within the potential study sample as:

$$m(\mathbf{D_n}, n_{max}, \pi_C, \pi_E) = \sum_{i=1}^{n}(1 - Y_i) + (n_{max} - n) \cdot [(1 - \pi_C) - d_l(\mathbf{D_n})(\pi_E - \pi_C)] \quad (3.1)$$

where $d_l(\mathbf{D_n}) = \mathrm{I}\{\text{reject } H_0\}$ is an indicator for whether the treatment is declared efficacious in design T_l. That is, (3.1) reflects the number of failures in the actual trial plus the expected number of failures among the remaining individuals in the potential study sample under the arm the trial recommends. Because the observed data and decision rule depend on the

design, we define the proposed metric as follows:

$$\begin{aligned} m_l = \mathrm{E}[L(m)|T_l] = \mathrm{E}\left[\mathrm{E}[L(m)|\mathbf{D_n}, T_l]\Big|T_l\right] = \int_{\mathbf{Y_n}}\int_{\mathbf{Z_n}} L(m)\cdot dF(\mathbf{Y_n}, \mathbf{Z_n}|T_l) \\ = \int_{\mathbf{Y_n}}\int_{\mathbf{Z_n}} L(m)\cdot dF(Y_n|Z_n)\cdot dF(Z_n|\mathbf{Y_{n-1}}, \mathbf{Z_{n-1}}, T_l)\cdots dF(Y_1|Z_1)\cdot dF(Z_1|T_l) \end{aligned} \tag{3.2}$$

for some loss function $L(m)$. We focus on $L(m) = m$; however, other loss functions may be used to prioritize different objectives. For example, using $L(m) = m^2$ (which is analogous to the traditional L2 loss function) would penalize designs that have a greater chance of yielding many failures in the potential study sample by squaring them. Alternatively, the median rather than mean loss could be used to compare designs while limiting the impact of outliers. The second line in (3.2) emphasizes that designs under consideration may use RAR, in which case the assignments are dependent on T_l and all previous assignments and outcomes, whereas the outcomes are only dependent on the assignment.

We approximate m_l using Monte Carlo integration; that is, we iteratively simulate $\mathbf{Z_n}$ and $\mathbf{Y_n}$ under design T_l, calculate $m(\mathbf{D_n}, n_{max}, \pi_C, \pi_E)$ for each realization of the data, and then take the average. We make the working assumption that participants who would have enrolled in the trial had it continued will receive the treatment when it is shown to be superior and the control otherwise. This approach provides a way to reward a design that stops early to recommend treatment when it is in fact efficacious. Design T_{l^*} with l^* = argmin$_l$ m_l is optimal with respect to the proposed metric among the set of $\mathbf{T_L}$ designs under consideration that provide the same type I and II error rates.

Our approach to finding the optimal group sequential design is similar to the horizon problem, which seeks a design that minimizes the expected number of patients receiving the inferior treatment within the trial and in the future, initially proposed by Anscombe. [53] Our potential study sample may be viewed as a fixed patient horizon that prioritizes individuals who are implicitly impacted by the decision to end a trial early or proceed with a smaller design. Larger patient horizons could have been considered; however, because we

only consider designs with the same type I and II error rates, the long-run impact of these designs is similar, with each recommending treatment at the same rates under both the null and alternative hypotheses.

3.3 Bayesian Response-Adaptive Randomization Group Sequential Designs

Response-adaptive randomization alters the allocation ratio throughout the trial based on accumulated data (i.e. past treatment assignments and patient outcomes) in order to achieve particular design objectives. In this paper, we consider various Bayesian RAR approaches that aim to maximize efficiency or reduce exposure to an inferior treatment. We implement each approach in a group sequential context that analyzes data up to $j = 1, ..., J$ times and allows a trial to be stopped early for efficacy or harm. When the trial continues, each Bayesian RAR approach modifies the randomization ratio for the next group of participants accordingly. Letting p_j denote the target randomization probability to the treatment arm for group j, we assume $p_1 = 0.5$ and thereafter modify p_j using data from the preceding interim analysis until the trial is stopped. We implement the Bayesian RAR designs using one probability model, four boundary shapes, eight adaptive modifications to the randomization probability, and one method for generating assignments under the targeted randomization probability.

3.3.1 Probability Model

Two-arm Bayesian RAR designs with a binary primary outcome have been conventionally implemented using the independent beta-binomial probability model for π_E and π_C. We instead employ a logistic regression probability model with weakly informative t-distribution priors. This model has been shown to have increased power and reduced type I error rate sensitivity to the underlying response probability compared to a Beta-Binomial modeling

approach due to shrinkage that arises from placing less prior density on extreme values of the response probabilities. [54] A plot of the prior distributions for the Beta-Binomial and logistic regression models on the probability and log-odds scales is provided in Section 1 of Appendix B, which indicates that the prior distribution for the Beta-Binomial model is asymmetric on the log-odds scale and places more mass on very small values of π_E. Letting $k = E$ and C denote the experimental treatment and control arms, respectively, the logistic regression probability model arises from a logit transformation of π_k:

$$\text{logit}(\pi_k) = \log\left(\frac{\pi_k}{1-\pi_k}\right) = \beta_0 + \beta_1 \cdot (\text{I}\{k = E\} - 0.5) \tag{3.3}$$

and is formally defined as follows:

$$\begin{aligned} Y_k|\beta_0, \beta_1 &\sim \text{Binomial}\left(n_k, \frac{\exp(\beta_0 + \beta_1 \cdot (\text{I}\{k = E\} - 0.5))}{1 + \exp(\beta_0 + \beta_1 \cdot (\text{I}\{k = E\} - 0.5))}\right),\ k = E, C \\ \text{and } \beta_0 &\sim t_7\left(\log\left(\frac{\pi_C}{1-\pi_C}\right), 2.5\right) \perp\!\!\!\perp \beta_1 \sim t_7(0, 2.5), \end{aligned} \tag{3.4}$$

where $t_\nu(\mu, \sigma)$ denotes a generalized t-distribution with ν, μ, and σ representing the degrees of freedom, location, and scale parameters, respectively. [54] The ν parameter advantageously allows investigators to induce statistical models that are robust to outliers by altering the kurtosis, or thickness of the tails, of the prior. [55] It may also be used to control the influence of prior beliefs on statistical inference. As shown in equation (3.4), we use a prior intercept location of $\log(\pi_C/(1-\pi_C))$; this value represents an estimate of the intercept parameter under the hypothesized null scenario and is easily derived by setting $\pi_k = \pi_C$ and $\beta_1 = 0$ in equation (3.3). Following JGhosh2015, we use a conditionally conjugate Polya-Gamma Gibbs sampler to jointly sample β_0 and β_1 from the posterior distribution.

3.3.2 Adaptive Modifications

Thompson Sampling

The first two adaptive modifications to the randomization probability that we consider are variations of Thompson sampling. We define $P_{E>C}$ as the proportion of posterior samples where $\beta_1 > 0$, which signifies that the treatment arm has greater odds of response than the control. Although it is intuitively appealing to set p_j=$P_{E>C}$, the variability of $P_{E>C}$ with small sample sizes may reduce power and exacerbate the nontrivial probability of a treatment imbalance in the wrong direction. [4] We implement two modifications proposed by Thall and Wathen that stabilize $P_{E>C}$ prior to using it as a randomization probability. [4,25] The first modification randomizes the first group using 1:1 allocation and sets $p_{rst,j} = \min\{0.75,\ \max\{0.25, P_{E>C}\}\}\ \forall j = 2, ..., J$. This restriction limits power loss and selection bias that may result from extreme allocation to one arm. The second modification adapts $P_{E>C}$ as follows:

$$p_{tun,j} = \frac{[P_{E>C}]^{C_j}}{[P_{E>C}]^{C_j} + [1 - P_{E>C}]^{C_j}} \tag{3.5}$$

where C_j is a positive tuning parameter equal to 0 for $j = 1$ and $0.5 \cdot (j-1)/(J-1)\ \forall j = 2, ..., J$. Thall and Wathen [4] originally proposed the use of a positive tuning parameter in a fully sequential trial; however, our expression in (B.2) is an adaptation of their proposal to a group sequential context. This modification is conservative in the beginning of the trial when it constrains p_j near 0.5 (e.g., when $j = 1$, $C_j = 0$ and hence $p_{tun,1} = 0.5$), but becomes more aggressive later on as C_j incrementally approaches 0.5. A plot of the p_j values arising from these two adaptations for various $P_{E>C}$ is provided in Section 2 of Appendix B.

Optimal Allocations

We consider two optimal allocation ratios for two-arm clinical trials with a binary outcome. The first is Neyman allocation, which maximizes power by minimizing the variance of the

test statistic. [56] It is important to note that when $\pi_E + \pi_C > 1$, Neyman allocation undesirably allocates more participants to the arm with the lower response probability to maximize power. The second is the optimal allocation ratio proposed by Rosenberger et al. [50] that, for a fixed variance of the test statistic, minimizes the expected number of non-responders. We herein refer to this procedure as RSIHR allocation. For a fixed number of trial participants, $n = n_E + n_C$, and assuming the difference in sample proportions is the test statistic of interest, one may show that the optimality criterion for Neyman allocation is satisfied by allocating the following proportion of participants to the treatment arm: [56]

$$p_{ney} = \frac{n_E}{n} = \frac{\sqrt{(\pi_E)(1-\pi_E)}}{\sqrt{(\pi_E)(1-\pi_E)} + \sqrt{(\pi_C)(1-\pi_C)}} \tag{3.6}$$

and the expected number of non-responders is minimized using. [50]:

$$p_{rsihr} = \frac{n_E}{n} = \frac{\sqrt{\pi_E}}{\sqrt{\pi_E} + \sqrt{\pi_C}} \tag{3.7}$$

Because the response probabilities are unknown, these optimal allocations must be approximated using sample data from an adaptive sequential design. Rosenberger et al. [50] proposed estimators for (3.6) and (3.7) that replace π_E and π_C with the current sample proportions and achieve the optimal allocations in the limit. They suggest randomizing the n^{th} participant to the treatment arm using the optimal allocation estimator based on the data from the first $n-1$ participants. Instead, we generalize these optimal allocation ratios to a Bayesian RAR group sequential setting as described in Section 3.2.4.

Efficient Adaptations

The final two adaptive modifications that we consider desirably lower the variance of the randomization procedure, which is inversely related to the average power of a randomization procedure for a given allocation proportion. [57] We first consider the doubly adaptive biased coin design (DBCD) proposed by Hu and Zhang, [51] which tends to the targeted allocation

in the limit and has a smaller variance than the RPW rule and the adaptive randomized design. This procedure uses the following allocation function, g, to precisely target any desired allocation proportion:

$$g(x,y) = \frac{y(y/x)^\gamma}{y(y/x)^\gamma + (1-y)((1-y)/(1-x))^\gamma} \tag{3.8}$$

where x is the realized allocation proportion, y is the targeted allocation proportion, $g(0,y) = 1$, $g(1,y) = 0$, and $\gamma \geq 0$ is a constant chosen to control the trade-off between allocation-randomness and the asymptotic variance of the design. [51] We set $\gamma = 2$ throughout our simulation study. Implementing this design in a fully-sequential fashion, Hu and Zhang [51] suggest randomizing the n^{th} participant to the treatment arm with probability $p_{dbcd} = g(r_{n-1}, \rho(\widehat{\pi}_{E,n-1}, \widehat{\pi}_{C,n-1}))$, where r_{n-1} is the proportion of participants receiving the treatment after $n-1$ allocations, $\rho(\pi_E, \pi_C)$ is the targeted allocation for the treatment arm, and $\widehat{\pi}_{k,n-1}$ is an estimate of π_k based on data for the first $n-1$ participants.

We next consider the Efficient Randomized Adaptive Design (ERADE) proposed by Hu et al.,ERADE which attains the Cramer-Rao lower bounds on the allocation variances for any allocation proportions. After using a restricted randomization procedure to allocate an initial n_0 participants to treatment or control, ERADE allocates the n^{th} participant to the treatment arm with the following probability:

$$p_{erade} = f(r_{n-1}, \widehat{\rho}) = \begin{cases} \alpha\rho(\widehat{\pi}_{E,n-1}, \widehat{\pi}_{C,n-1}), & \text{if } r_{n-1} > \rho(\widehat{\pi}_{E,n-1}, \widehat{\pi}_{C,n-1}) \\ \rho(\widehat{\pi}_{E,n-1}, \widehat{\pi}_{C,n-1}), & \text{if } r_{n-1} = \rho(\widehat{\pi}_{E,n-1}, \widehat{\pi}_{C,n-1}) \\ 1 - \alpha(1 - \rho(\widehat{\pi}_{E,n-1}, \widehat{\pi}_{C,n-1})), & \text{if } r_{n-1} < \rho(\widehat{\pi}_{E,n-1}, \widehat{\pi}_{C,n-1}) \end{cases} \tag{3.9}$$

where $0 \leq \alpha \leq 1$ is a constant reflecting the degree of randomization. Per Hu et al., [52] we use $\alpha = 2/3$ throughout our simulation study. We generalize both DBCD and ERADE to a Bayesian RAR group sequential setting and use them to target Neyman and RSIHR allocation as described in Section 3.2.4.

Generalizing to a Bayesian RAR Group Sequential Setting

Neyman allocation, RSIHR allocation, DBCD, and ERADE were originally proposed for a fully sequential design. Instead, we generalize these adaptive modifications to a Bayesian RAR group sequential setting. We randomize the first group of participants using $p_1 = 0.5$ and adapt $p_j \ \forall j = 2, ..., J$ using posterior mean estimates of the optimal allocations in (3.6) and (3.7) which arise as:

$$p_{ney,j} = \mathrm{E}\left[\frac{\sqrt{(\pi_E)(1-\pi_E)}}{\sqrt{(\pi_E)(1-\pi_E)} + \sqrt{(\pi_C)(1-\pi_C)}}\middle| \boldsymbol{D_{(j-1)b}}\right] \tag{3.10}$$

$$p_{rsihr,j} = \mathrm{E}\left[\frac{\sqrt{\pi_E}}{\sqrt{\pi_E} + \sqrt{\pi_C}}\middle| \boldsymbol{D_{(j-1)b}}\right] \tag{3.11}$$

where b is the sequential group size. These posterior means are estimated using Gibbs sampling with algorithmic details provided in Section 3 of Appendix B.

3.3.3 Randomization Method

The weighted coin randomization method is generally used to allocate participants among treatments in Bayesian RAR trials. This method assumes $p_{ji} = p_j \ \forall j = 1, ..., J$, where p_{ji} is the conditional probability that the i^{th} participant in the j^{th} group is allocated to the treatment. We instead use an alternative randomization method proposed by Zhao [34] called the mass weighted urn design (MWUD). Proper et al. [54] have previously shown that, in conjunction with the logistic regression probability model, this method substantially reduces the probability of a treatment imbalance in favor of the inferior arm and more precisely targets the desired allocation throughout the trial.

The MWUD uses an urn randomization scheme with one ball per treatment in the urn. The randomization schedule is generated by consecutively drawing one ball from the urn with replacement, where the probability of selecting a ball is proportional to its mass. The initial mass of the balls are αp_j and $\alpha(1 - p_j)$ for the treatment and control, respectively.

After every selection, the chosen ball loses one unit of mass that is redistributed in proportion to the target allocations. This approach can be implemented using the following equation:

$$p_{ji} = \frac{\max[\alpha p_j - n^j_{i-1,E} + (i-1)p_j, 0]}{\max[\alpha p_j - n^j_{i-1,E} + (i-1)p_j, 0] + \max[\alpha(1-p_j) - n^j_{i-1,C} + (i-1)(1-p_j), 0]} \tag{3.12}$$

where $n^j_{i-1,k}$ is the number of participants on treatment k prior to the i^{th} allocation in group j, and α is a treatment imbalance parameter that restricts how far the realized allocation is permitted to deviate from the targeted allocation throughout the trial. [34] For example, when the current treatment allocation, $n^j_{i-1,E}$, exceeds the targeted allocation by more than αp_j, the next participant is assigned to the control with probability 1. Similar to Proper et al., [54] we use $\alpha = 3$ throughout our simulation study.

3.3.4 Interim Monitoring Boundaries

We perform interim monitoring as follows. For the two-sided setting, we compute $P_{E>C}$ using all available outcome data at each interim analysis. When $P_{E>C} \geq a_j$ or $P_{E>C} \leq b_j$, the trial is stopped early for efficacy or harm, respectively. For the one-sided setting, the trial is stopped early for efficacy when $P_{E>C} \geq a_j$ or futility when $P_{E>C} \leq b_j$. In either setting, when $b_j < P_{E>C} < a_j$ no decision is made and the trial continues. We consider both a two-sided hypothesis test setting with symmetric boundaries where $b_j = 1 - a_j$ and a one-sided hypothesis test setting with asymmetric boundaries where $b_J = a_J$ and $b_j \neq 1 - a_j$. Similar to Shi and Yin, [58] we use computer simulation to find the maximum sample sizes and $\{a_j, b_j\}$ $\forall j = 1, ..., J$ that control the frequentist properties of the Bayesian RAR group sequential designs (i.e. maintain type I error rate and power at the desired levels). We consider setting $a_j = a$ $\forall j = 1, ..., J$ to implement flat boundaries throughout the trial that are similar to Pocock. We also consider setting $a_j = \Phi\left(\sqrt{J/j} \cdot c\right)$, where Φ is the standard normal cumulative distribution function and c is an arbitrary constant. Because

this quantity becomes smaller with increasing j, these boundaries are similar to O'Brien-Fleming (OBF) and become more aggressive as the trial proceeds. A description of how to find the maximum sample sizes and posterior probability stopping boundaries for a given design is provided in Section 4 of Appendix B.

3.4 Simulation Study

3.4.1 Design Considerations for the ARREST and ACCESS Trials

Motivated by the ARREST and ACCESS trials, we performed a simulation study with 10,000 simulated trials per scenario to find the optimal two-arm group sequential design with respect to m_l in (3.2). For each simulated trial, we generated a data set consisting of two potential outcomes for each individual in the potential study sample: one assuming they received the treatment and one assuming they received the control. For each trial participant, we set their observed outcome equal to the potential outcome corresponding to their randomization assignment. We then computed the expected number of failures among the group of individuals within the potential study sample but not the trial itself using the trial conclusion, as in (3.1). The same data sets were used to evaluate each design in each scenario.

To reflect the ARREST trial, we considered a hypothesized null scenario of $\pi_C = \pi_E = 12\%$ and a hypothesized alternative scenario of $\pi_C = 12\%$ and $\pi_E = 37\%$. To reflect the ACCESS trial, we considered a hypothesized null scenario of $\pi_C = \pi_E = 50\%$ and a hypothesized alternative scenario of $\pi_C = 50\%$ and $\pi_E = 65\%$. For completeness, we also considered contexts with hypothesized null scenarios of $\pi_C = \pi_E = 5\%$ and $\pi_C = \pi_E = 80\%$ with corresponding hypothesized alternative scenarios of $\pi_C = 5\%$ and $\pi_E = 20\%$ and $\pi_C = 80\%$ and $\pi_E = 95\%$, respectively.

For each treatment effect configuration, we considered two design sets, $\mathbf{T_{S,L}}$ and $\mathbf{T_{A,L}}$,

comprised of 48 different Bayesian RAR group sequential designs using symmetric or asymmetric posterior probability stopping boundaries, respectively. Each RAR design used:

- The logistic regression probability model to estimate $P_{E>C}$.
- One of eight adaptive modifications in Section 3.2 to obtain randomization probabilities.
- The MWUD to limit deviations from the target allocation.
- $J = 3, 5$, or 10 interim analyses.
- One of two posterior probability stopping boundaries in Section 3.4.

The eight adaptive modifications were chosen due to their familiarity in the literature or desirable statistical properties, though designs using other adaptive sampling methods could have been considered (see, for example, Hu and Rosenberger [59]). More flexible group sequential procedures, such as the Lan-Demets alpha spending function, are available and common in practice. [60] However, we implement Pocock and OBF-like stopping boundaries due to their simplicity and widespread popularity. They also represent the two extremes typically encountered in practice: investigators tend not to use more aggressive boundaries than Pocock, or less aggressive boundaries than OBF. Notably, we consider a design with Pocock-like boundaries and $J = 10$ interim looks, which reflects intensive monitoring with a substantial sample size inflation, to attain an upper bound for n_{max} and a standard against which to assess the performance of more feasible designs.

For symmetric boundaries, designs were calibrated to maintain a two-sided type I error rate, defined as the proportion of simulated trials where $P_{E>C} \geq a_j$ or $P_{E>C} \leq 1 - a_j$ for some j under the hypothesized null scenario, of $\alpha = 0.05$ and a power, defined as the proportion of simulated trials where $P_{E>C} \geq a_j$ or $P_{E>C} \leq 1 - a_j$ for some j under the hypothesized alternative scenario, of $1 - \beta = 0.90$. For asymmetric boundaries, designs were calibrated to maintain a one-sided type I error rate, defined as the proportion of

simulated trials where $P_{E>C} \geq a_s$ for some $j = s$ and $P_{E>C} > b_j \; \forall j = 1, ..., s-1$ under the hypothesized null scenario, of $\alpha = 0.025$ and power, defined as the proportion of simulated trials where $P_{E>C} \geq a_s$ for some $j = s$ and $P_{E>C} > b_j \; \forall j = 1, ..., s-1$ under the hypothesized alternative scenario, of $1 - \beta = 0.90$.

We compared each Bayesian RAR design to analogous frequentist designs using the MWUD to target a 1:1 allocation, increasing the size of $\mathbf{T_{S,L}}$ and $\mathbf{T_{A,L}}$ to 54. For the frequentist designs, we used the gsDesign R package (v3.1.1) [61] to find the maximum sample sizes and stopping boundaries required to achieve the desired error rates. The accrued data at each interim analysis was analyzed using an unconditional exact test via the uncondExact2x2 function in the exact 2x2 R package (version 1.6.5). [62,63] Additional details pertaining to the frequentist designs are provided in Section 5 of Appendix B. We ran all simulations with R version 4.0.2.

3.4.2 Results

Figure 3.1 presents the maximum (N_{Max}) and average (N_{Avg}) trial sample sizes under the targeted alternative for designs with $J = 5$ in $\mathbf{T_{S,L}}$ for the ARREST and ACCESS trials. The size of the potential study sample was defined by the restricted Thompson sampling design with $J = 10$ and Pocock-like stopping boundaries for both ARREST ($n^1_{max} = 179$) and ACCESS ($n^2_{max} = 702$). In the ARREST context, the Bayesian RAR designs generally had larger average sample sizes than the frequentist design and required more participants to achieve 90% power. Conversely, in the ACCESS context, the Bayesian RAR designs using $p_{rsihr,j}$ or ERADE or DBCD to target this optimal allocation ratio engendered as or more favorable sample size distributions than the frequentist design.

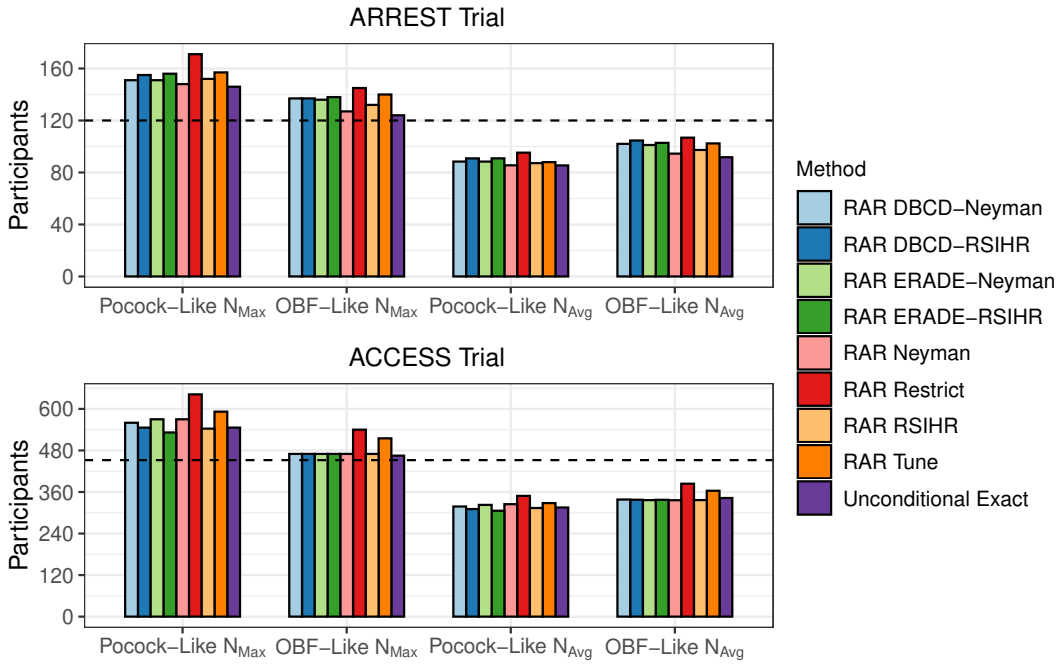


Figure 3.1: Maximum (N_{Max}) and average (N_{Avg}) sample sizes under the targeted alternative for the ARREST and ACCESS trials with $J = 5$. The terms "Pocock-like" and "OBF-like" respectively refer to group sequential designs using symmetric Pocock-like and OBF-like efficacy and harm boundaries. A dashed line is used to denote the sample size required for a fixed sample design with 1:1 allocation. "Unconditional Exact" denotes the frequentist group sequential design using 1:1 allocation. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Figure 3.2 displays the absolute difference between the expected number of trial failures in the group sequential designs with $J = 5$ and in a fixed sample design using 1:1 allocation. The restricted Thompson sampling design engendered the greatest number of trial failures at the targeted alternative and when the observed treatment response rate was less than hypothesized. Although the frequentist design was consistently one of the top performers at the targeted alternative, at least one Bayesian RAR design matched or yielded

fewer trial failures than this design for both ARREST and ACCESS. Differences between designs became negligible when the observed treatment response rate greatly exceeded the hypothesized value.

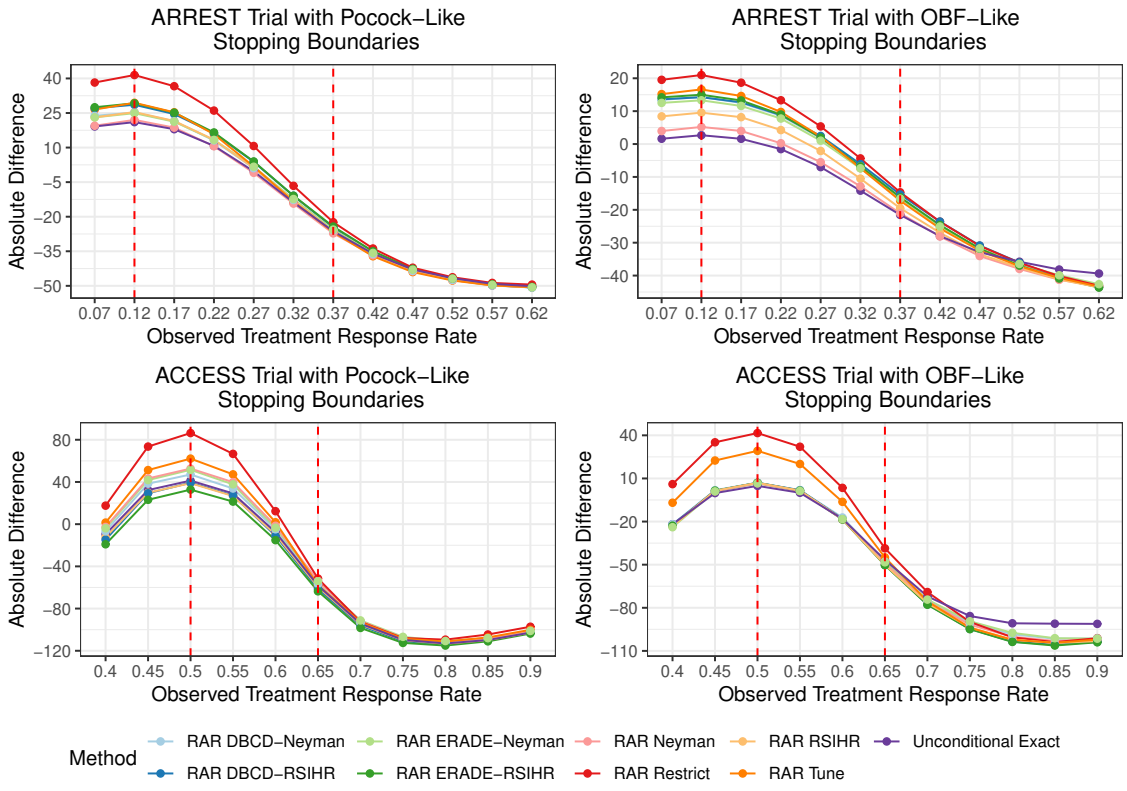


Figure 3.2: Absolute difference in the average number of trial failures relative to a fixed sample design with 1:1 allocation for various observed response rates in the treatment arm. The vertical red lines denote the hypothesized null and alternative response rates for the ARREST and ACCESS trials. These differences correspond to designs using $J = 5$ and symmetric stopping boundaries. "Unconditional Exact" denotes the frequentist group sequential design using 1:1 allocation. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Figure 3.3 contains an analogous figure for our proposed design comparison metric,

m_l, using either n^1_{max} or n^2_{max} as appropriate. In contrast to the average number of trial failures, the frequentist design generally performed the worst with respect to m_l whereas the restricted Thompson sampling design performed the best. For designs implementing Pocock-like stopping boundaries in the ARREST context, the expected reduction in m_l relative to a fixed sample design was 7.9 for the restricted Thompson sampling design and 5.1 for the frequentist design at the targeted alternative, indicating RAR would prevent an additional 2.8 (35%) failures in the potential study sample on average. When the treatment response rate was smaller than hypothesized at 27%, the restricted Thompson sampling design prevented 3.7 (70%) additional failures, on average, relative to 1:1 allocation with group sequential monitoring. When the treatment response rate was larger than hypothesized at 47%, these expected marginal reductions were smaller at 1.7 (14%) additional failures. The potential patient benefit of using RAR over fixed 1:1 allocation appears greatest when the observed treatment effect is slightly smaller than hypothesized and diminishes when the observed treatment effect is greater than hypothesized. Using our proposed decision-theoretic framework from Section 2, the restricted Thompson sampling design with $J = 10$ and Pocock-like stopping boundaries was the optimal design among the set of $\mathbf{T_{S,L}}$ designs under consideration for both ARREST and ACCESS.

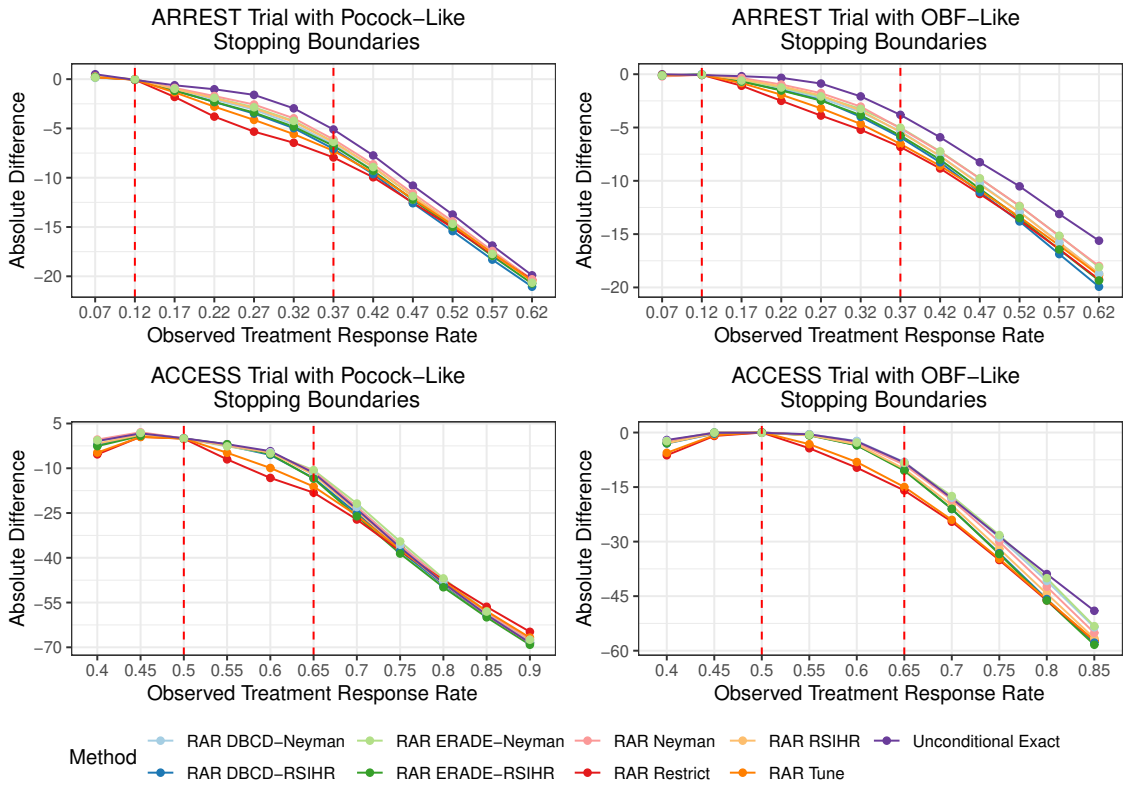


Figure 3.3: Absolute difference in expected failures among potential participants (m_l) relative to a fixed sample design with 1:1 allocation for various observed response rates in the treatment arm. The vertical red lines denote the hypothesized null and alternative response rates for the ARREST and ACCESS trials. These differences correspond to designs using $J = 5$ and symmetric stopping boundaries. "Unconditional Exact" denotes the frequentist group sequential design using 1:1 allocation. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Although the differences in m_l between the Bayesian RAR designs and the frequentist design are modest, these reductions are on par with those achieved by group sequential monitoring. This is evidenced in Table 3.1, which contains the value of m_l for each design in $\mathbf{T}_{\mathbf{S},\mathbf{L}}$ for the ARREST and ACCESS trials. Consider the average number of failures in the

potential study sample prevented by using the restricted Thompson sampling design with $J = 10$ and Pocock-like stopping boundaries over the fixed sample design in the ARREST context: $130 - 120.2 = 9.8$. Because $m_l = 124.0$ for the corresponding frequentist design, we can infer that 6.0, or 61.2%, of the failures prevented by the Bayesian RAR design were prevented due to interim monitoring, whereas 3.8, or 38.8%, were prevented due to the Bayesian RAR procedure itself. Nearly identical percent reductions were observed for the ACCESS trial.

Table 3.1: Expected failures among potential participants (m_l) under the targeted alternative for group sequential designs using symmetric stopping boundaries in the ARREST and ACCESS contexts. "Unconditional Exact" denotes the frequentist group sequential design using 1:1 allocation. For reference, a fixed sample design would have $m_l = 130$ for ARREST and $m_l = 284$ for ACCESS.

Method	Boundaries	J = 3	J = 5	J = 10
ARREST				
RAR DBCD-Neyman	Pocock-like	123.8	122.7	121.6
	OBF-like	124.4	123.9	123.2
RAR DBCD-RSIHR	Pocock-like	123.3	122.2	121.1
	OBF-like	124.0	123.3	122.6
RAR ERADE-Neyman	Pocock-like	124.0	122.9	121.7
	OBF-like	125.1	124.3	123.2
RAR ERADE-RSIHR	Pocock-like	123.8	122.5	121.3
	OBF-like	124.4	123.5	122.8
RAR Neyman	Pocock-like	124.2	123.2	122.0
	OBF-like	125.2	124.2	123.5
RAR Restrict	Pocock-like	122.7	121.3	120.2
	OBF-like	123.5	122.4	121.5
RAR RSIHR	Pocock-like	123.8	122.9	121.8
	OBF-like	124.8	123.8	123.1
RAR Tune	Pocock-like	122.8	122.0	121.4
	OBF-like	123.4	122.7	122.3
Unconditional Exact	Pocock-like	125.0	124.2	124.0
	OBF-like	126.5	125.5	125.0
ACCESS				
RAR DBCD-Neyman	Pocock-like	273.3	272.0	271.0
	OBF-like	277.0	275.3	273.9
RAR DBCD-RSIHR	Pocock-like	272.4	270.1	269.2
	OBF-like	274.7	273.2	271.8
RAR ERADE-Neyman	Pocock-like	274.6	272.7	271.8
	OBF-like	278.8	275.0	273.7
RAR ERADE-RSIHR	Pocock-like	271.4	269.9	269.0
	OBF-like	274.5	272.9	272.1
RAR Neyman	Pocock-like	273.8	272.3	270.4
	OBF-like	276.9	274.5	273.9
RAR Restrict	Pocock-like	268.7	265.2	262.9
	OBF-like	270.5	267.4	265.6
RAR RSIHR	Pocock-like	272.7	271.2	270.1
	OBF-like	275.7	273.4	271.8
RAR Tune	Pocock-like	269.1	267.3	265.6
	OBF-like	269.9	268.4	267.5
Unconditional Exact	Pocock-like	273.4	271.8	270.9
	OBF-like	277.0	275.1	273.9

The findings discussed herein generally held across all scenarios considered, however differences between designs with respect to trial failures and m_l were more apparent for designs targeting low or high null response rates. The restricted Thompson sampling design with $J = 10$ and Pocock-like stopping boundaries was found to be the optimal design among $\mathbf{T_{S,L}}$ and $\mathbf{T_{A,L}}$ for each treatment effect configuration.

3.5 Discussion

This paper proposes a design comparison metric that investigators may use to select a group sequential design that minimizes harm to potential participants, which may be appealing in trials where the primary outcome is mortality or some other outcome with severe implications to quality of life. This metric is simple to implement, may be applied to any set of designs and extended to a multi-arm context, and provides a means to reasonably evaluate the potential patient benefit of response-adaptive randomization against classical frequentist group sequential designs. Our simulations indicate that the restricted Thompson sampling Bayesian RAR design tends to perform the best with respect to our metric across a variety of scenarios, and that Bayesian RAR offers modest reductions in the number of failures in the potential study sample relative to group sequential monitoring alone. We also found that Bayesian RAR group sequential designs exhibit greater marginal gains relative to group sequential designs using 1:1 allocation when the treatment effect is slightly smaller than hypothesized. Conversely, these gains diminish when the treatment effect is much larger than hypothesized, likely because both designs will stop very early with high probability which precludes adaptations to the allocation ratio. Whether these gains are worthwhile needs to be assessed within the context of each trial based on the severity of the primary outcome, the plausibility of observing various effect sizes, and the increased complexity that comes with Bayesian RAR implementation.

This research has several limitations. First, due to the motivating context for this work, we only focused on two-arm trials with a binary primary outcome and a select set

of Bayesian RAR procedures. Our conclusions regarding the potential patient gains arising from RAR with respect to our proposed metric may vary in different design settings or when alternative RAR procedures are considered. Next, we did not account for the presence of time trends including patient drift, which has been shown to inflate type I error rate in response-adaptive designs. [6] We also made the working assumption that individuals in the potential study sample who did not participate in the trial would have received the treatment recommended by the trial. While this is a reasonable assumption for the motivating ACCESS and ARREST trials, this assumption may deviate from practice when the recommended treatment cannot be readily applied to the entire target population due to regulatory requirements and time required for dissemination and uptake of trial results. Yet, this assumption provides a reasonable way to reward a design that stops early to recommend treatment when it is in fact efficacious and enables our proposed metric to facilitate a fairer design comparison than existing metrics based on the realized trial sample size which often fail to 1) compare designs with equal type I and II error rates or 2) assess the impact of the design with respect to relevant, equal-sized populations. We also did not consider multi-arm trials in which Bayesian RAR has been shown to increase efficiency relative to balanced randomization. An interesting area for future work would be to study our metric in the context of platform trials or multi-arm multi-stage (MAMS) designs (see Lin and Bunn [64] and Watson and Trippa [65]). Finally, although we declare the optimal group sequential design to be the one that minimizes the average number of failures in the potential study sample among the designs considered, we did not fully optimize this criterion. Future work, including the evaluation of frequentist group sequential designs whose characteristics have been optimized such that the expected number of failures in the potential study sample is minimal, is needed to identify the design that fully minimizes the proposed metric.

Chapter 4

Network Meta Analysis to Predict the Efficacy of an Approved Treatment in a New Indication

4.1 Introduction

Drug repurposing (also known as drug repositioning or drug reprofiling) is the process of evaluating FDA-approved treatments in an additional clinical indication. This process occurs when drug developers hypothesize that another indication will respond similarly to a licensed drug due to biological similarities or comparable mechanisms of action, and can advantageously be used to satisfy unmet clinical needs or maximize a treatment's therapeutic ability. [66, 67] It also involves studying treatments with well established safety profiles, which can dramatically shorten the drug development timeline and get treatments to new patient populations quicker. [11] Two indications for which drug re-purposing has been successful are psoriasis and psoriatic arthritis. These indications are psoriatic diseases arising from malfunctioning immune responses that lead to inflammation and may be successfully treated with several of the same treatments (e.g., Stelera and Cosyntyx). [68, 69]

The decision to pursue supplemental indications is costly, challenging, and complex. There is a lack of methods that can reliably quantify whether a repurposed molecule is likely to succeed in a future trial, making it difficult for drug developers, particularly those with extensive drug pipelines that may have multiple candidate drugs for the same indication, to determine which compound to prioritize next. To inform this decision, companies may use existing approaches to drug repurposing including experimental screening, the analysis of large-scale omics data, and the use of real-world data to identify novel associations between treatment and disease. [70,71] However, these methods are time-consuming, involve complex and computationally intensive analytical methods (e.g., artificial intelligence and deep learning), and are not widely used. Repurposing decisions are instead often based on human judgement and limited empirical evidence, which can lead to inefficient drug development, waste valuable resources, and increase the time required for successful drugs to get to market. [12]

A well-established tool that is often used to facilitate informed decision-making in the healthcare industry is meta-analysis. Meta-analysis (MA) is a statistical method that combines data from multiple studies to reduce bias, improve efficiency, and identify patterns and sources of disagreement in treatment effect estimates. By incorporating a larger evidence base, MA increases the statistical power of a treatment comparison and offers more definitive answers to scientific questions compared to individual trials, which may exhibit conflicting evidence. [72, 73] Traditional methods for MA are limited to pairwise comparisons and may not be applied to contexts where the treatments under evaluation differ across trials. Network meta-analysis (NMA) was developed to simultaneously compare multiple treatments and strengthen inference through the incorporation of both direct and indirect evidence, with the latter facilitating the estimation of relative effects for treatments that were never compared in a single study. [19–21] Consider a network comprised of two-arm trials comparing treatments A vs. B or B vs. C. In contrast to standard MA, NMA yields an estimate of the population-averaged effect for the unstudied A vs. C comparison through

the common comparator, B. [74, 75]

Although traditionally applied to data from late-stage studies for comparative effectiveness research or regulatory submissions, NMA is useful in different stages of drug development. Mawdsley et al. [22] developed model-based NMA (MBNMA) by incorporating dose-response models into an NMA framework. Their method uses plausible physiological dose-response models to predict the effect of multiple treatments across a range of doses. Pedder et al. [23] expanded MBNMA to model the time-course relationship of multiple treatments using a continuous function. This method can similarly predict treatment efficacy at unstudied time points and enhance the understanding of pharmacodynamic profiles of new compounds. Importantly, predictions from both methods can be used to inform the design of future clinical trials, evaluate the competitive landscape, and accelerate compound development by mitigating trial failures. [24] Bujkiewicz et al. [76] have also applied bivariate NMA (bvNMA) to surrogate endpoint evaluation. They use a bvNMA model that allows the association between effects on surrogate endpoints and final outcomes to differ by treatment and show how this methodology can be used to predict the effect of a new treatment on a final outcome given its effect on the surrogate endpoint.

Despite these applications, NMA has not been used to predict the efficacy of an approved treatment in a new indication and thereby identify repurposable treatments. For this reason, we develop a novel Bayesian NMA framework that can assist drug developers in answering the commercial decision of "what to study next" in the context of drug repurposing. Our proposed method predicts whether a treatment is likely to succeed in a new indication based on data from other treatments studied in both the new and approved indications, and is applicable to situations where at least 3 drugs (e.g, placebo and two active treatments) have been studied in two indications that can reasonably be assumed to be related. We obtain predictions using two main steps: first, we use traditional NMA methods to estimate the population-averaged relative effects from a network comprised of treatments studied in both indications in addition to one treatment studied in only one indication. Then, we model

the correlation between relative effects across indications to predict how a treatment would perform if studied in the additional indication (i.e. if repurposed). The goal of our work is somewhat analogous to that of Bujkiewicz et al., [76] however here we consider a treatment that has yet to be studied in one indication and model the correlation between treatments across indications.

This paper is organized as follows. In section 2, we discuss the standard Bayesian random effects NMA model for one indication. In section 3, we introduce our proposed NMA framework for synthesizing data from two related indications and predicting treatment efficacy in an unstudied indication. We consider nine different models that differ in how they model treatments across indications and within the same drug class. In section 4, we report an extensive simulation study comparing models with respect to their predictive performance under a variety of scenarios. In section 5, we demonstrate how the proposed methodology can be applied in practice using a case study in psoriasis and psoriatic arthritis. We conclude by discussing our findings and directions for future research.

4.2 Bayesian Network Meta-Analysis for One Indication

4.2.1 Background

Our use of the Bayesian paradigm for NMA provides several advantages. First, prior distributions can be used to keep inference plausible in scenarios where there are weakly identified parameters, such as in NMAs with a limited number of studies. Second, Bayesian hierarchical modeling allows us to investigate between-study heterogeneity using the posterior distribution of the variance components. This assessment is critical in establishing the validity of an NMA, since between-study differences may engender misleading treatment comparisons. [19, 74] Finally, the sampling-based framework of Bayesian inference readily produces the posterior predictive distribution of the treatment effect of interest. This is particularly important for drug repurposing, where investigators are often interested in learning

their likelihood of success in a future clinical trial.

The models we introduce use a contrast-based (CB) approach where relative treatment effects (e.g., log-odds ratios for binary outcomes or log-hazard ratios for time-to-event endpoints) are modeled across studies. Although the conditional OR for binary outcomes is a non-collapsible effect measure and is not "portable" in meta-analysis, the CB approach preserves within-trial randomization and generates relatively stable estimates. [77–79] In contrast, arm-based NMA (AB-NMA) combines absolute effects for treatment-specific arms (e.g., log odds for binary outcomes or log hazard for time-to-event endpoints). Although AB-NMA offers several benefits, including robustness to treatment exclusions and the estimation of marginal treatment effects (see, for example, Lin et al., [80] Wang et al., [81] and White et al. [82]), this approach may have some drawbacks as well, including breaking within-trial randomization and engendering biased treatment effect estimates when the assumption of exchangeable treatment arms is inappropriate or there is insufficient information to estimate the correlation parameters among treatments within studies. For these reasons, we proceed with the CB-NMA framework of Lu and Ades [74, 75] described below in Section 2.2.

4.2.2 The Standard Contrast-Based Model

Suppose we have binary outcome data from J RCTs evaluating the efficacy of a total of K different treatments in a given clinical indication (e.g., psoriasis). Let T_j be the subset of treatments compared in trial j such that data collected from each trial may be expressed as $\boldsymbol{D_J} = \{(r_{jk}, n_{jk}), k \in T_j, j = 1, ..., J\}$, where r_{jk} and n_{jk} respectively denote the number of successes and participants in the k^{th} arm of the j^{th} trial. Assuming we have a fully connected network, i.e. every treatment pairing has been evaluated directly or indirectly through a common comparator, we may analyze these data using the following Bayesian

random effects NMA (RE-NMA) model: [74, 75]

$$
\begin{aligned}
&r_{jk} \sim \text{Bin}(n_{jk}, p_{jk}), \\
&\text{logit}(p_{jk}) = \mu_j + \delta_{jb_jk} \cdot \text{I}(k \neq b_j), \text{ and} \\
&\delta_{jb_jk} \sim \text{N}(d_{b_jk}, \sigma^2_{b_jk}) \text{ for } k \in T_j \text{ and } j = 1, ..., J
\end{aligned}
\tag{4.1}
$$

where p_{jk} is the probability of response in arm k of study j, μ_j is the log-odds of response for the baseline treatment b_j in study j, δ_{jb_jk} is the study-specific log-odds ratio comparing treatment k to b_j, and d_{b_jk} is the pooled log-odds ratio comparing treatment k to b_j. Letting $k = 1$ denote the network reference treatment, we assume that $d_{b_jk} = d_{1k} - d_{1b_j} \forall k \in T_j$ and $j = 1, ..., J$; that is, we assume the pooled log-odds ratios follow the first-order consistency equations of Lu and Ades [74] which allows us to estimate relative treatment effects for all possible pairwise comparisons.

Using a fully Bayesian approach, we specify prior distributions for all unknown parameters. Let $t_\nu(\mu, \sigma)$ denote a generalized t-distribution with ν, μ, and σ representing the degrees of freedom, location, and scale parameters, respectively. After constraining $d_{11} = 0$, we follow the work of Ghosh et al. [83] and Gelman et al. [84] for prior specification in logistic regression modeling and specify $\mu_j \sim t_7(0, 2.5)$ for $j = 1, ...J$ and $d_{1k} \sim t_7(0, 2.5)$ for $k = 2, ..., K$. When estimating μ_j is of interest, a hierarchical model can be placed on the baseline effects corresponding to $k = 1$. [73] Instead, we treat the baseline effects as nuisance parameters since relative effect estimates can become biased when the baseline effect model is misspecified. [77]

As shown in the third line of equation (4.1), the RE-NMA model assumes that the study-specific relative effects are exchangeable, i.e. that the δ_{jb_jk} come from a common normal distribution with mean d_{b_jk} and variance $\sigma^2_{b_jk}$. In contrast to a fixed effects model, this approach considers between-study heterogeneity that may arise due to differences in patient populations, treatment regimens, eligibility criteria, or other reasons that are less easy to articulate or measure. [85] We make the simplifying assumption that $\sigma^2_{b_jk} = \sigma^2$

$\forall k \in T_j, j = 1, ..., J$, which will increase precision by borrowing strength across treatment comparisons when data is sparse. [86] It also simplifies incorporating data from multi-arm studies by ensuring that the correlation between any two study-specific relative effects arising from the same trial is 0.5. [87] We consider four different prior distributions for σ below.

4.2.3 Prior Distributions for the Between-Study Heterogeneity Parameter

Because posterior inference is sensitive to the choice of prior distribution for variance parameters in Bayesian hierarchical models, we evaluate 4 different prior distributions for the between-study standard deviation, σ: Uniform(0,5) (U(0,5)), Half-$t_7(0, 2.5)$ ($Ht_7(0, 2.5)$), standard Half-Normal (HN(0,1)), and Log-Normal(-2.70,1.52) (LN(-2.70,1.52)). [88,89] The U(0,5) prior reflects the belief that the true value for σ is equally likely to take on any value between 0 and 5 and places 95% of its probability mass between 0.12 and 4.88. Although widely used throughout the NMA literature, this prior is arguably inappropriate in a binary outcomes context where values greater than 1 are considered extremely unlikely. [90] For example, consider the 95% confidence interval (CI) for d_{b_jk} arising from a normal approximation for the log-odds ratio: $d_{b_jk} \pm 1.96\sigma$. When $\sigma = 1$, the ratio of the upper to lower CI bounds is $\exp(3.92\sigma) = 50.4$. In scenarios where the odds ratio is unlikely to vary by such an extreme amount, it may be desirable to use a prior that places more mass near 0. [90]

Following the recommendations of Gelman, [88] we consider a weakly informative $Ht_7(0, 2.5)$ prior distribution. This prior has a 95% credible interval of (0.08,7.11) and is generally preferable when the number of groups in the hierarchical model is low, in which case the Uniform distribution tends to overestimate σ, particularly when σ is small. [88] The LN(-2.70,1.52) prior is an empirically-derived informative prior distribution from Turner et al. [85] with a 95% credible interval of (0,1.33). This empirical prior was derived by simultaneously modeling data from a large number of binary outcome MAs and deriving predictive distributions for σ based on outcome and intervention comparison type. The LN(-2.70,1.52)

prior corresponds to subjective outcomes and pharmacological vs. placebo/control comparisons, which are commonly encountered in practice. [85] This prior will be most useful for NMAs conducted with a limited number studies for which it is difficult to precisely estimate the heterogeneity parameter. [89, 91] Finally, the HN(0,1) distribution represents an intermediate choice between the $\text{Ht}_7(0, 2.5)$ and LN(-2.70,1.52) priors with respect to informativeness; it has a mean of 0.8 and places 95% of its mass between 0.06 and 1.96. A plot of the 4 considered priors for σ in addition to the induced priors for variance (σ^2) and precision ($1/\sigma^2$) is provided in Appendix Figure C.1.

4.3 Bayesian Network Meta-Analysis for Two Indications

4.3.1 Background

In this section, we present an NMA framework that can synthesize data from two related indications. We assume we have data from multiple studies evaluating the same treatments in both indications as well as data for one treatment that has been approved and studied in only one indication. Our goal is to leverage the available data to predict whether the treatment approved in one indication is likely to be efficacious relative to a standard reference treatment (e.g., placebo) in the other indication. Below we present three different models that use an NMA framework to predict treatment efficacy in the new indication. These methods are applicable to scenarios where at least 3 treatments, and hence 2 treatment contrasts, have been studied in both indications which we herein refer to as overlapping treatments. This requirement allows our proposed models to estimate the correlation between the basic parameters, d_{1k}, in each indication, thus enabling predictions for how a treatment studied in one indication would perform if studied in the other.

4.3.2 Notation

Suppose we have binary outcome data from J_i studies in indication i, $i = 1, 2$, such that $J_1 + J_2 = J$. Let S_i denote the set of treatments studied in indication i such that $|S_1 \cup S_2| = K$ and $|S_1 \cap S_2| < K \geq 3$. That is, our data consists of K unique treatments where at least three of these treatments have been evaluated in both indications (e.g., $S_1 = \{1, 2, 3, 4, 5, 6\}$ and $S_2 = \{1, 2, 3, 4, 5\}$). Although the methods described herein are suitable for any network where $|S_1 \cap S_2| < K \geq 3$, we focus on networks with $|S_1 \cap S_2| = K - 1 \geq 3$ for simplicity. Then, we use the standard CB-NMA model from equation (4.1) now assuming:

$$\delta_{jb_jk} \sim \mathrm{N}(d^i_{b_jk}, \sigma_i^2) \; \forall k \in T_j \text{ and } j = 1, ..., J \tag{4.2}$$

where $i \in \{1, 2\}$ is the indication corresponding to trial j. The model parameters in (4.2) are interpreted as in (4.1) apart from $d^i_{b_jk}$, which is the pooled log-odds ratio comparing treatments k and b_j in indication i, and σ_i^2, which is the between-study heterogeneity parameter for indication i. Note that we treat the random effects δ_{jb_jk} as exchangeable across studies only within the same indication. Similar to before, we constrain $d^1_{11} = d^2_{11} = 0$ and follow the first-order consistency equations of Lu and Ades [74] by assuming $d^i_{b_jk} = d^i_{1k} - d^i_{1b_j}$ for $k = 2, ..., K$ and $i = 1, 2$. This assumption allows us to estimate all possible pairwise comparisons for each indication while only modeling the indication-specific basic parameters, d^i_{1k}.

As opposed to the standard RE-NMA model from section 2, the d^i_{1k} parameters are only directly estimable when all treatments of interest are overlapping, i.e $|S_1 \cap S_2| = K$. Because we are instead interested in the scenario where $|S_1 \cap S_2| = K - 1$, we do not apply independent and identically distributed priors to d^i_{1k} in order to avoid obtaining a posterior for d^i_{1k} that matches the prior. For example, suppose $S_1 = \{1, 2, 3, 4, 5\}$ and $S_2 = \{1, 2, 3, 4\}$. In this scenario, assuming $d^i_{1k} \sim t_7(0, 2.5)$ for $k = 2, ..., 5$ and $i = 1, 2$ would engender a meaningless estimate for d^2_{15} since there is no available data from which to estimate this

parameter. To overcome this limitation, we consider various models for each (d^1_{1k}, d^2_{1k}) pair that borrow strength across indications to enable informed predictions for the treatment that has yet to be studied in the second indication.

4.3.3 Proposed Models

The Bivariate Normal Approach

Our first model uses a bivariate normal distribution to model the correlation between the basic parameters across indications. This approach directly extends the standard RE-NMA model of Lu and Ades [74] to a two indication setting by assuming each (d^1_{1k}, d^2_{1k}) pair are exchangeable and follow a common bivariate normal distribution as follows:

$$\begin{pmatrix} d^1_{1k} \\ d^2_{1k} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_{1k}(\boldsymbol{\beta}) \\ \mu_{2k}(\boldsymbol{\beta}) \end{pmatrix}, \boldsymbol{\Sigma} \right), \text{with } \boldsymbol{\Sigma} = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \text{ for } k = 2, ..., K \text{ and } i = 1, 2 \tag{4.3}$$

where ρ is the correlation between d^1_{1k} and d^2_{1k}, τ_i^2 is the variance of d^i_{1k}, and $\mu_{ik}(\boldsymbol{\beta})$ is the mean of d^i_{1k} and a function of regression parameters. We assume $\tau_i^2 \sim$ HN(0,1) and $\rho \sim$ Uniform$(0, 1)$, with the latter reflecting the prior belief that treatments are likely to behave similarly relative to each other across indications, and consider three different structures for $\mu_{ik}(\boldsymbol{\beta})$ which we will discuss in section 3.3.4. Notably, this model assumes $\text{Corr}(d^1_{1k}, d^2_{1k}) = \rho$, $\text{Corr}(d^i_{1k}, d^i_{1l}) = 0$, and $\text{Corr}(d^1_{1k}, d^2_{1l}) = 0 \ \forall k \neq l \in \{2, ..., K\}$ and $i = 1, 2$. Namely, that only basic parameters corresponding to the same treatment but different indications are correlated. Letting K be the treatment that has yet to be studied in the second indication, we obtain a posterior distribution for d^2_{1K} by fitting this model using MCMC methods, which automatically imputes d^2_{1K} due to the specification of the bivariate normal prior in (4.3).

The Multivariate Normal Approach

It may also be argued that the vector of parameters $(d^i_{12}, d^i_{13}, ..., d^i_{1K})^T$ for $i = 1, 2$ are correlated since they are relative to the same network reference treatment, $k = 1$, in which case it would be inappropriate to assume that each (d^1_{1k}, d^2_{1k}) pair are exchangeable. Under this assumption, our second model follows the approach of Bujkiewicz et al. [76] and first decomposes the pooled contrast-level effects into their respective arm-level effects as follows:

$$d^i_{1k} = \alpha^i_k - \alpha^i_1 \tag{4.4}$$

where α^i_k is the log-odds of response in treatment arm k for indication i. Because $\alpha^i_k \perp\!\!\!\perp \alpha^i_l \; \forall k \neq l \in \{2, ..., K\}, i = 1, 2.$ and thus an assumption of exchangeability is now reasonable, we model each (α^1_k, α^2_k) pair using a common bivariate normal distribution:

$$\begin{pmatrix} \alpha^1_k \\ \alpha^2_k \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \phi_{1k} \\ \phi_{2k} \end{pmatrix}, \frac{\boldsymbol{\Sigma}}{2} \right) \text{ for } k = 2, ..., K \tag{4.5}$$

where ϕ_{ik} is the mean of α^i_k and Σ is the covariance matrix from (4.3). Notably, the α^i_k parameters are unidentifiable because we are modeling our data using a contrast-based approach. However, the advantage of the decomposition in (4.4) is that it, together with the joint prior distribution in (4.5), induces a $2(K-1)$ by $2(K-1)$ multivariate normal distribution for $\boldsymbol{d_1} = (\boldsymbol{d_{12}}, \boldsymbol{d_{13}}, ..., \boldsymbol{d_{1K}})^T$ with:

$$\begin{aligned} &\mathrm{E}[\boldsymbol{d_{1k}}] = \boldsymbol{\mu_{1k}}(\boldsymbol{\beta}), \mathrm{Var}(\boldsymbol{d_{1k}}) = \boldsymbol{\Sigma} \text{ for } k = 2, ..., K \text{ and} \\ &\mathrm{Cov}(\boldsymbol{d_{1k}}, \boldsymbol{d_{1l}}) = \begin{pmatrix} \frac{\tau_1^2}{2} & \frac{\rho}{2}\tau_1\tau_2 \\ \frac{\rho}{2}\tau_1\tau_2 & \frac{\tau_2^2}{2} \end{pmatrix} \text{ for } l \neq k = 2, ..., K \end{aligned} \tag{4.6}$$

where $\boldsymbol{d_{1k}} = (d^1_{1k}, d^2_{1k})^T$, $\boldsymbol{\mu_{1k}} = (\mu_{1k}(\boldsymbol{\beta}), \mu_{2k}(\boldsymbol{\beta}))^T$, and each basic parameter is correlated, as desired. Specifically, (4.6) suggests that $\mathrm{Corr}(d^1_{1k}, d^2_{1k}) = \rho$, $\mathrm{Corr}(d^i_{1l}, d^i_{1k}) = 1/2$, and

$\text{Corr}(d^1_{1l}, d^2_{1k}) = \rho/2\ \forall k \neq l \in \{2, ..., K\}$ and $i = 1, 2$. Rather than treating $\text{Corr}(d^i_{1l}, d^i_{1k})$ as an additional unknown parameter to be modeled, we use $\text{Corr}(d^i_{1l}, d^i_{1k}) = 1/2$ because both d^i_{1l} and d^i_{1k} are relative to α^i_1 and $\text{Var}(\alpha^i_l) = \text{Var}(\alpha^i_k) = \tau^2_i/2$. Similar to the bivariate normal approach, we specify $\tau_1, \tau_2 \sim \text{HN}(0,1)$, $\rho \sim \text{Uniform}(0, 1)$, and predict the posterior distribution for d^2_{1K} using MCMC imputation.

The Mixed Effects Approach

The modeling approaches from Sections 3.3.1 and 3.3.2 directly model the correlation between basic parameters across indications by specifying a prior distribution for ρ. In contrast, our final approach indirectly models the correlation between each (d^1_{1k},d^2_{1k}) pair by introducing random effects into the linear predictor for d^i_{1k} as follows:

$$\begin{aligned} d^i_{1k} &= \mu_{ik}(\boldsymbol{\beta}) + \gamma_k + \epsilon_{ki}, \\ \gamma_k &\overset{iid}{\sim} \text{N}(0, \sigma^2_\gamma), \text{ and} \\ \epsilon_{ki} &\overset{iid}{\sim} \text{N}(0, \sigma^2_\epsilon) \end{aligned} \tag{4.7}$$

for $k = 2, ..., K$ and $i = 1, 2$. This formulation implies that the correlation between d^1_{1k} and d^2_{1k} is $\sigma^2_\gamma/(\sigma^2_\gamma + \sigma^2_\epsilon)$, where σ^2_γ is the variance of a treatment-specific random effect, γ_k and σ^2_ϵ is the variance of a treatment- and indication-specific random error term, ϵ_{ki}. Notably, as $\sigma^2_\epsilon \to 0$, $\text{Corr}(d^1_{1k},d^2_{1k}) \to 1$ and as $\sigma^2_\epsilon \to \infty$, $\text{Corr}(d^1_{1k},d^2_{1k}) \to 0$. Therefore, the indication-specific basic parameters are highly correlated when the random error variance is small relative to σ^2_γ and weakly correlated when the random error variance is large, and it is likely that the predictive performance of each proposed model will improve as σ^2_ϵ approaches 0. Although this model requires $\text{Corr}(d^1_{1k},d^2_{1k}) > 0$ and is thus less flexible than the previous two approaches that allow the use of any reasonable prior for ρ, this approach may be more appealing to those who prefer mixed effects modeling. It also provides an intuitive framework for inducing correlation between additional basic parameters by adding random effects to the linear predictor for d^i_{1k}. After specifying $\sigma_\gamma, \sigma_\epsilon \sim \text{HN}(0, 1)$, we use

MCMC methods to facilitate posterior sampling and obtain a posterior distribution for d^2_{1K} by adding together the MCMC draws for $\mu_{2K}(\boldsymbol{\beta})$, γ_K, and ϵ_{K2}.

Incorporating Indication and Drug Class Effects

For each model, we evaluate three different structures for $\mu_{ik}(\boldsymbol{\beta}) = \mathrm{E}[d^i_{1k}]$ that are functions of indication and/or drug class. We first allow $\mu_{ik}(\boldsymbol{\beta})$ to vary by indication since treatments $k = 2, ..., K$ may perform systematically better in one indication than the other. We then consider two expressions for $\mu_{ik}(\boldsymbol{\beta})$ that assume treatments within the same drug class have a similar effect on the endpoint of interest. This additional assumption may increase precision though the borrowing of information in networks with a limited number of studies but many treatments under consideration. [92,93] It may also improve predictions for missing treatments effects by allowing the relationship between indications to differ by class. We specifically evaluate the following three mean structures:

$$\begin{aligned}
&1)\ \mu_{ik}(\boldsymbol{\beta}) = \beta_0 + \beta_1(i - 1.5), \\
&2)\ \mu_{ik}(\boldsymbol{\beta}) = \beta_0 + \beta_1(i - 1.5) + \beta_2(c - 1.5), \text{ and} \\
&3)\ \mu_{ik}(\boldsymbol{\beta}) = \beta_0 + \beta_1(i - 1.5) + \beta_2(c - 1.5) + \beta_3\{(i - 1)(c - 1) - 0.5\}
\end{aligned} \tag{4.8}$$

where $c \in \{1, ..., C\}$ is the drug class corresponding to treatment k. We use contrast coding to reduce the autocorrelation between posterior samples and assume $C = 2$ for simplicity; however, the expressions in (4.8) may be straightforwardly modified to accommodate additional classes. In expression 1), $\mu_{ik}(\boldsymbol{\beta})$ is a function of indication only where β_1 reflects the expected difference in d^2_{1k} and d^1_{1k} for $k = 2, ..., K$. In expression 2), $\mu_{ik}(\boldsymbol{\beta})$ is a function of both indication and drug class where β_2 reflects the expected difference in basic parameters from the same indication but different drug classes. In expression 3), $\mu_{ik}(\boldsymbol{\beta})$ is still a function of both indication and drug class, however now the drug class effect differs by indication. Here, β_2 is the drug class effect for indication 1 and $\beta_2 + \beta_3$ is the drug class effect for indication 2.

4.4 Simulation Study

4.4.1 Design

We performed a simulation study to assess the ability of each proposed model to predict d^2_{1t} under a variety of settings, where $k = t$ is the candidate treatment for drug repurposing. We generated data according to the network diagram in Figure 4.1, which consists of $J_1 = 40$ two-arm studies in indication 1, $J_2 = 35$ studies in indication 2, and $K = 9$ overlapping treatments. Letting $k = 1$ denote placebo, we generated data from a vector of basic parameters, $\boldsymbol{d} = (d^1_{12}, d^1_{13}, ..., d^1_{19}, d^2_{12}, d^2_{13}, ..., d^2_{19})^T$ by solving for $d^i_{b_jk}$ using the consistency equations of Lu and Ades. [74, 75] We then sampled $\delta_{jb_jk} \sim \mathrm{N}(d^i_{b_jk}, 0.25^2)$ and simulated r_{jk} using the first two lines of equation (4.1) $\forall k \in T_j, j = 1, ..., J$. We fixed n_{jk} at 60 and 38 for indications 1 and 2, respectively, which provides 90% power to detect log-odds ratios of 2 in each indication. Additional details pertaining to data generation, including how we simulated values for the baseline parameters, μ_j, are provided in Section C.2 of the Appendix.

Instead of selecting one $\boldsymbol{d}$ to use in this simulation study, we randomly generated different values for $\boldsymbol{d}$ using the mixed effects model with $\mu_{ik}(\boldsymbol{\beta}) = \beta_0 + \beta_1(i-1) + \beta_2(c-1)$ in order to compare model performance across a wide range of treatment effects. We considered three sets of regression coefficients: $\boldsymbol{\beta_1} = \boldsymbol{\beta_2} = (3.4, 1.3, -1.6)$, $\boldsymbol{\beta_1} = \boldsymbol{\beta_2} = (3.4, 0.325, -0.4)$, and $\boldsymbol{\beta_1} = (3.4, 1.3, -1.6)$ but $\boldsymbol{\beta_2} = (3.4, 0, -1.6)$, where $\boldsymbol{\beta_i} = (\beta_0, \beta_1, \beta_2)$ denotes the regression coefficients used for indication i. In comparison to set 1, set 2 generates data using smaller indication and drug class effects and set 3 employs a heterogeneous class effect where treatments in different drug classes differ, on average, in indication 1 but not in indication 2. For each set of regression coefficients, we examined five different pairs of random effect variances, $(\sigma^2_\gamma, \sigma^2_\epsilon)$, that keep the total variation constant at 0.50: (0.05,0.45), (0.15,0.35), (0.25,0.25), (0.35,0.15), and (0.45,0.05). These pairs correspond to $\mathrm{Cor}(d^1_{1k}, d^1_{2k}) = 0.10$, 0.30, 0.50, 0.70, and 0.90, respectively, since the correlation between d^1_{1k} and d^1_{2k} in the

mixed effects model is $\sigma_\gamma^2/(\sigma_\gamma^2 + \sigma_\epsilon^2)$. We randomly generated 50 $\boldsymbol{d}$'s under each of these settings.

Using each $\boldsymbol{d}$, we generated 200 data sets excluding treatment t from the indication 2 network; that is, prior to generating data, we removed all studies evaluating treatment t in indication 2 from Figure C.1. We considered $t = 2, 4, 6, 9$ in order to assess the impact of the amount of available information on treatment t in indication 1 on the posterior distribution for d_{1t}^2. Note that treatment $k = 2$ was heavily studied in indication 1 (13 times), treatments $k = 4$ and 6 were moderately studied in indication 1 (7-8 times), and treatment $k = 9$ was sparsely studied in indication 1 (2 times). For each data set, we predicted d_{1t}^2 using the bivariate normal, multivariate normal, and mixed effects models with the mean structures in (4.8) and $\sigma \sim \text{U}(0, 1), \text{Ht}_7(0.2, 5), \text{HN}(0, 1)$, and LN(-2.70,1.52). We implemented each model using JAGS via the R package rjags (version 4.13), and approximated each posterior distribution using 2 chains run for 10,000 iterations after a burn-in of 5,000. [94] Posterior convergence was monitored using the Gelman-Rubin convergence diagnostic with values less than 1.1 considered adequate. [95] We used the posterior median and 95% equal-tailed credible intervals as point and interval estimates, respectively, and compared model performance using root mean squared error (RMSE), average credible interval width, and coverage probability. We ran all simulations with R version 4.1.0.

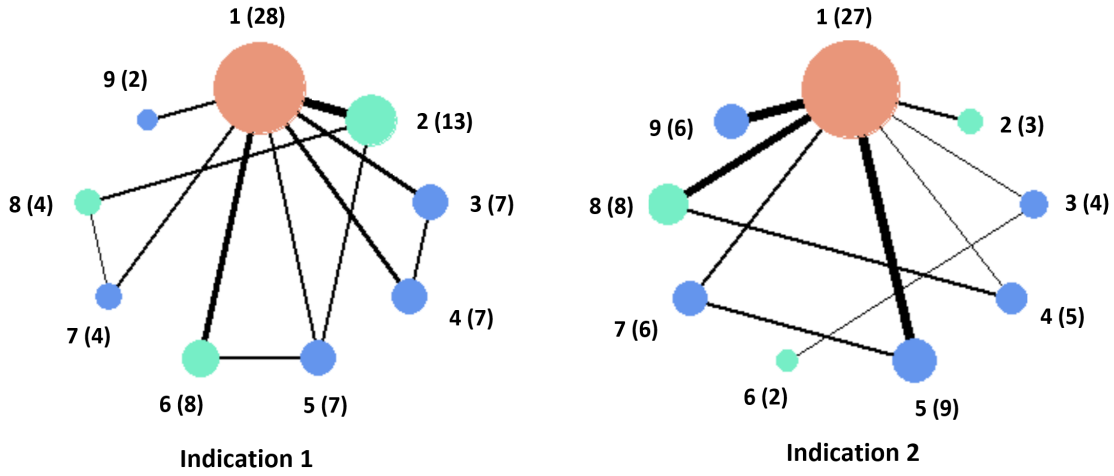


Figure 4.1: Network plots for indications 1 ($J_1 = 40$) and 2 ($J_2 = 35$) with 9 overlapping treatments. Each node represents a treatment and each edge represents a direct comparison between two treatments. The size of each node is proportional to the number of studies evaluating that treatment whereas the width of each edge is proportional to the number of direct comparisons. The network reference treatment is placebo ($k = 1$) and treatments 2-9 are active agents. Nodes that are green correspond to drug class 1 and nodes that are blue correspond to drug class 2.

4.4.2 Results

Figure 4.2 displays the RMSE of the posterior median of d^2_{1t} arising from $\boldsymbol{d}$'s simulated using coefficient set 1. All models were fit using the $\text{Ht}_7(0, 2.5)$ prior for the between-study heterogeneity parameter and results are presented by treatment and the correlation between d^1_{1k} and d^2_{1k}, $\text{Cor}(d^1_{1k}, d^2_{1k})$. As expected, the predictive performance of each model tended to improve as the correlation increased; that is, RMSE decreased, on average, as $\text{Cor}(d^1_{1k},d^2_{1k})$ approached 0.9. For each class of models, RMSE was generally smallest for models incorporating a homogeneous drug class effect and largest for models incorporating a heterogeneous drug class effect. This is unsurprising given that $\boldsymbol{d}$ was generated assuming a homogeneous

drug class effect, and therefore models assuming a similar structure exhibited reductions in both bias and variance. In contrast, reductions in bias observed for models assuming a heterogeneous drug class effect were often offset by increases in variance associated with increasing model complexity.

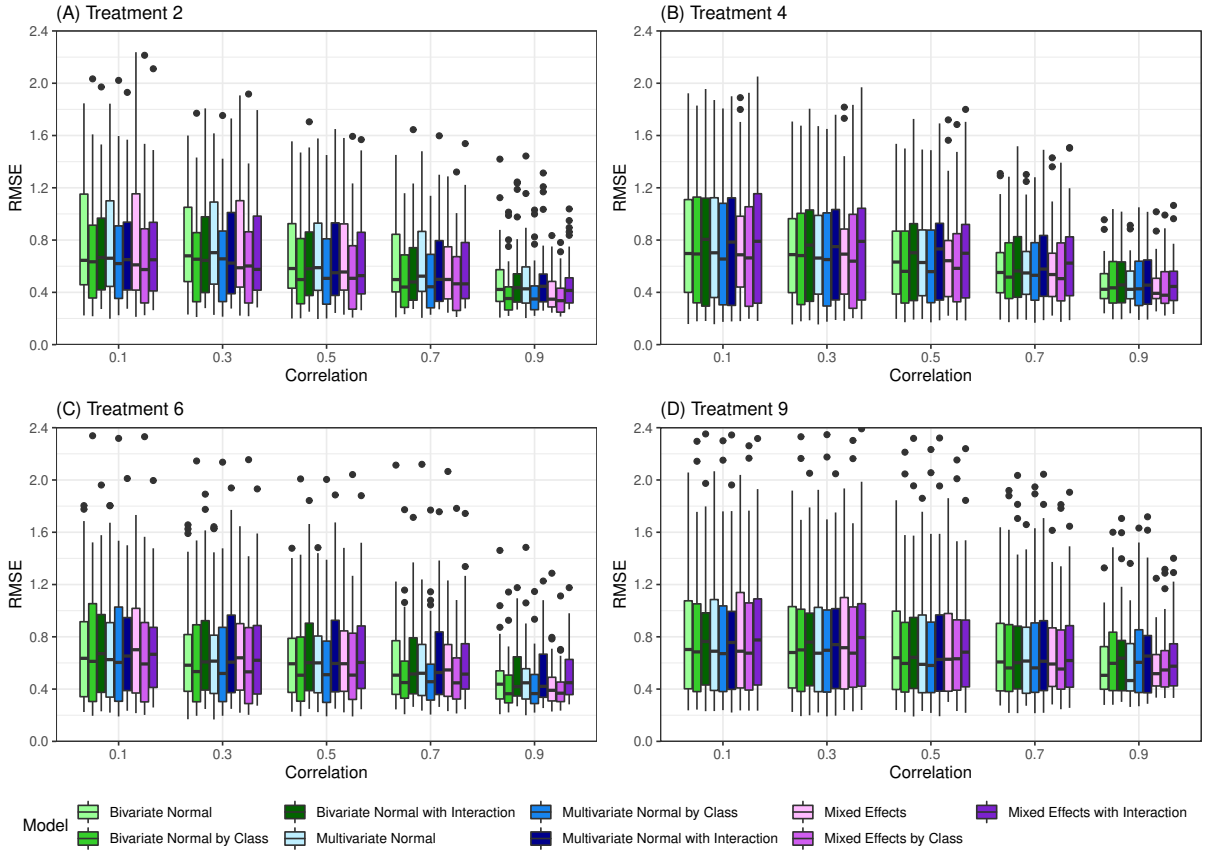


Figure 4.2: RMSE of the posterior median of d_{1t}^2 arising from $\boldsymbol{d}$'s simulated using coefficient set 1 (i.e. $(\beta_0, \beta_1, \beta_2) = (3.4, 1.3, -1.6)$). All models were fit using the $\text{Ht}_7(0, 2.5)$ prior for the between-study heterogeneity parameter and results are presented by treatment and $\text{Cor}(d_{1k}^1, d_{1k}^2)$.

A plot of the ranking distribution for RMSE is shown in Appendix Figure C.2. We

assigned each model a rank of 1 through 9 for each $\boldsymbol{d}$ vector, with rank 1 indicating the model with the smallest RMSE. The model with the smallest average rank varied by the amount of available information for treatment t in indication 1. For example, when $\text{Cor}(d^1_{1k},d^2_{1k})$=0.10 and a substantial amount of information was available for treatment t ($t = 2$), the mixed effects model with a homogeneous class effect had the smallest average rank. In contrast, the corresponding multivariate normal model had the smallest average rank when only a moderate ($t = 4$ or 6) or small amount ($t = 9$) of information was available for treatment t. The bivariate normal model tended to perform intermediately.

Figure 4.3 contains a similar figure to that in Figure 4.2 but with respect to the average width of the 95% credible interval for d^2_{1t}. As the correlation between d^1_{1k} and d^2_{1k} increased, the credible intervals became smaller, on average, and the variability of the distribution decreased. Within each class of models, predictions were generally most precise when a homogeneous class effect was incorporated. Within each mean structure, predictions were generally most precise for the multivariate normal model when the correlation was less than or equal to 0.5 and the mixed effects model when the correlation was greater than 0.5. For example, consider $t = 4$; when the correlation was 0.3, the average width of the 95% credible interval was 3.49, 3.35, and 3.14 for the bivariate normal, mixed effects, and multivariate normal models with homogeneous class effects, respectively. When the correlation was 0.9, the average widths were 2.84, 2.44, and 2.61, respectively. A plot of the average coverage probability of the 95% credible interval is also provided in Appendix Figure C.3, which ranged between 0.86 and 0.99 with a mean of 0.96 for all scenarios considered. The average coverage probability tended to increase with $\text{Cor}(d^1_{1k},d^2_{1k})$ and the amount of available information on treatment t.

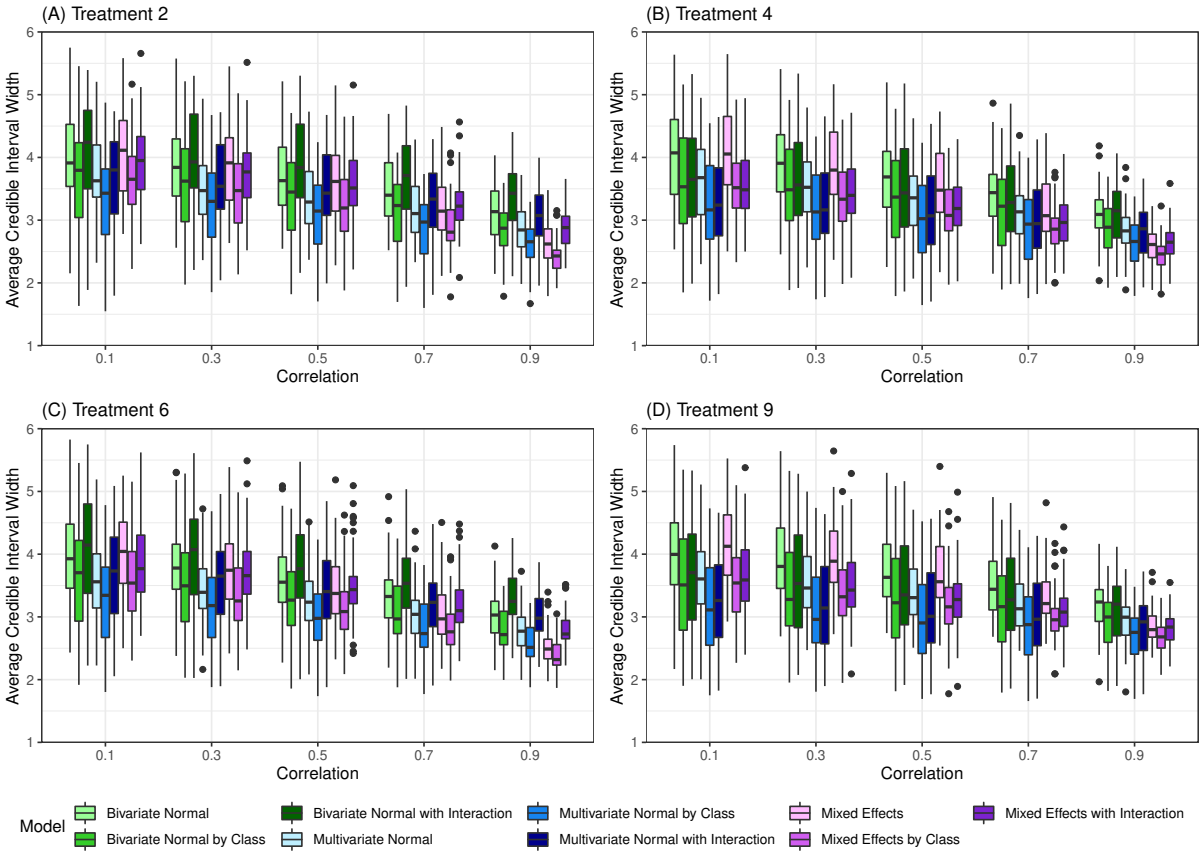


Figure 4.3: Average width of the 95% credible interval for d^2_{1t} arising from $\boldsymbol{d}$'s simulated using coefficient set 1 (i.e. $(\beta_0, \beta_1, \beta_2) = (3.4, 1.3, -1.6)$). All models were fit using the $\text{Ht}_7(0, 2.5)$ prior for the between-study heterogeneity parameter and results are presented by treatment and $\text{Cor}(d^1_{1k}, d^2_{1k})$.

The RMSE and average width of the 95% credible interval for d^2_{12} arising from $\boldsymbol{d}$'s simulated using a heterogeneous class effect are presented in panel (A) of Figure 4.4. Models incorporating a homogeneous class effect performed notably worse since data were generated using an interaction between drug class and indication. In contrast, models incorporating a heterogeneous class effect had the smallest RMSE, on average, when the correlation was

greater than or equal to 0.5; otherwise, models assuming a null class effect performed the most favorably. This is likely attributable to the fact that, for the interaction models, reductions in bias only exceeded increases in variances when correlation was high and hence random error was low, at which point model parameters can be estimated more precisely. A plot of the corresponding ranking distribution for RMSE is shown in Appendix Figure C.4. Average rank was minimized by the multivariate normal model incorporating a null class effect when the correlation was less than 0.9 but by the mixed effects model with heterogeneous class effect when the correlation equaled 0.9. Predictions were consistently most precise when using the multivariate normal model.

Panel (B) of Figure 4.4 displays the RMSE and average width of the 95% credible interval for d_{12}^2 arising from $\boldsymbol{d}$'s simulated using smaller class and indication effects. In contrast to coefficient set 1, RMSE was generally smallest for models incorporating a null drug class effect, particularly when correlation was low. This is likely because improvements in bias resulting from correctly incorporating a homogeneous drug class effect were smaller in this scenario, and hence increases in variance ultimately increased RMSE. A plot of the corresponding ranking distribution for RMSE is shown in Appendix Figure C.6. Average rank was minimized by the multivariate normal model incorporating a null class effect when the correlation was less than 0.9 but by the mixed effects model with a homogeneous class effect when the correlation equaled 0.9. The same was true for precision, where the average width of the 95% credible interval was minimized by the multivariate normal model when the correlation was less than 0.9 but by the mixed effects model when the correlation equaled 0.9.

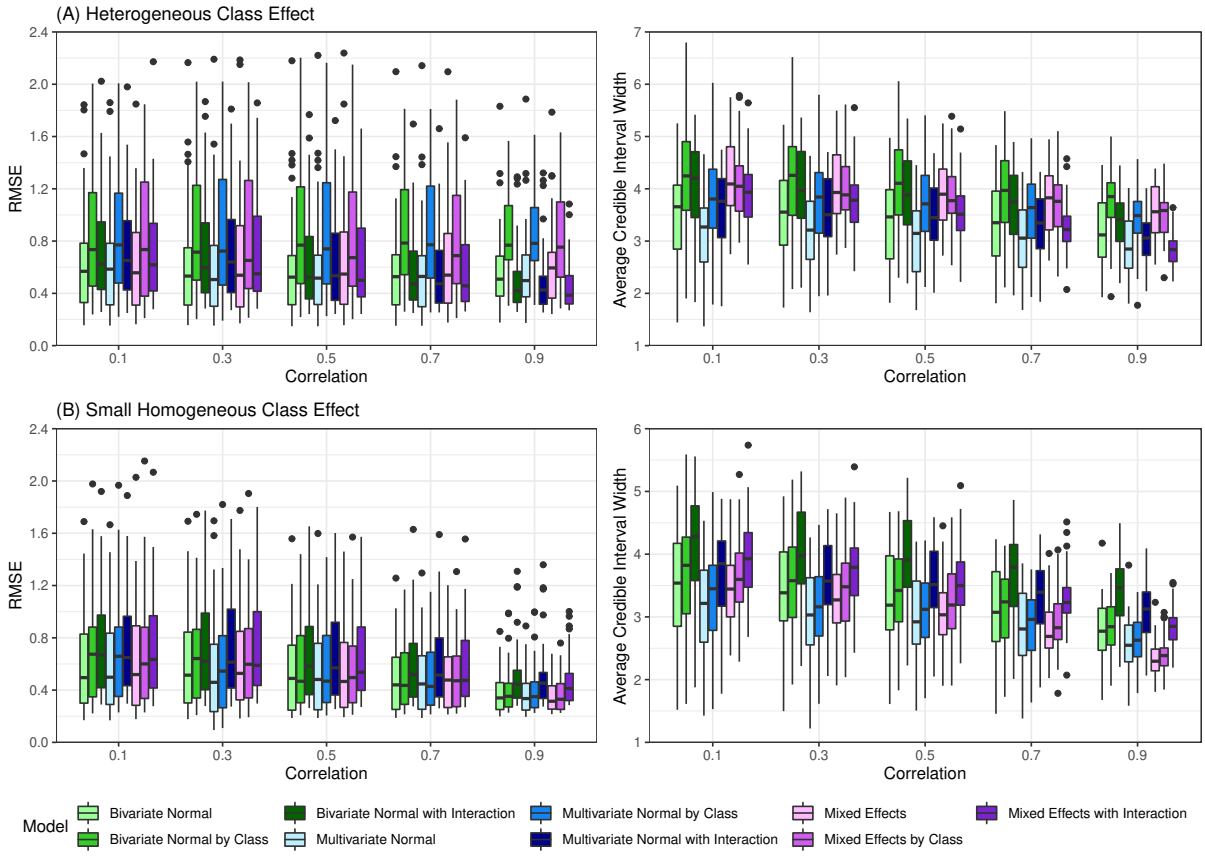


Figure 4.4: RMSE of the posterior median and average 95% credible interval width of d_{12}^2 arising from $\boldsymbol{d}$'s simulated using a (A) heterogeneous or (B) small homogeneous class effect. All models were fit using the $\text{Ht}_7(0, 2.5)$ prior for the between-study heterogeneity parameter and results are presented by $\text{Cor}(d_{1k}^1, d_{1k}^2)$.

Comparisons between prior distributions for the between-study heterogeneity parameter are made in Section C.3.4 of the Appendix using the mixed effects model incorporating a homogeneous class effect. Predictions did not meaningfully differ by prior distribution with respect to RMSE, average width of the 95% credible interval, or coverage probability.

4.5 Case Study

To demonstrate the use of our methods in practice, we considered an unpublished data set comprised of publicly available data from phase II, III, or IV studies in psoriasis, a chronic inflammatory autoimmune disease characterized by red scaly patches on the skin, and psoriatic arthritis, a chronic inflammatory autoimmune disease characterized by joint pain, stiffness, and swelling. [68] The endpoint was a binary outcome indicating a 75% or greater reduction in Psoriasis Area and Severity Index scores from baseline (PASI75) and hence excellent disease improvement. PASI75 was reported at 12 weeks for each study in psoriasis and 24 or 26 weeks for each study in psoriatic arthritis. Prior to analysis, we pruned the data to focus on overlapping treatments by removing extraneous arms from multi-arm studies that were either not of interest or required to maintain a fully connected network. The resulting network diagram is shown in Figure 4.5. For confidentiality reasons, we use A to refer to placebo and B through J to refer to active treatments. Notably, treatment E has only been studied and approved in psoriasis. The goal of this case study was to determine if treatment E is also worth pursuing in psoriatic arthritis.

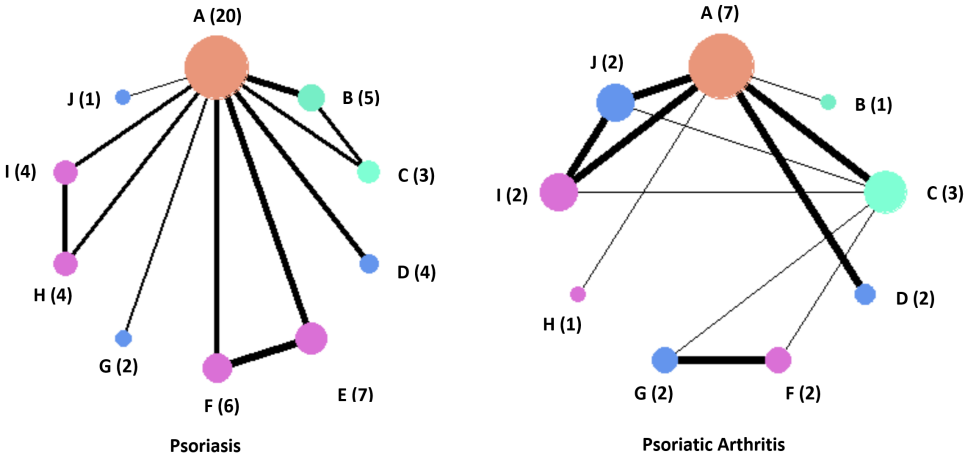


Figure 4.5: Network plots for psoriasis ($J_1 = 23$) and psoriatic Arthritis ($J_2 = 9$) with 9 overlapping treatments. Each node represents a treatment and each edge represents a direct comparison between two treatments. The size of each node is proportional to the number of studies evaluating that treatment whereas the width of each edge is proportional to the number of direct comparisons. The network reference treatment is placebo ($k = A$) and treatments B through J are active agents. Nodes that are green correspond to drug class 1, nodes that are blue correspond to drug class 2, and nodes that are purple correspond to drug class 3.

We began by fitting the standard RE-NMA model in (4.1) to each indication separately. We used the LN(-2.70,1.52) prior for the between-studies heterogeneity parameter given the limited number of studies in psoriatic arthritis. The posterior medians of each basic parameter are shown in Appendix Table C.1. The mean difference in posterior medians for psoriasis and psoriatic arthritis is 1.3 and the average posterior medians for drug classes 1, 2, and 3 are 2.6, 3.0, and 4.5 and 2.2, 2.1, and 2.5 for psoriasis and psoriatic arthritis, respectively. Moreover, the empirical correlation between the posterior medians of the overlapping basic parameters is 0.68. Given that treatment efficacy appears to differ by drug class only for psoriasis and the empirical correlation between indications is moderate,

we opted to use the multivariate normal model with a null class effect to predict d^2_{AE}.

The estimated log-odds ratio (95% credible intervals) for each active treatment relative to placebo in both indications is shown in Figure 4.6. Estimates are provided for both the standard RE-NMA and multivariate normal models. Notably, the point estimates corresponding to overlapping treatments are very similar between models, and estimates engendered by the proposed model are also more precise, particularly for treatments that were sparsely studied (e.g., treatment J in psoriasis and treatment B in psoriatic arthritis). The predicted log-odds ratio comparing treatment E to A in psoriatic arthritis is 2.52 with a 95% credible interval of (1.91, 3.24), suggesting treatment E may warrant further study in this indication.

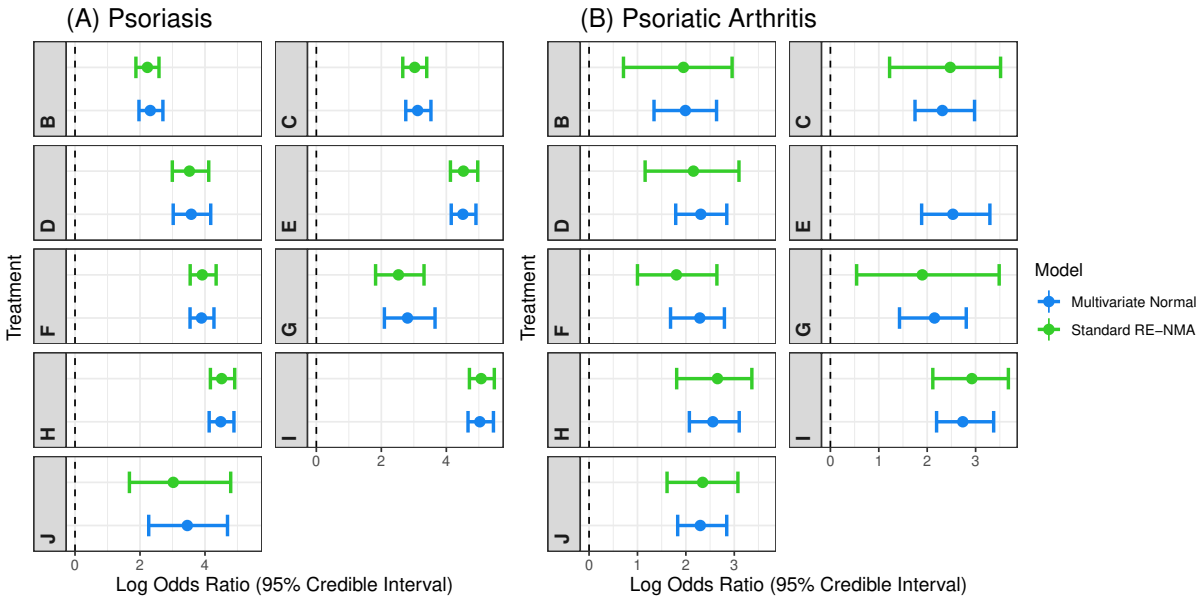


Figure 4.6: The estimated log-odds ratios (95% credible intervals) for treatments B through J relative to placebo in psoriasis and psoriatic arthritis by model type. Note that the log-odds ratio comparing treatment E to A in psoriatic arthritis could only be estimated by our proposed multivariate normal model.

To determine whether resources should be expended to pursue this treatment in a future study, we computed a metric known as probability of success (PoS). PoS is broadly defined as the probability of achieving a "success" in a future trial, where "success" may be defined as achieving statistical significance, observing a clinically meaningful effect, or attaining a favorable safety profile for the treatment under study. It is computed by marginalizing the traditional power function over a prior distribution for the unknown treatment effect, which is often estimated using observed data from previous trials. [96–98] Since d^2_{AE} is a population-level effect that fails to account for between-study variability, we instead computed PoS using:

$$\text{PoS} = \int_{\delta_{j^*AE}} \int_{d^2 AE} \Pr(\text{Success}|\delta_{j^*AE}) \text{p}(\delta_{j^*AE}|d^2_{AE}, \sigma^2_2) \text{p}(d^2_{AE}|\{\boldsymbol{D_j}\}_{j=1}^J)\ d\delta_{j1t} dd^2_{1t} \quad (4.9)$$

where δ_{j^*AE} is the study-specific log-odds ratio comparing treatment E vs. A in a future trial, j^*. We defined "success" as the probability of having a two-sided p-value less than 0.05 and estimated $\text{p}(\delta_{j^*AE}|d^2_{AE}, \sigma^2_2)$ and $\text{p}(d^2_{AE}|\{\boldsymbol{D_j}\}_{j=1}^J)$ from (4.9) using the multivariate normal model. We assumed that the future phase II study would enroll 20 participants to each arm to provide 90% power to detect a significant difference in response rates (with $p_A = 0.09$ and $p_E = 0.55$) and a two-sided type I error rate of 0.05. Additional computational details are provided in Appendix Section C.4.1. The resulting PoS was 83%, suggesting investigators should proceed with the planned design.

4.6 Discussion

This paper presents a novel NMA framework that can predict the effect of an approved treatment in a new indication by modeling the correlation between average conditional effects for a subset of treatments that have been studied in both indications. Importantly, our proposed models impute a posterior distribution for the aforementioned effect that can be used to identify repurposable treatments during drug development via a data-driven

approach. When the performance of the approved treatment in a future trial is of interest, our case study also demonstrates that our methods can be seamlessly integrated into probability of success (PoS) calculations. In contrast to the traditional power function, PoS advantageously considers the uncertainty associated with the unknown treatment effect and therefore more accurately reflects the true likelihood of success in a future trial. [96–98] It can also assist investigators with trial planning: in the event that PoS is lower than anticipated, investigators can use this methodology to modify the planned design until the desired level of assurance is attained.

Our simulations indicate that the predictive performance of each model improves with respect to RMSE and precision as the correlation between indications increases. The model that minimizes RMSE, on average, depends on the following factors: the amount of available information on the candidate treatment, the level of correlation between indications, the presence of a homogeneous or heterogeneous class effect, and the magnitude of the class effect if present. To select an appropriate model in practice, we recommend following a similar approach to that in our case study. That is, we recommend fitting a standard RE-NMA model to each indication separately and assessing the resulting posterior medians for the basic parameters. The absolute mean difference in posterior medians from different drug classes can be used to assess the nature of the drug class effect, whereas the empirical correlation between posterior medians from different indications can be used to evaluate the strength of the between-indication relationship. Then, a model can be selected using the decision aid provided in Section C.5 of the Appendix. Interestingly, the multivariate normal model appears to perform most favorably when little information is available. This is likely because it imposes greater regularization on the basic parameters compared to the other models, which is particularly beneficial in low information settings where the basic parameters are highly variable.

This research has several limitations. We considered studies with binary endpoints;

however, this methodology can be easily extended to other data types by modifying equation (4.1). We focused on networks where all but one treatment have been previously studied in both indications. When there are multiple candidates for drug repurposing, both the bivariate and multivariate normal approaches can simultaneously impute multiple missing basic parameters via MCMC sampling, whereas the mixed effects approach can simultaneously estimate multiple missing basic parameters using the linear predictor from equation (4.7). However, the conclusions from our simulation study may no longer hold in this scenario, and as such it would be interesting to assess model performance when predicting the effect of multiple treatments in future work. We only considered one information-rich network in our simulation study. When fewer studies are available for one or both indications, we would expect both RMSE and the average width of the 95% credible interval to increase for each proposed model. Finally, our proposed methodology only considers two indications and a common endpoint. An interesting area for future work would be to expand these methods to consider multiple indications and endpoints to determine if increasing our evidence base in this fashion yields more accurate or precise predictions.

Chapter 5

Conclusion

5.1 Summary of Major Findings

This thesis proposed several methods to reduce patient exposure to ineffective treatments in early phase drug development pertaining to response-adaptive randomization and drug repurposing. Chapter 2 introduced a novel probability model and randomization strategy for implementing Bayesian RAR in a binary outcomes setting. Simulation studies demonstrated that the proposed logistic regression model has increased power, more favorable treatment arm sample size distributions, and reduced type I error rate sensitivity to the underlying response probability compared to the traditional beta-binomial modeling approach. These improvements in performance are likely due to shrinkage that arises from placing less prior density on extreme values of the response probabilities. Moreover, our simulations demonstrated that the proposed modified permuted block design has a negligible chance of a sample size imbalance in the wrong direction and more effectively targets the evolving allocation ratio throughout the trial compared to complete randomization. The modified permuted block design also achieves the allocation ratio preserving property of Kuznetsova and Tymofyeyev, which advantageously protects against evaluation, selection, and accidental biases that may arise even in double-blind trials. [35]

Chapter 3 proposed a new design comparison metric that investigators may use to select a group sequential design that minimizes harm to potential participants, which may be appealing in trials where the primary outcome is mortality or some other outcome with severe implications to quality of life. This metric is simple to implement, may be applied to any set of designs, and provides a means to reasonably evaluate the potential patient benefit of RAR against classical frequentist group sequential designs. Our simulations indicated that the restricted Thompson sampling Bayesian RAR design tends to perform the best with respect to our metric across a variety of scenarios, and that Bayesian RAR offers modest reductions in the number of failures in the potential study sample relative to group sequential monitoring alone. We also found that Bayesian RAR group sequential designs exhibit greater marginal gains relative to group sequential designs using 1:1 allocation when the treatment effect is slightly smaller than hypothesized. Conversely, these gains diminish when the treatment effect is much larger than hypothesized, likely because both designs will stop very early with high probability which precludes adaptations to the allocation ratio. Whether these gains are worthwhile needs to be assessed within the context of each trial based on the severity of the primary outcome, the plausibility of observing various effect sizes, and the increased complexity that comes with Bayesian RAR implementation.

Finally, Chapter 4 introduced a novel NMA framework that can predict the effect of an approved treatment in a new indication by modeling the correlation between average conditional effects for a subset of treatments that have been studied in both indications. Importantly, the proposed models impute a posterior distribution for the aforementioned effect that can be used to identify repurposable treatments during drug development via a data-driven approach. When the performance of the approved treatment in a future trial is of interest, our case study also demonstrated that these methods can be seamlessly integrated into probability of success (PoS) calculations. Our simulations indicated that the predictive performance of each model improves with respect to RMSE and precision as the correlation between indications increases. The model that minimizes RMSE, on average,

depends on the following factors: the amount of available information on the candidate treatment, the level of correlation between indications, the presence of a homogeneous or heterogeneous class effect, and the magnitude of the class effect if present. To select an appropriate model in practice, we recommend following a similar approach to that in our case study. That is, we recommend fitting a standard RE-NMA model to each indication separately and assessing the resulting posterior medians for the basic parameters. Then, a model can be selected using the decision aid provided in Figure C.11.

5.2 Future Research

The findings described in the previous sections lead to many opportunities for future work, which are described briefly below.

5.2.1 Future Work Stemming from Chapter 2:

Recall that this chapter assessed the impact of implementing Bayesian RAR using a logistic regression probability model and modified permuted block design in the context of the ARREST trial, which is a recently conducted phase II clinical trial with a binary primary outcome. The hypothesized treatment effect for this trial was 37% and block sizes of 15, 30, and 50 were considered. An extension of this work would be to assess the operating characteristics of the proposed design using different block and treatment effect sizes. For scenarios where the the targeted effect size is smaller than 37%, we would expect simulation studies to indicate larger average sample sizes and smaller differences in treatment arm sample sizes. For scenarios with smaller block sizes, we would expect to observe greater variability early on in the trial, which could exacerbate type I error rate inflation and the probability of assigning more participants tot he inferior arm.

Further research is also needed to evaluate our recommendations in a multi-arm trial setting. As discussed by Trippa et al., another way to improve the performance of multi-arm

response-adaptive design is by incorporating tuning parameters into the randomization algorithm that regulate the exploration versus exploitation trade-off. [37] A possible extension of our work would be to assess the performance of a multi-arm response-adaptive design using our recommended probability model and a tuning randomization algorithm. In the context of multi-arm bandit models, Villar et al. have also demonstrated that non-myopic randomization procedures may influence the operating characteristics of response-adaptive designs. [38, 39] Defining our randomization procedure in terms of the patient horizon, or the total patient population, is another possible direction for future work. [39, 40]

5.2.2 Future Work Stemming from Chapter 3:

Due to the motivating context for this work, we only focused on two-arm trials with a binary primary outcome and a select set of Bayesian RAR procedures. Our conclusions regarding the potential patient gains arising from RAR with respect to our proposed metric may vary in different design settings or when alternative RAR procedures are considered. We also did not account for the presence of time trends including patient drift, which has been shown to inflate type I error rate in response-adaptive designs [6]. Therefore, it may be beneficial to conduct additional simulation studies to assess the potential patient gains arising from RAR under additional settings.

We did not consider multi-arm trials in which Bayesian RAR has been shown to increase efficiency relative to balanced randomization. An interesting area for future work would be to study our metric in the context of platform trials or multi-arm multi-stage (MAMS) designs (see Lin and Bunn [64] and Watson and Trippa [65]). Finally, although we declare the optimal group sequential design to be the one that minimizes the average number of failures in the potential study sample among the designs considered, we did not fully optimize this criterion. Future work, including the evaluation of frequentist group sequential designs whose characteristics have been optimized such that the expected number of failures in the potential study sample is minimal, is needed to identify the design that fully minimizes

the proposed metric.

5.2.3 Future Work Stemming from Chapter 4:

Chapter 4 focused on networks where all but one treatment have been previously studied in both indications. When there are multiple candidates for drug repurposing, both the bivariate and multivariate normal approaches can simultaneously impute multiple missing basic parameters via MCMC sampling, whereas the mixed effects approach can simultaneously estimate multiple missing basic parameters using the linear predictor from equation (4.7). However, the conclusions from our simulation study may no longer hold in this scenario, and as such it would be interesting to assess model performance when predicting the effect of multiple treatments in future work. In addition, we only considered one information-rich network in our simulation study. When fewer studies are available for one or both indications, we would expect both RMSE and the average width of the 95% credible interval to increase for each proposed model, however this should be verified via additional simulation studies. Finally, our proposed methodology only considers two indications and a common endpoint. An interesting area for future work would be to expand these methods to consider multiple indications and endpoints to determine if increasing our evidence base in this fashion yields more accurate or precise predictions.

References

[1] A. Sertkaya, H. H. Wong, A. Jessup, and T. Beleche. Key cost developers of pharmaceutical clinical trials in the united states. *Clinical Trials*, 13(2):117–126, 2016.

[2] S. M. Berry, B. P. Carlin, J. J. Lee, and P. Muller. *Bayesian Adaptive Methods for Clinical Trials*. Champman Hall/CRC Biostatistics Series. CRC Press: Taylor and Francis Group, LLC, Boca Raton, FL, 2011.

[3] P. Pallmann, A. W. Bedding, B. Choodari-Oskooei, M. Dimairo, L. Flight, L. V. Hampson, J. Holmes, A. P. Mander, L. Odondi, M. R. Sydes, S. S. Villar, J. M. S. Wason, C. J. Weir, C. M. Wheeler, and T. Jaki C. Yap. Adaptive designs in clinical trials: Why use them, and how to run and report them. *BMC Medicine*, 16(1), 2018.

[4] P. F. Thall and J. K. Wathen. Practical Bayesian adaptive randomization in clinical trials. *European Journal of Cancer*, 43:859–866, 2007.

[5] P. Thall, P. Fox, and J. Wathen. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Annals of Oncology*, 26:1621–1628, 2015.

[6] Peter F. Thall, Patricia S. Fox, and J. Kyle Wathen. Some caveats for outcome adaptive randomization in clinical trials. In Oleksandr Sverdlov, editor, *Modern adaptive randomized clinical trials: Statistical and practical aspects*, chapter 13, pages 287–305. CRC Press, Boca Raton, 2016.

[7] E. L. Korn and B. Freidlin. Outcome-adaptive randomization - is it useful? *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 29:771–776, 2010.

[8] S. P. Hey and J. Kimmelman. Are outcome-adaptive allocation trials ethical? *Clinical Trials*, 12:102–106, 2015.

[9] Sofia S. Villar, Jack Bowden, and James Wason. Response-adaptive designs for binary responses: How to offer patient benefit while being robust to time trends? *Pharmaceutical Statistics*, 17:182–197, 2018.

[10] Richard Simon and Noah Robin Simon. Using randomization tests to preserve type I error with response-adaptive and covariate-adaptive randomization. *Statistics Probability Letters*, 81:767–772, 2010.

[11] Babak Sahragardjoonegani, Reed F. Beall, Aaron S. Kesselheim, and Aidan Hollis. Repurposing existing drugs for new uses: a cohort study of the frequency of FDA-granted new indication exclusivities since 1997. *Journal of Pharmaceutical Policy and Practice*, 14(1):3, 2021.

[12] Antii Jekunen. Decision-making in product portfolios of pharmaceutical research and development – managing streams of innovation in highly regulated markets. *Drug Design, Development and Therapy*, 8:2009–2016, 2014.

[13] Demetris Yannopoulos, Rajat Kalra, Marinos Kosmopoulos, Emily Wasler, Jason A. Bartos, Thomas A. Murray, John E. Connett, and Tom P. Aufderheide. Rationale and methods of the Advanced R^2Eperfusion STrategies for Refractory Cardiac Arrest (ARREST) trial. *American Heart Journal*, 229:29–39, 2020.

[14] Demetris Yannopoulos, Jason Bartos, Ganesh Raveendran, Emily Walser, John Connett, Thomas A. Murray, Gary Collins, Lin Zhang, Rajat Kalra, Marinos Kosmopoulos, Ranjit John, Andrew Shaffer, RJ Frascone, Keith Wesley, Marc Conterato,

Michelle Biros, Jakub Tolar, and Tom P. Aufderheide. Advanced reperfusion strategies for patients with out-of-hospital cardiac arrest and refractory ventricular fibrillation (ARREST): a phase 2, single centre, open-label, randomised controlled trial. *The Lancet*, 2020.

[15] Christopher Jennison and Bruce W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman Hall, Boca Raton, FL, 2000.

[16] Theodore G. Karrison, Dezheng Huo, and Rick Chappell. A group sequential, response-adaptive design for randomized clinical trials. *Controlled Clinical Trials*, 24:506–522, 2003.

[17] Caroline C. Morgan and D. Stephen Coad. A comparison of adaptive allocation rules for group-sequential binary response clinical trials. *Statistics in Medicine*, 26:1937–1954, 2007.

[18] Hongjian Zhu and Feifang Hu. Sequential monitoring of response-adaptive randomized clinical trials. *The Annals of Statistics*, 38:2218–2241, 2010.

[19] Georgia Salanti, Julian PT Higgins, AE Ades, and John PA Ioannidis. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research*, 17(3):279–301, 2008.

[20] J. Zhang, B. P. Carlin, J. D. Neaton, G. G. Soon, L. Nie, R. Kane, B. A. Virnig, and H. Chu. Network meta-analysis of randomized clinical trials: reporting the proper summaries. *Clinical Trials*, 11(2):246–262, 2014.

[21] H. Hong, H. Chu, J. Zhang, and B. P. Carlin. A bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research Synthesis Methods*, 7(1):6–22, 2016.

[22] D. Mawdsley, M. Bennetts, S. Dias, M. Boucher, and N. J. Welton. Model-based network meta-analysis: A framework for evidence synthesis of clinical trial data. *CPT: pharmacometrics & systems pharmacology*, 5(8):393–401, 2016.

[23] Hugo Pedder, Sofia Dias, Margherita Bennetts, Martin Boucher, and Nicky J. Welton. Modelling time-course relationships with multiple treatments: Model-based network meta-analysis for continuous summary outcomes. *Research Synthesis Methods*, 10(2):267–286, 2019.

[24] P A Milligan, M J Brown, B Marchant, S W Martin, P H van der Graaf, N Benson, G Nucci, D J Nichols, R A Boyd, J W Mandema, S Krishnaswami, S Zwillich, D Gruben, R J Anziano, T C Stock, and R L Lalonde. Model-based drug development: A rational approach to efficiently accelerate drug development. *Clinical Pharmacology & Therapeutics*, 93(6):502–514, 2013.

[25] J. K. Wathen and P. F. Thall. A simulation study of outcome adaptive randomization in multi-arm clinical trials. *Clinical Trials*, 14:432–440, 2017.

[26] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of the two samples. *Biometrika*, 25:285–294, 1933.

[27] J. Ghosh, Y. Li, and R. Mitra. On the use of cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, 13, 2015.

[28] A. Gelman, A. Jakulin, M. G. Pittau, and Y. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

[29] S. Morita, P. F. Thall, and P. Muller. Determining the effective sample size of a parametric prior. *Biometrics*, 64:595–602, 2008.

[30] Wenle Zhao and Yanqiu Weng. Block urn design — a new randomization algorithm for sequential trials with two or more treatments and balanced or unbalanced allocation. *Contemporary Clinical Trials*, 32:953–961, 2011.

[31] Anastasia Ivanova. A play-the-winner-type urn design with reduced variability. *Metrika*, 58:1–13, 2003.

[32] Olga M. Kuznetsova and Yevgen Tymofyeyev. Wide brick tunnel randomization – an unequal allocation procedure that limits the imbalance in treatment totals. *Statistics in Medicine*, 33:1514–1530, 2014.

[33] Yevgen Ryeznik and Oleksandr Sverdlov. A comparative study of restricted randomization procedures for multiarm trials with equal or unequal treatment allocation ratios. *Statistics in Medicine*, 37:3056–3077, 2018.

[34] W. Zhao. Mass weighted urn design - a new randomization algorithm for unequal allocations. *Contemporary Clinical Trials*, 43:209–216, 2015.

[35] Olga M. Kuznetsova and Yevgen Tymofyeyev. Expansion of the modified Zelen's approach randomization and dynamic randomization with partial block supplies at the centers to unequal allocation. *Contemporary Clinical Trials*, 32:962–972, 2011.

[36] J. J. Lee, N. Chen, and G. Yin. Worth adapting? Revisiting the usefulness of outcome-adaptive randomization. *Clinical Cancer Research*, 18:4498–4507, 2012.

[37] Lorenzo Trippa, Eudocia Q. Lee, Patrick Y. Wen, Tracy T. Batchelor, Timonthy Cloughesy, Giovanni Parmigiani, and Brian M. Alexander. Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 30:3258–3263, 2012.

[38] Sofia S. Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30:199–215, 2015.

[39] Sofia S. Villar, James Wason, and Jack Bowden. Response-adaptive randomization for multi-arm clinical trials using the forward looking gittins index rule. *Biometrics*, 71:969–978, 2015.

[40] Yi Cheng and Donald A. Berry. Optimal adaptive randomized designs for clinical trials. *Biometrika*, 94:673–689, 2007.

[41] J. J. Lee and D. D. Liu. A predictive probability design for phase II cancer clinical trials. *Clinical Trials*, 5(2):93–106, 2008.

[42] N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using polya-gamma latent variables. *Journal of the American Statistical Association*, 108:1339–1349, 2013.

[43] Douglas G. Altman and J Martin Bland. Treatment allocation in controlled trials: why randomize? *BMJ (Clinical Research ed.)*, 318:1209, 1999.

[44] Christopher R. Palmer. Ethics and statistical methodology in clinical trials. *Journal of medical ethics*, 19:219–222, 1993.

[45] Christopher R. Palmer and William F. Rosenberger. Ethics and practice: Alternative designs for phase iii randomized clinical trials. *Controlled Clinical Trials*, 20:172–186, 1999.

[46] Demetris Yannopoulos and Tom P. Aufderheide. Access to the cardiac cath lab in patients without STEMI resuscitated from out-of-hospital vt/vf cardiac arrest.

[47] Q. Yao and L. J. Wei. Play the winner for phase II/III clinical trials. *Statistics in Medicine*, 15:2413–2423, 1996.

[48] D.S. Coad and Willliam F. Rosenberger. A comparison of the randomized play-the-winner rule and the triangular test for clinical trials with binary responses. *Statistics in Medicine*, 18:761–769, 1999.

[49] David S. Robertson, Kim May Lee, Boryana C. Lopez-Kolkovska, and Sofia S. Villar. Response-adaptive randomization in clinical trials from myths to practical considerations. Technical report, 2020.

[50] W. F. Rosenberger, N. Stallard, A. Ivanova, C. N. Harper, and M. L. Ricks. Optimal adaptive designs for binary response trials. *Biometrics*, 57:909–913, 2001.

[51] Feifang Hu and Li-Xin Zhang. Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *The Annals of Statistics*, 322:268–301, 2004.

[52] Feifang Hu, Li-Xin Zhang, and Xuming He. Efficient randomized-adaptive designs. *The Annals of Statistics*, 37:2543–2560, 2009.

[53] F. J. Anscombe. Sequential medical trials. *Journal of the American Statistical Association*, 58:365–383, 1963.

[54] J. Proper, J. Connett, and T. Murray. Alternative models and randomization techniques for Bayesian response-adaptive randomization with binary outcomes. *Clinical Trials*, 18:417–426, 2021.

[55] K. Lange, R. J. A. Little, and J. Taylor. Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association*, 84:881–896, 1989.

[56] V. Melfi and C. Page. Variability in adaptive designs for estimation of success probabilities. In N. Flournoy, W. F. Rosenberger, and W. K. Wong, editors, *New Developments and Applications in Experimental Design*, pages 106–114. Institute of Mathematical Statistics, Hayward, CA, 1998.

[57] Feifang Hu and William F. Rosenberger. Optimality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association*, 98:671–678, 2003.

[58] Haolun Shi and Guosheng Yin. Control of type I error rates in Bayesian sequential designs. *Bayesian Analysis*, 14:399–425, 2019.

[59] Feifang Hu and William F. Rosenberger. *The theory of response-adaptive randomization in clinical trials*. John Wiley Sons, Inc, Hoboken, NJ, USA, 2006.

[60] K. K. Gordan Lan and David L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70:659–663, 1983.

[61] Keaven Anderson. *gsDesign: Group Sequential Design*, 2020. R package version 3.1.1.

[62] Karim F. Hirji. *Exact Analysis of Discrete Data*. Champman and Hall/CRC. New York, 2006.

[63] Michael P. Fay and Sally A. Hunsberger. Unconditional exact tests in the exact2x2 R package, 2020.

[64] Jianchang Lin and Veronica Bunn. Comparison of multi-arm multi-stage design and adaptive randomization in platform clinical trials. *Contemporary Clinical Trials*, 54:48–59, 2017.

[65] James M. S. Watson and Lorenzo Trippa. A comparison of bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Statistics in Medicine*, 33:2206–2221, 2014.

[66] Bo Wang and Aaron S. Kesselheim. Characteristics of efficacy evidence supporting approval of supplemental indications for prescription drugs in united states, 2005-14: systematic review. *BMJ*, 351, 2015.

[67] Meera Dhodapkar, Audrey D. Zhang, Jeremy Puthumana, Nicholas S. Downing, Nilay D. Shah, and Joseph S. Ross. Characteristics of clinical studies used for US food and drug administration supplemental indication approvals of drugs and biologics, 2017 to 2019. *JAMA Network Open*, 4(6), 2021.

[68] Jennifer Clay Cather, Melodie Young, and Martin Jan Bergman. Psoriasis and psoriatic arthritis. *The Journal of Clinical and Aesthetic Dermatology*, 10(3):S16–S25, 2017.

[69] Harsha Jain, Aditi Rajan Bhat, Harshita Dalvi, Chandraiah Godugu, Shashi Bala Singh, and Saurabh Srivastava. Repurposing approved therapeutics for new indication: Addressing unmet needs in psoriasis treatment. *Current Research in Pharmacology and Drug Discovery*, 2, 2021.

[70] Deisy Morselli Gysi, Ítalo do Valle, Marinka Zitnik, Asher Ameli, Xiao Gan, Onur Varol, Susan Dina Ghiassian, J. J. Patten, Robert A. Davey, Joseph Loscalzo, and Albert-László Barabási. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences*, 118(19), 2021.

[71] Y Cha, T Erez, I J Reynolds, D Kumar, J Ross, G Koytiger, R Kusko, B Zeskind, S Risso, E Kagan, S Papapetropoulos, I Grossman, and D Laifenfeld. Drug repurposing from the perspective of pharmaceutical companies. *British Journal of Pharmacology*, 175(2):168–180, 2018.

[72] R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986.

[73] Sofia Dias, Nicky J. Welton, Alex J. Sutton, and A. E. Ades. Evidence synthesis for decision making 5: the baseline natural history model. *Medical Decision Making:*

An International Journal of the Society for Medical Decision Making, 33(5):657–670, 2013.

[74] Guobing Lu and A. E Ades. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association*, 101(474):447–459, 2006.

[75] G. Lu and A. E. Ades. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine*, 23(20):3105–3124, 2004.

[76] Sylwia Bujkiewicz, Dan Jackson, John R. Thompson, Rebecca M. Turner, Nicolas Städler, Keith R. Abrams, and Ian R. White. Bivariate network meta-analysis for surrogate endpoint evaluation. *Statistics in Medicine*, 38(18):3322–3341, 2019.

[77] S. Dias and A. E. Ades. Absolute or relative effects? arm-based synthesis of trial data. *Research Synthesis Methods*, 7(1):23–28, 2016.

[78] M. Xiao, H. Chu, S. R. Cole, Y. Chen, R. F. MacLehose, D. B. Richardson, and S. Greenland. Controversy and debate : Questionable utility of the relative risk in clinical research: Paper 4: Odds ratios are far from "portable" - a call to use realistic models for effect variation in meta-analysis. *Journal of Clinical Epidemiology*, 142:294–304, 2022.

[79] M. Xiao, Y. Chen, S. R. Cole, R. F. MacLehose, D. B. Richardson, and H. Chu. Controversy and debate: Questionable utility of the relative risk in clinical research: Paper 2: Is the odds ratio "portable" in meta-analysis? time to consider bivariate generalized linear mixed model. *Journal of Clinical Epidemiology*, 142:280–287, 2022.

[80] Lifeng Lin, Haitao Chu, and James S. Hodges. Sensitivity to excluding treatments in network meta-analysis:. *Epidemiology*, 27(4):562–569, 2016.

[81] Zhenxun Wang, Lifeng Lin, James S. Hodges, and Haitao Chu. The impact of covariance priors on arm-based bayesian network meta-analyses with binary outcomes. *Statistics in Medicine*, 39(22):2883–2900, 2020.

[82] Ian R. White, Rebecca M. Turner, Amalia Karahalios, and Georgia Salanti. A comparison of arm-based and contrast-based models for network meta-analysis. *Statistics in Medicine*, 38(27):5197–5213, 2019.

[83] Joyee Ghosh, Yingbo Li, and Robin Mitra. On the use of cauchy prior distributions for bayesian logistic regression. *Bayesian Analysis*, 13(2):359–383, 2018.

[84] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

[85] Rebecca M. Turner, Dan Jackson, Yinghui Wei, Simon G. Thompson, and Julian P. T. Higgins. Predictive distributions for between-study heterogeneity and simple methods for their application in bayesian meta-analysis. *Statistics in Medicine*, 34(6):984–998, 2015.

[86] Sofia Dias, Alex J. Sutton, A. E. Ades, and Nicky J. Welton. Evidence synthesis for decision making 2. *Medical Decision Making*, 33(5):607–617, 2013.

[87] Julian P. T. Higgins and Anne Whitehead. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine*, 15(24):2733–2749, 1996.

[88] Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534, 2006.

[89] Kristine J. Rosenberger, Aiwen Xing, Mohammad Hassan Murad, Haitao Chu, and Lifeng Lin. Prior choices of between-study heterogeneity in contemporary bayesian

network meta-analyses: an empirical study. *Journal of General Internal Medicine*, 36(4):1049–1057, 2021-04-01.

[90] D. Spiegelhalter, K. Abrams, and J. Myles. *Bayesian approaches to clinical trials and health-care evaluation*. Wiley, New York, 2004.

[91] Rebecca M Turner, Jonathan Davey, Mike J Clarke, Simon G Thompson, and Julian PT Higgins. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the cochrane database of systematic reviews. *International Journal of Epidemiology*, 41(3):818–827, 2012.

[92] Fiona C. Warren, Keith R. Abrams, and Alex J. Sutton. Hierarchical network meta-analysis models to address sparsity of events and differing treatment classifications with regard to adverse outcomes. *Statistics in Medicine*, 33(14):2449–2466, 2014.

[93] Helen A. Dakin, Nicky J. Welton, A.e. Ades, Sarah Collins, Michelle Orme, and Steven Kelly. Mixed treatment comparison of repeated measurements of a continuous endpoint: an example using topical treatments for primary open-angle glaucoma and ocular hypertension. *Statistics in Medicine*, 30(20):2511–2535, 2011.

[94] Martyn Plummer. *rjags: Bayesian Graphical Models using MCMC*, 2022. R package version 4-13.

[95] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Sciences*, 7(4):557–611, 1992.

[96] C. Chuang-Stein. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286, 2006.

[97] R. L. Lalonde and C. C. Peck. Probability of success: A crucial concept to inform decision making in pharmaceutical research and development. *Clinical Pharmacology Therapeutics*, 111(5):1001–1003, 2022.

[98] Y. Wang, F. Fu, P. Kulkami, and C. Kaiser. Evaluating and utilizing probability of study success in clinical development. *Clinical Trials*, 10:407–413, 2013.

[99] Stuart J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64:191–199, 1977.

[100] Peter C. O'Brien and Thomas R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35:549–556, 1979.

[101] V. Hasselblad and Yulia Lokhnygina. Tests for 2x2 tables in clinical trials. *Journal of Modern Applied Statistical Methods*, 6:456–468, 2007.

Appendix A

Supplementary Materials for "Alternative Models and Randomization Techniques for Bayesian Response-Adaptive Randomization with Binary Outcomes"

A.1 The Density Function of the Generalized t-distribution

The density function of the generalized t-distribution, where v, μ, and σ signify the degrees of freedom, location, and scale parameters, respectively, arises as:

$$p(x|v,\mu,\sigma) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})\sqrt{\pi v}\sigma}\left(1+\frac{1}{v}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{v+1}{2}}$$

A.2 Derivation of Equation 6

To see how the probabilities in equation (2.6) were obtained, note that we want $E[E_m]$, the expected number of subjects to be randomly assigned to the experimental treatment in block m, to equal bp_m exactly to facilitate accurate targeting of general allocation ratios. The following equation was used to determine the probability that U_m=1, denoted p_u, that ensures $E[E_m] = bp_m$:

$$E[E_m] = (\lceil bp_m \rceil) \cdot p_u + \lfloor bp_m \rfloor \cdot (1 - p_u) \stackrel{\text{set}}{=} bp_m$$

After some rearranging, it is clear that $p_u = bp_m$ - $\lfloor bp_m \rfloor$. Let a_i further denote the block assignment for subject i. We will now show that P($Z_i = E|a_i$=m) = p_m when $p_u = bp_m - \lfloor bp_m \rfloor$:

$$\begin{aligned}
P(Z_i = E|a_i = m) &= P(Z_i = E|U = 1) \cdot P(U = 1) + P(Z_i = E|U = 0) \cdot P(U = 0) \\
&= \frac{\lceil bp_m \rceil}{b} \cdot (bp_m - \lfloor bp_m \rfloor) + \frac{\lfloor bp_m \rfloor}{b} \cdot (\lceil bp_m \rceil - bp_m) \\
&= \frac{\lceil bp_m \rceil \cdot bp_m}{b} - \frac{\lceil bp_m \rceil \cdot \lfloor bp_m \rfloor}{b} + \frac{\lceil bp_m \rceil \cdot \lfloor bp_m \rfloor}{b} - \frac{bp_m \cdot \lfloor bp_m \rfloor}{b} \\
&= \frac{\lceil bp_m \rceil \cdot bp_m - bp_m \cdot \lfloor bp_m \rfloor}{b} \\
&= \frac{bp_m \cdot (\lceil bp_m \rceil - \lfloor bp_m \rfloor)}{b} \\
&= p_m
\end{aligned}$$

A.3 Type I Error at Various Null Response Rates Across Randomization Methods

All plotted values are from the logistic regression probability model with a Student-t prior intercept location of $\log(\frac{0.12}{0.88})$. The $P_{E>C}$ stopping boundaries under the key null scenario were 0.9931, 0.9931, and 0.9930 for the weighted coin, mass-weighted urn, and modified

permuted block designs, respectively.

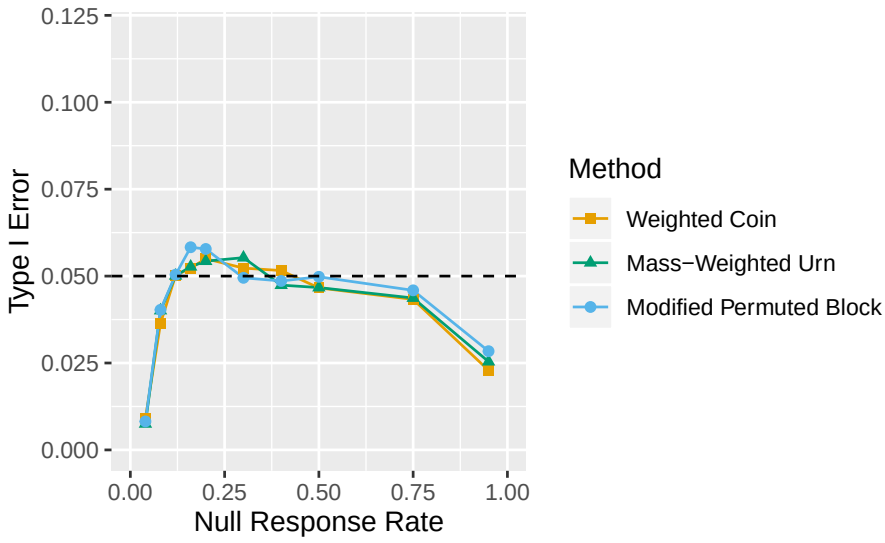
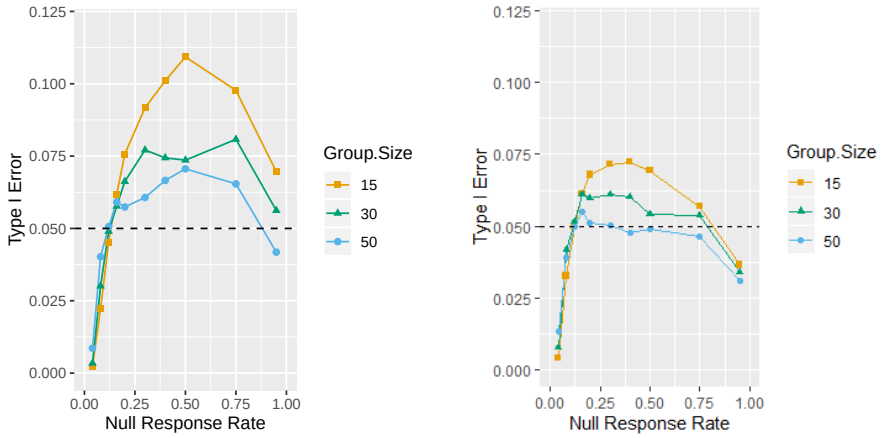


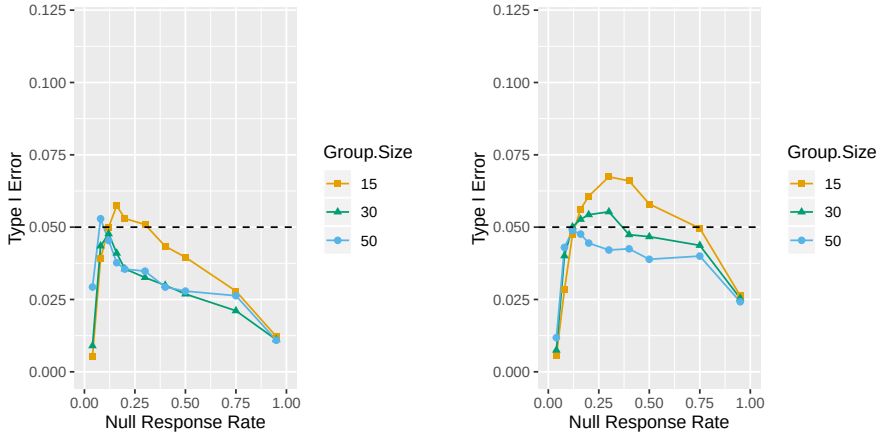
Figure A.1: Type I Error at Various Null Response Rates by Randomization Method for the Logistic Regression Model with a Student-t Prior Intercept Location of log(0.12/0.88) and Sequential Group Size of 30

A.4 Type I Error Plots: Mass-Weighted Urn Design



(a) Independent Beta-Binomial Model with Prior Mean 0.50

(b) Logistic Regression Model with Location = 0

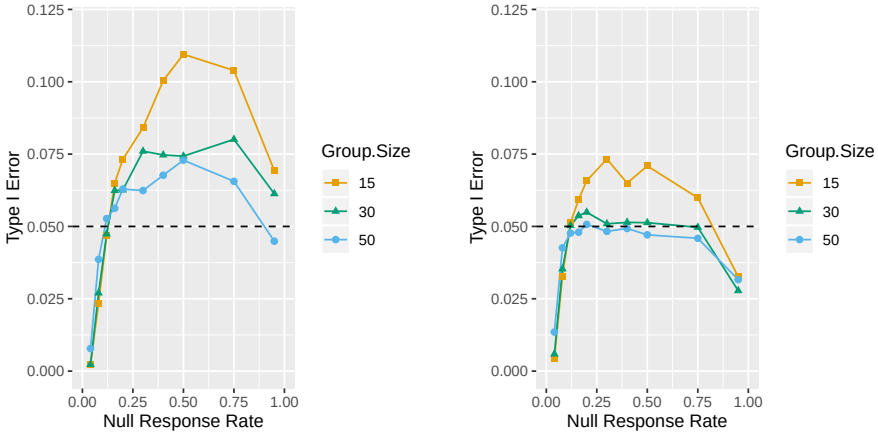


(c) Independent Beta-Binomial Model with Prior Mean 0.12

(d) Logistic Regression Model with Location = log(0.12/0.88)

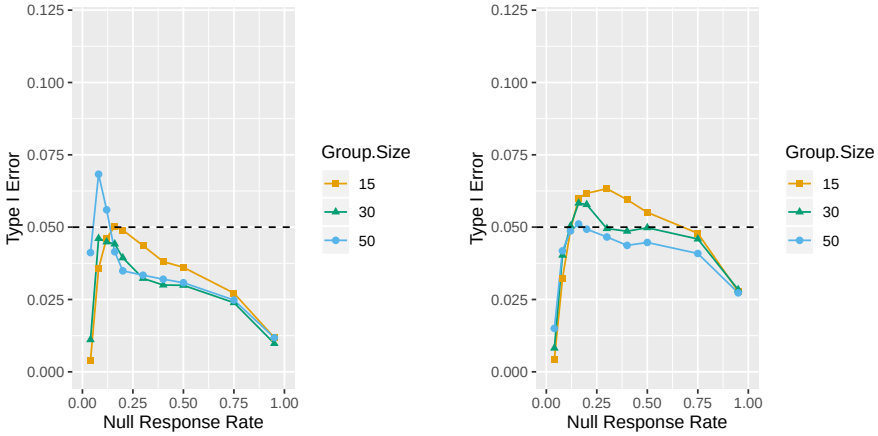
Figure A.2: Type I Error at Various Null Response Rates and Sequential Group Sizes by Probability Model (Mass-Weighted Urn Design)

A.5 Type I Error Plots: Modified Permuted Block Design



(a) Independent Beta-Binomial Model with Prior Mean 0.50

(b) Logistic Regression Model with Location = 0



(c) Independent Beta-Binomial Model with Prior Mean 0.12

(d) Logistic Regression Model with Location = log(0.12/0.88)

Figure A.3: Type I Error at Various Null Response Rates and Sequential Group Sizes by Probability Model (Modified Permuted Block Design)

Appendix B

Supplementary Materials for "An Alternative Metric for Evaluating the Potential Patient Benefit of Response-Adaptive Randomization Procedures"

B.1 Prior Distributions for the Bayesian RAR Probability Models

In this paper, we employ a logistic regression probability model with weakly informative t-distribution priors on the regression coefficients. Suppose the hypothesized null response rate that we are interested in is 12%, as in the ARREST trial. The logistic regression probability model assumes the following prior distributions for β_0 and β_1: $\beta_0 \sim t_7(\log\left(\frac{0.12}{0.88}\right), 2.5) \perp\!\!\!\perp \beta_1$

$\sim t_7(0, 2.5)$, whereas the conventional independent-beta binomial probability model would,

as in Proper et al., [54] assume $\pi_E \sim \text{Beta}(0.24, 1.76)$. The induced prior distributions on π_E for both probability models on the probability and log-odds scales are displayed in the plot below. Note that the prior distribution associated with the IBB model is asymmetric on the log-odds scale. Also note that the prior distribution associated with the logistic regression model places less density on extreme values of the response probabilities.

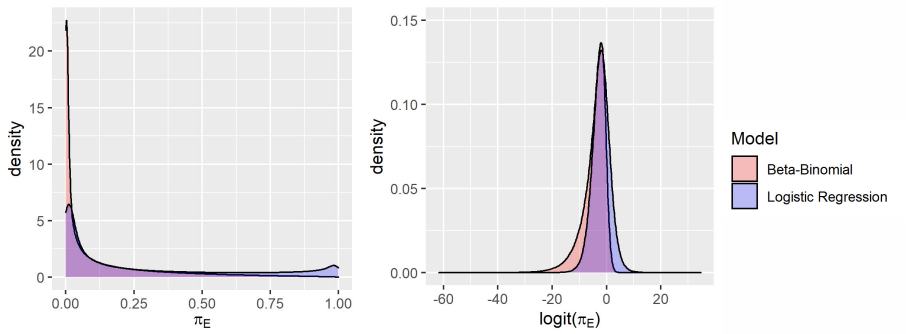


Figure B.1: Prior distributions for π_E on the probability and log-odds scale for the independent-beta binomial and logistic regression probability models.

B.2 Variations to Thompson Sampling

The figure below displays the p_j values arising from the following two adaptive modifications for $P_{E>C}$:

$$p_{rst,j} = \min\{0.75,\ \max\{0.25, P_{E>C}\}\} \tag{B.1}$$

$$p_{tun,j} = \frac{[P_{E>C}]^{C_j}}{[P_{E>C}]^{C_j} + [1 - P_{E>C}]^{C_j}} \tag{B.2}$$

The modification in (B.2) is plotted using $j = 1$, $j = 0.5J$ and $j = J$. Note that this modification is conservative in the beginning of the trial when it constrains p_j near 0.5, but becomes more aggressive later on.

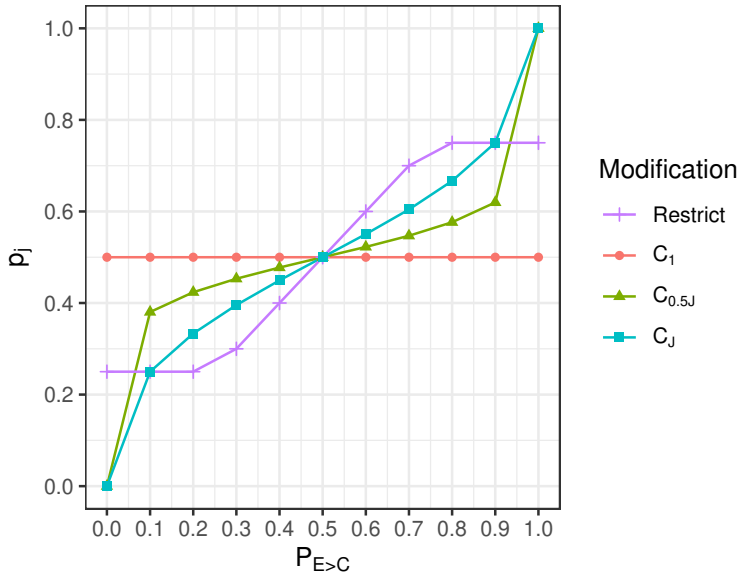


Figure B.2: Modifications to stabilize $P_{E>C}$ prior to randomization. C_j and "Restrict" refer to $p_{tun,j}$ and $p_{rst,j}$, respectively.

B.3 Gibbs Sampling

The posterior means of p_{ney} and p_{rsihr} are estimated using Gibbs draws and p_j is obtained as follows:

1. Draw T posterior samples for β_0 and β_1 via Gibbs sampling where T is a sufficiently large number (e.g., 2000) using the logistic regression probability model and the trial data up to and including group $j-1$.

2. Obtain T Gibbs draws for π_E and π_C using the identity $\tilde{\pi}_{k,i} = \text{expit}(\tilde{\beta}_{0,i} + \tilde{\beta}_{1,i} \cdot (\text{I}\{k = E\} - 0.5))$ $\forall i = 1, ..., T$, where $\text{expit}(x) = \frac{\exp x}{1+\exp x}$ and $\tilde{\beta}_{0,i}$ and $\tilde{\beta}_{1,i}$ are the i^{th} Gibbs draws for the regression coefficients.

3. Obtain T Gibbs draws for the optimal allocation ratio of interest, $\tilde{p}_{rsihr,i}$ or $\tilde{p}_{ney,i}$, by plugging in $\tilde{\pi}_{E,i}$ and $\tilde{\pi}_{C,i}$ for π_E and π_C in the expressions for p_{ney} and p_{rsihr}

$\forall i = 1, ..., T$.

4. Randomize the j^{th} group of participants to the treatment arm using:

$$p_j = \begin{cases} \frac{1}{T}\sum_{i=1}^{T} \tilde{p}_{ney,i}, & \text{for Neyman allocation} \\ \frac{1}{T}\sum_{i=1}^{T} \tilde{p}_{rsihr,i}, & \text{for RIHSR allocation} \\ g(r_{n-1}, \frac{1}{T}\sum_{i=1}^{T} \tilde{p}_i^*), & \text{for DBCD} \\ f(r_{n-1}, \frac{1}{T}\sum_{i=1}^{T} \tilde{p}_i^*), & \text{for ERADE} \end{cases} \tag{B.3}$$

where $\tilde{p}_i^*$ may equal $\tilde{p}_{ney,i}$ or $\tilde{p}_{rsihr,i}$ depending on the targeted allocation.

B.4 Finding the Maximum Sample Sizes and Posterior Probability Stopping Boundaries of the Bayesian RAR Group Sequential Designs

Below are instructions for finding the maximum sample size and posterior probability stopping boundaries that control the frequentist properties of a given Bayesian RAR group sequential design.

1. Determine the sample size required for the two-sample binomial proportion test to achieve the targeted power of the trial.

2. After determining an appropriate sequential group number and size, simulate 10,000 trials under the hypothesized null scenario using the logistic regression probability model with weakly informative t-distribution priors, the chosen adaptive modification to the randomization probability, and the chosen randomization procedure (e.g., the mass-weighted urn design).

3. When symmetric boundaries are desired, perform steps 3.1-3.4. Otherwise, proceed to step 4.

(a) Choose a Pocock-like or OBF-like boundary shape. For Pocock-like boundaries, set $a_j = a \ \forall j = 1, ..., J$ and find the value a for which 5% of simulated trials had either $P_{E>C}$ or $1 - P_{E>C}$ exceed a. For OBF-like boundaries, define a_j by $\Phi\left(\sqrt{J/j} \cdot c\right)$ and find the constant c for which 5% of simulated trials had either $P_{E>C}$ or $1 - P_{E>C}$ exceed a_j.

(b) Simulate 10,000 trials under the hypothesized alternative scenario using the same design as in step 2.

(c) Determine the proportion of simulated trials where $P_{E>C}$ exceeds a_j (i.e. power) found in step 3.1.

(d) Proceed to step 5.

4. When asymmetric boundaries are desired, perform the steps below. We only consider an OBF-like shape for the futility boundary in this paper.

(a) Choose a Pocock-like or OBF-like shape for the upper stopping boundary. For a Pocock-like upper boundary, set $a_j = \Phi(c) \ \forall j = 1, ..., J$. For an OBF-like upper boundary, set $a_j = \Phi\left(\sqrt{J/j} \cdot c\right) \ \forall j = 1, ..., J$.

(b) Set $b_j = \Phi\left(2c - \sqrt{J/j} \cdot c\right)$.

(c) Find the constant c for which 2.5% of the simulated trials had $P_{E>C} \geq a_s$ for some $j = s$ and $P_{E>C} > b_j \ \forall j = 1, ..., s - 1$.

(d) Simulate 10,000 trials under the hypothesized alternative scenario using the same design as in step 2.

(e) Determine the proportion of simulated trials where $P_{E>C} \geq a_s$ for some $j = s$ and $P_{E>C} > b_j \ \forall j = 1, ..., s - 1$ (i.e. power).

5. If the targeted power has not been met, repeat steps 2-4 using an updated sample size estimate. We recommend using the following identity to update the sample size estimate:

$$N_{new} = \text{ceiling}\left(N_{current} \cdot \frac{(\Phi(0.975)+\Phi(\text{Targeted Power}))^2}{(\Phi(0.975)+\Phi(\text{Current Power}))^2}\right)$$

where $N_{current}$, N_{new}, and Φ are the current sample size, updated sample size, and quantile function of the standard normal distribution, respectively.

B.5 Frequentist Group Sequential Designs

B.5.1 Pocock and O'Brien-Fleming Testing Procedures

Repeated significance testing in group sequential designs renders trials susceptible to substantial type I error inflation. In this section, we discuss two of the first sequential testing procedures proposed by Pocock [99] and O'Brien and Fleming [100] to control type I error at the nominal significance level.

Consider a two-arm trial with a binary primary outcome. Suppose accumulated data is analyzed after the responses for b new participants have been attained, and that a restricted randomization procedure is employed to guarantee that, on average, $\frac{b}{2}$ participants are allocated to the treatment and control within each group of size b. A group sequential design generally proceeds as follows. At each interim analysis j, where $j = 1, ..., J$, a standardized test statistic Z_j is computed using the accrued outcome data up to and including group j. The trial is terminated early for efficacy if $Z_j \geq a_j$ or for harm or futility if $Z_j \leq b_j$, where a_j and b_j denote upper and lower stopping boundaries, respectively. When $b_j < Z_j < a_j$, no decision is made and the trial continues to the next interim analysis. In this paper, we consider symmetric stopping boundaries (i.e. when $a_j = -b_j$) that allow early stopping for efficacy or harm and asymmetric stopping boundaries (i.e. when $a_j \neq -b_j$ and $a_J = b_J$) that allow early stopping for efficacy or futility.

Pocock [99] and O'Brien and Fleming [100] proposed setting a_j equal to $C_P(J, \alpha)$ and $C_B(J, \alpha)\sqrt{J/j}$, respectively, where $C_P(J, \alpha)$ and $C_B(J, \alpha)$ are functions of J that maintain type I error at the desired significance level, α. Pocock boundaries remain unchanged throughout the trial and engender a final critical value, a_J, that exceeds the critical value

of an α-level two-sided test. OBF boundaries facilitate conservative testing early on that becomes more aggressive as the trial proceeds through the use of an inflation factor, $\sqrt{J/j}$. At the final interim analysis, $\sqrt{J/j} = 1$ and the final critical value, a_J, is comparable to the critical value of an α-level two-sided test.

B.5.2 Tests for Two Independent Binomial Proportions

Numerous tests are available for comparing two independent binomial proportions and no one test is regarded as universally superior. [101] We evaluate the operating characteristics of various group sequential designs using an unconditional exact test, Fisher's exact test, [**?**] and the Chi-square test. [**?**] We focus on the unconditional exact test as it generally induces the most powerful analysis. Further discussion of Fisher's exact test and the Chi-square test is included in Section 1 of the Supporting Information.

Let $k =$ E and C denote the experimental treatment and control arms, respectively, and suppose there are y_k responders among the n_k participants assigned to treatment k. We are interested in testing H_0: $\pi_E = \pi_C$, where π_k are the true, but unknown, treatment arm response rates. An unconditional exact test analyzes 2x2 contingency tables conditioning only on n_E and n_C, as opposed to Fisher's exact test which assumes the marginal totals of the table are fixed. There are several variations of the unconditional exact test; however, we use the version that orders the sample space by the mid p-value from a one-sided Fisher's exact test. Letting $\mathbf{y}$ denote the set of observed data $\{y_E, y_C, n_E, n_C\}$, the unconditional exact test statistic is defined as:

$$T(\mathbf{y}) = \mathrm{F}(y_E, n_E, n_C, y_C + y_E) - 0.5 \cdot \mathrm{f}(y_E, n_E, n_C, y_C + y_E), \tag{B.4}$$
$$\text{where } \mathrm{f}(y_E, n_E, n_C, y_C + y_E) = \frac{\binom{n_E}{y_E}\binom{n_C}{y_C}}{\binom{n_E+n_C}{y_C+y_E}}$$

and $\mathrm{F}(\cdot)$ is the distribution function for the hypergeometric distribution. A two-sided exact p-value may be obtained using $P(\mathbf{y}) = \min\{1, 2 \cdot P_U(\mathbf{y}), 2 \cdot P_L(\mathbf{y}\}$, where $P_L(\mathbf{y})$ and $P_U(\mathbf{y})$ are

the one sided p-values for respectively testing $H_0 : \pi_E \leq \pi_C$ and $H_0 : \pi_E \geq \pi_C$, computed as follows: [62, 63]

$$P_L(\mathbf{y}) = \sup_{\pi:\pi_E \leq \pi_C} \Pr_\pi(T(\mathbf{Y}) \geq T(\mathbf{y})) \text{ and} \tag{B.5}$$
$$P_U(\mathbf{y}) = \sup_{\pi:\pi_E \geq \pi_C} \Pr_\pi(T(\mathbf{Y}) \leq T(\mathbf{y}))$$

We implement this test using the uncondExact2x2 function in the exact2x2 R package (version 1.6.5).

Appendix C

Supplementary Materials for "Network Meta Analysis to Predict the Efficacy of an Approved Treatment in a New Indication"

C.1 Prior Distributions for the Between-Studies Standard Deviation

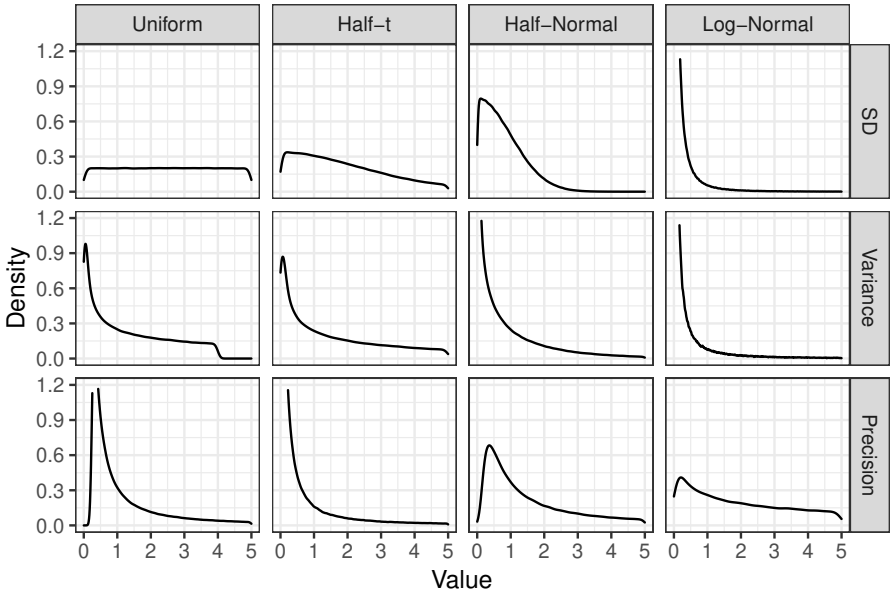


Figure C.1: The 4 considered prior distributions for the between-studies standard deviation (SD), σ: U(0,5), $\text{Ht}_7(0, 2.5)$, HN(0,1), and LN(-2.70,1.52). The induced prior distributions for variance (σ^2) and precision ($1/\sigma^2$) are also provided.

C.2 Data Generation

For each vector of basic parameters, $\boldsymbol{d} = (d^1_{12}, d^1_{13}, ..., d^1_{19}, d^2_{12}, d^2_{13}, ..., d^2_{19})^T$, we generated data according to the network diagram in Figure 1 of the main manuscript as follows.

1. Solve for $d^i_{b_j k}$ using the consistency equations of Lu and Ades $\forall k \in T_j, j = 1, ..., J_i$, and $i = 1, 2$.

2. Generate study-specific effects by sampling $\delta_{jb_jk} \sim \text{N}(d^i_{b_j k}, 0.25^2)$ $\forall k \in T_j, j = 1, ..., J_i$, and $i = 1, 2$.

3. Assume $p_{j1} = 0.05$ for indication 1 and $p_{j1} = 0.09$ for indication 2. Then, use the relation $d^i_{b_j k} = \log(p_{jk}/(1-p_{jk}) - \log(p_{jb_j}/(1-p_{jb_j})$ to solve for $p_{jb_j} \forall k \in T_j, j = 1, ..., J$.

4. Generate baseline effects by sampling $\mu_j \sim \mathrm{N}(\log(p_{jb_j}/(1-p_{jb_j}), 0.25^2))$ for $j = 1, ..., J$.

5. Solve for p_{jk} using $\mathrm{expit}(\mu_j + \delta_{jb_jk} \cdot \mathrm{I}(k \neq b_j))$ $\forall k \in T_j, j = 1, ..., J$.

6. Sample $r_{jk} \sim \mathrm{Bin}(n_{jk}, p_{jk})$ $\forall k \in T_j, j = 1, ..., J$, letting $n_{jk} = 60$ for studies corresponding to indication 1 and $n_{jk} = 38$ for studies corresponding to indication 2.

C.3 Additional Figures

C.3.1 Larger Indication and Drug Class Effect Sizes

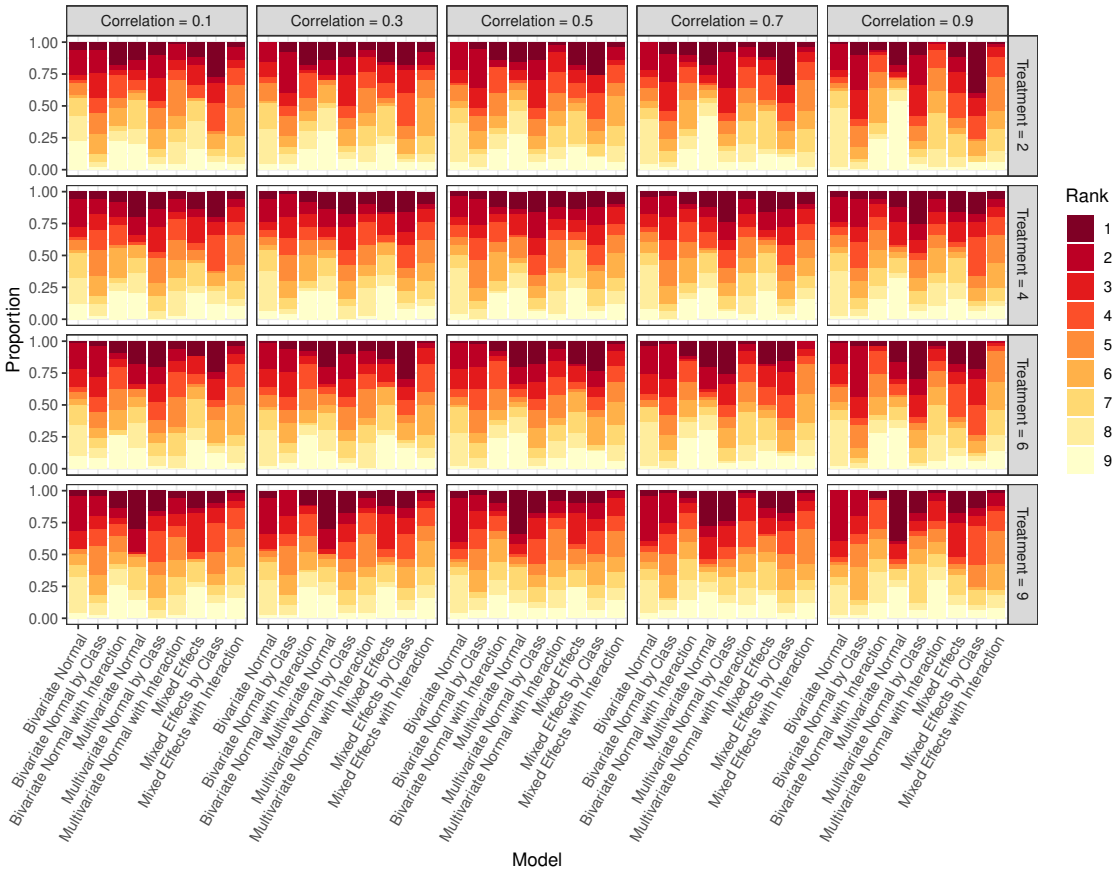


Figure C.2: Ranking distribution of RMSE of the posterior median of d_{1t}^2 arising from $\boldsymbol{d}$'s simulated using coefficient set 1. All models were fit using the $\text{Ht}_7(0, 2.5)$ prior for the between-study heterogeneity parameter and results are presented by treatment and $\text{Cor}(d_{1k}^1, d_{1k}^2)$.

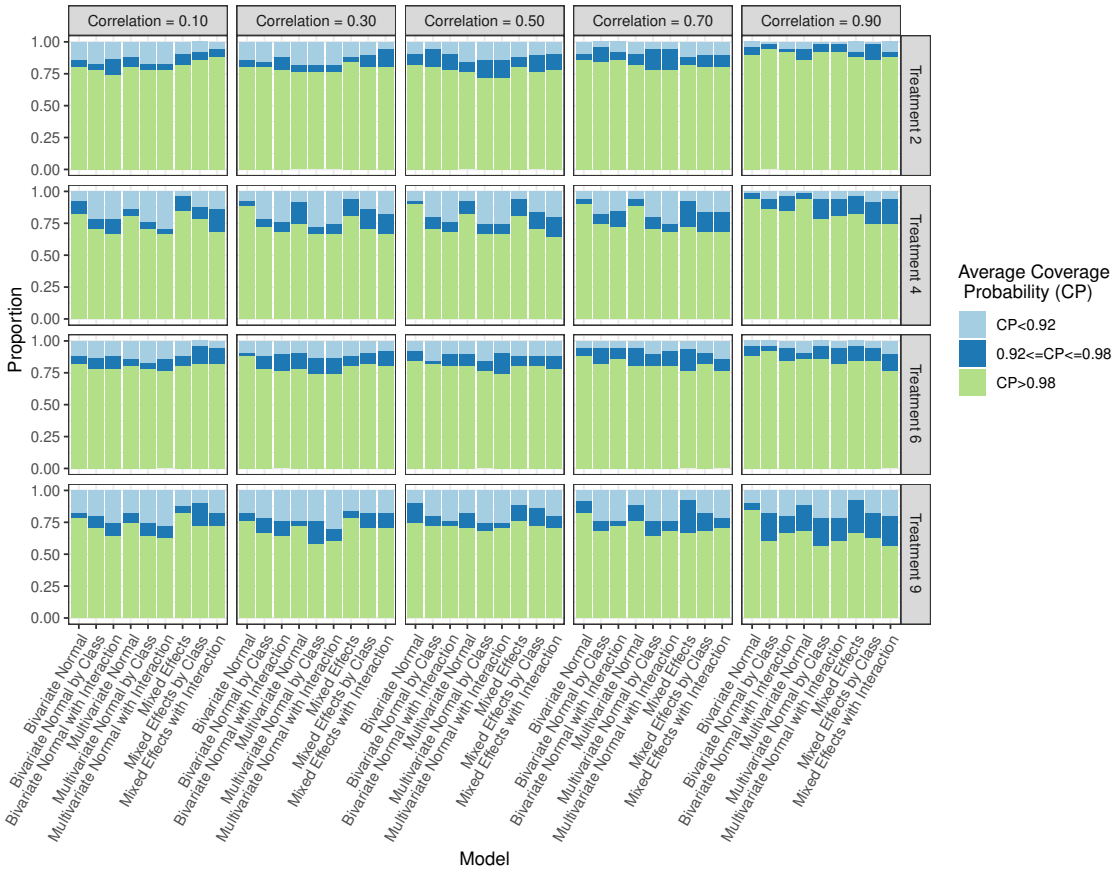


Figure C.3: Average coverage probability of the 95% credible interval for d_{1t}^2 arising from $\boldsymbol{d}$'s simulated using coefficient set 1. All models were fit using the $\text{Ht}_7(0, 2.5)$ prior for the between-study heterogeneity parameter and results are presented by treatment and $\text{Cor}(d_{1k}^1, d_{1k}^2)$.

C.3.2 Heterogeneous Class Effect

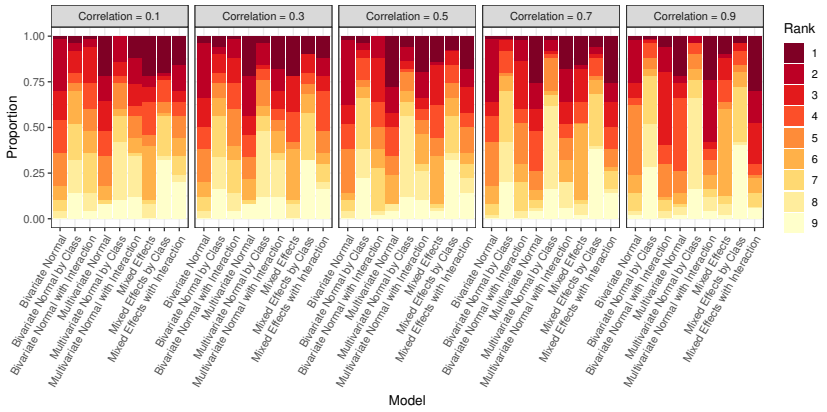


Figure C.4: Ranking distribution of RMSE of the posterior median of d_{12}^2 arising from $\boldsymbol{d}$'s simulated using a heterogeneous class effect. All models were fit using the $\text{Ht}_7(0, 2.5)$ prior for the between-study heterogeneity parameter and results are presented by $\text{Cor}(d_{1k}^1, d_{1k}^2)$.

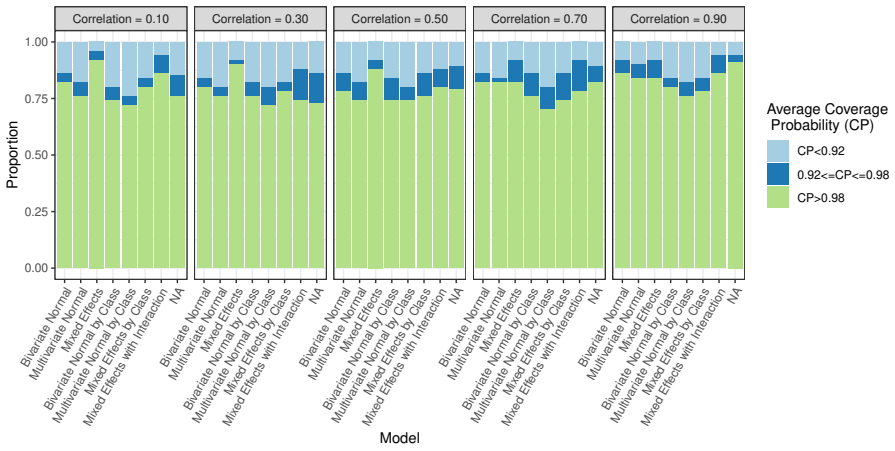


Figure C.5: Average coverage probability of the 95% credible interval for d_{1t}^2 arising from $\boldsymbol{d}$'s simulated using a heterogeneous class effect. All models were fit using the $\text{Ht}_7(0, 2.5)$ prior for the between-study heterogeneity parameter and resu2ts are presented by $\text{Cor}(d_{1k}^1, d_{1k}^2)$.

C.3.3 Smaller Indication and Drug Class Effect Sizes

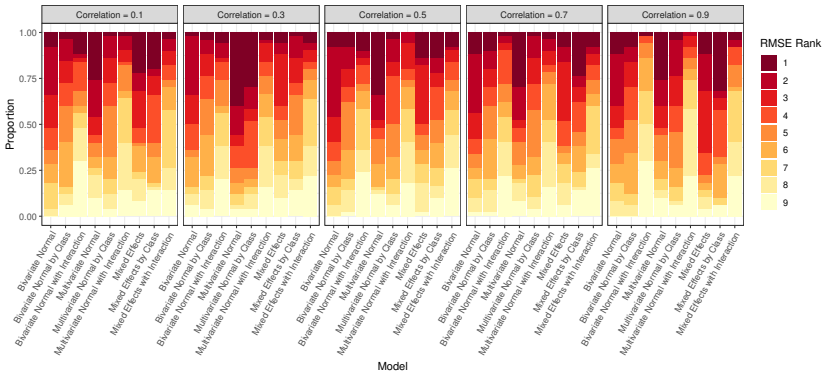


Figure C.6: Ranking distribution of RMSE of the posterior median of d_{12}^2 arising from $\boldsymbol{d}$'s simulated using smaller class and indication effects. All models were fit using the $\text{Ht}_7(0, 2.5)$ prior for the between-study heterogeneity parameter and results are presented by $\text{Cor}(d_{1k}^1, d_{1k}^2)$.

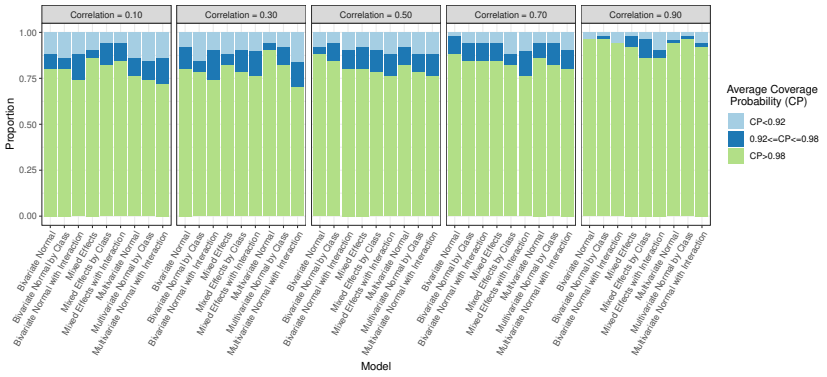


Figure C.7: Average coverage probability of the 95% credible interval for d_{1t}^2 arising from $\boldsymbol{d}$'s simulated using smaller class and indication effects. All models were fit using the $\text{Ht}_7(0, 2.5)$ prior for the between-study heterogeneity parameter and resu2ts are presented by $\text{Cor}(d_{1k}^1, d_{1k}^2)$.

C.3.4 Prior Distributions

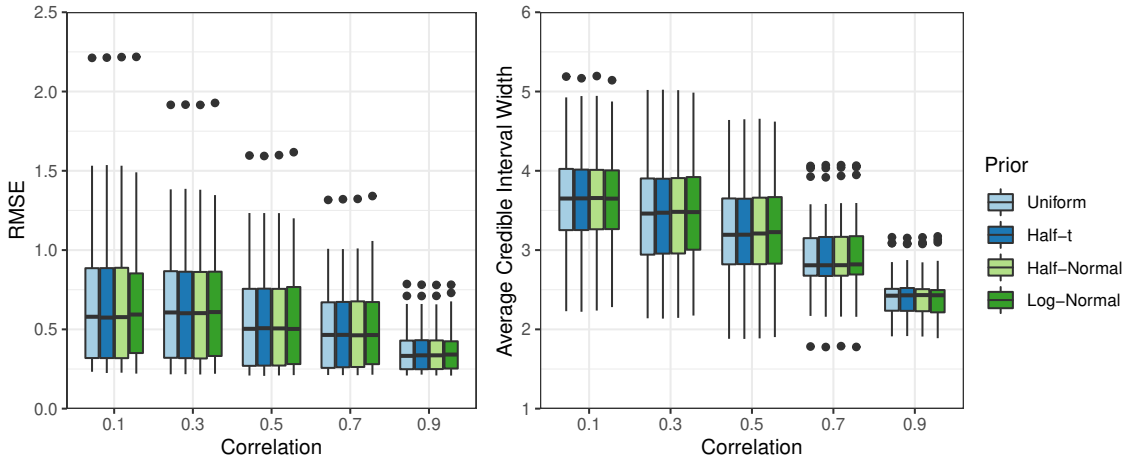


Figure C.8: RMSE of the posterior median and average 95% credible interval width of d_{12}^2 arising from $\boldsymbol{d}$'s simulated using coefficient set 1 by prior distribution for the between-study heterogeneity parameter. These results correspond to the mixed effects model incorporating a homogeneous class effect.

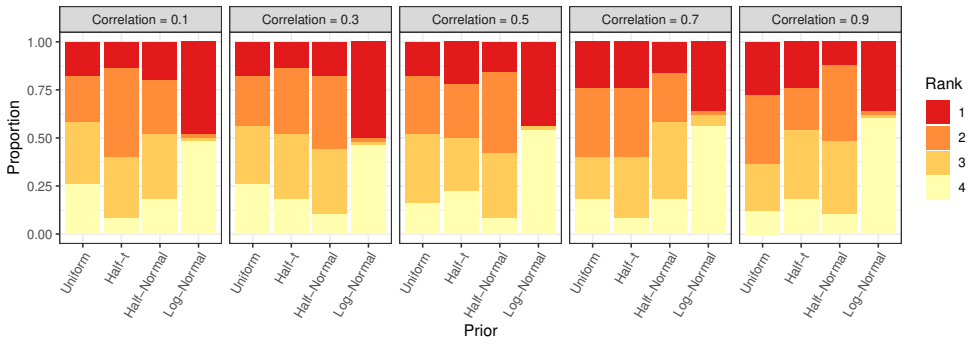


Figure C.9: Ranking distribution of RMSE of the posterior median of d_{12}^2 arising from $\boldsymbol{d}$'s simulated coefficient set 1 by prior distribution for the between-study heterogeneity parameter. These results correspond to the mixed effects model incorporating a homogeneous class effect

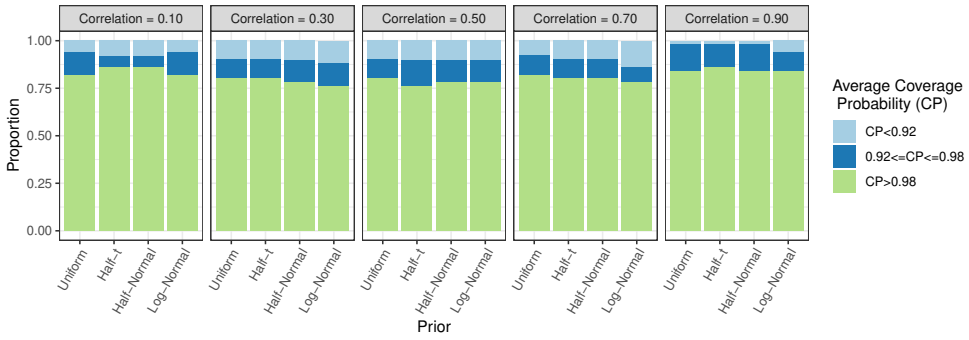


Figure C.10: Average coverage probability of the 95% credible interval for d_{1t}^2 arising from $\boldsymbol{d}$'s simulated coefficient set 1 by prior distribution for the between-study heterogeneity parameter. These results correspond to the mixed effects model incorporating a homogeneous class effect

C.4 Case Study

Table C.1: The posterior median of each basic parameter, d_{Ak}, arising from the standard RE-NMA model with the LN(-2.70,1.52) prior fitted to each indication separately. Parameter d_{AE} could not be estimated for psoriatic arthritis as treatment E has not been studied in this indication.

		Psoriasis	**Psoriatic Arthritis**
Drug Class 1	d_{AB}	2.23	1.95
	d_{AC}	3.03	2.48
Drug Class 2	d_{AD}	3.52	2.16
	d_{AG}	2.53	1.90
	d_{AJ}	3.02	2.35
Drug Class 3	d_{AE}	4.53	N/A
	d_{AF}	3.92	1.80
	d_{AH}	4.51	2.65
	d_{AI}	5.07	2.92

C.4.1 Probability of Success

We defined "success" as the probability of having a two-sided p-value less than 0.05 and computed PoS using:

$$\text{PoS} = \int_{\delta_{j^*AE}} \int_{d^2AE} \Pr(\text{Success}|\delta_{j^*AE}) \text{p}(\delta_{j^*AE}|d^2_{AE}, \sigma^2_2) \text{p}(d^2_{AE}|\{\boldsymbol{D_j}\}_{j=1}^{J})\, d\delta_{j1t} dd_{1t2} \quad \text{(C.1)}$$

where δ_{j^*AE} is the study-specific log-odds ratio comparing treatment E vs. A in a future trial, j^*. Because the double integral in (C.1) is intractable, we approximated PoS using MCMC simulation as follows:

1. Obtain B posterior samples for d^2_{AE} and σ^2_2 using the proposed mixed effects model with a heterogeneous class effect, where B is a sufficiently large number (e.g., 10000).

2. Obtain B draws from the posterior predictive distribution by sampling $\delta^{(b)}_{j^*AE} \sim \text{N}(d^{2(b)}_{AE}, \sigma^{2(b)}_2)$ for $b = 1, ..., B$.

3. For $b = 1, ..., B$, estimate the probability of response for treatment E, p_E, implied by $\delta^{(b)}_{j^*AE}$ using $\hat{p}^{(b)}_E = \frac{\exp(\delta^{(b)}_{j^*AE}) \cdot \hat{p}_A/(1-\hat{p}_A)}{1+\exp(\delta^{(b)}_{j^*AE}) \cdot \hat{p}_A/(1-\hat{A}_A)}$, where $\hat{p}_A$ is an estimate of the probability of response for treatment A obtained from external cohort studies or expert opinion. In this paper, we use $p_A = 0.09$.

4. For $b = 1, ..., B$, simulate a trial designed to detect a clinically meaningful treatment effect with $(1 - \beta)\%$ power and a type I error rate of α using $\hat{p}^{(b)}_E$ and $\hat{p}_A$ from step 3. Determine whether the study was successful (i.e. whether the null hypothesis was rejected).

5. Approximate PoS using $\text{PoS} = \frac{1}{B} \sum_{b=1}^{B} \text{I}\{b^{th} \text{ study successful}\}$.

C.5 Decision Aid for Selecting a Model in Practice

To select an appropriate model in practice, we recommend following a similar approach to that in our case study. That is, we recommend fitting a standard RE-NMA model to each indication separately and assessing the resulting posterior medians for the basic parameters. The absolute mean difference in posterior medians from different drug classes can be used to assess the nature of the drug class effect, whereas the empirical correlation between posterior medians from different indications can be used to evaluate the strength of the between-indication relationship. Then, a model can be selected using the decision aid in Figure C.11 below.

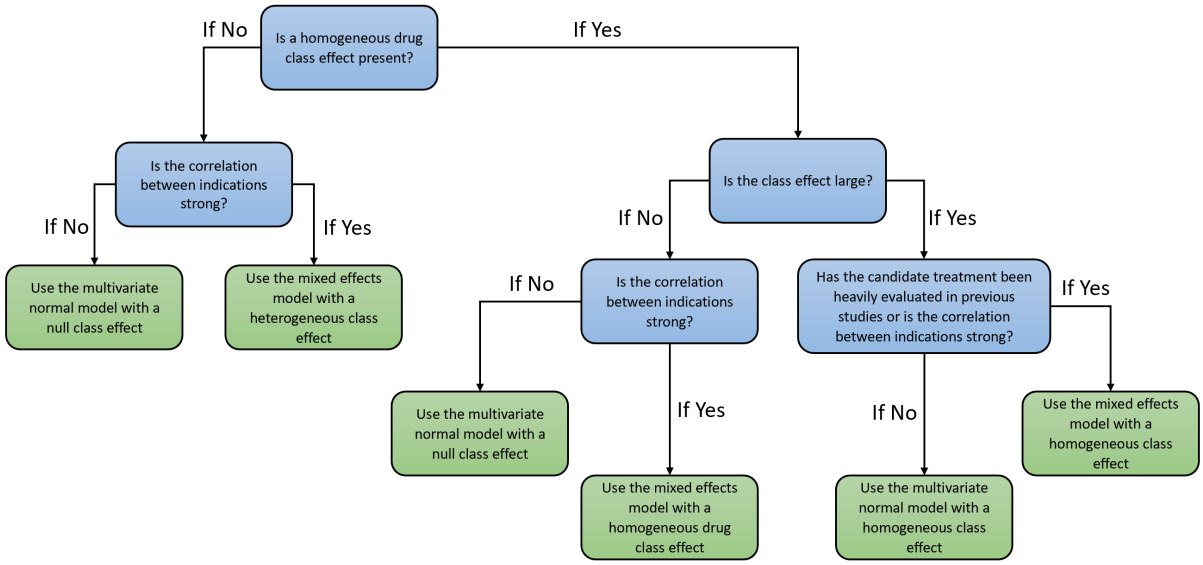


Figure C.11: Decision aid for selecting an appropriate model in practice.