Bayesian methodological extensions for comparative effectiveness, dose-response, and cluster randomized trials

By

Fengming Tang

Submitted to the graduate degree program in Biostatistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Committee Chair: Byron J. Gajewski, Ph.D.

_____

Jinxiang Hu, Ph.D.

_____

Milind Phadnis, Ph.D.

_____

Jo Wick, Ph.D.

_____

Susan Carlson, Ph.D.

Date Defended: 24 May 2021

The dissertation committee for Fengming Tang certifies that this is
the approved version of the following dissertation:

Bayesian methodological extensions for comparative effectiveness, dose-response, and cluster randomized trials

_____
Chair: Byron J. Gajewski, Ph.D.

_____
Graduate Director: Jo Wick, Ph.D.

Date Approved: 2 June 2021

Abstract

In this dissertation, we explored three Bayesian methodological extensions, including an adaptive Bayesian design featuring participant reuse for comparative effectiveness clinical trials, an innovative Bayesian dose-response EMAX model for a mixture of normal distributions, and a Bayesian analysis of weight loss for a cluster randomized clinical trial.

We first developed an adaptive Bayesian clinical trial design in the setting of comparative effectiveness clinical research where multiple treatments are of interest and the accrual rate is slow. Our proposed design mimics the real-world clinical practice that allows patients to switch treatments when the desired outcome is not achieved. As a result, each participant can have more than one observation, and hence it is possible to control for participant-specific variability which in turn results in a reduced number of participants needed. Additionally, response adaptive randomization is employed to improve trial efficiency by allocating more participants to the promising arms.

We also developed an innovative Bayesian dose-response EMAX mixture model incorporating finite mixture distributions into the EMAX framework. It is the first time that an EMAX model being extended to a finite mixture distribution. The model was motivated by a proposal investigating the dose effect of DHA supplementation on preterm birth rate (< 37 weeks of gestation), where gestational age was analyzed as continuous with a normal mixture distribution. We compared our proposed EMAX mixture model with an EMAX logistic model and an independent doses logistic model for a dichotomized endpoint using extensive simulations. Across the scenarios under consideration, the EMAX mixture model achieved higher power in detecting the effect of DHA supplementation on the PTB rate. It also resulted in smaller mean squared errors (MSE) in PTB rate estimates.

Lastly, we reanalyzed the percent weight loss data from Rural Engagement in Primary Care for Optimizing Weight Reduction (REPOWER), a cluster randomized clinical trial, using a Bayesian hierarchical model. We showed that the Bayesian approach can derive probability estimates of direct clinical interest and can provide additional insights into data interpretation by utilizing posterior distributions for parameters of interest. We also demonstrated that the Bayesian approach can easily handle complex problems using the same statistical framework.

Acknowledgements

I would like to first thank my advisor, Dr. Byron J. Gajewski, who helped me formulate the research questions and methodologies. Without his valuable guidance and support I would not have been able to complete my dissertation.

I also want to thank Dr. Jo Wick, my graduate director. She continuously supported my growth and success and made efforts to accommodate my needs and answer my questions. I also thank her for her insightful suggestions on the paper we coauthored.

Additionally, I would like to thank the committee members, Dr. Jinxiang Hu, Dr. Milind Phadnis, and Dr. Susan Carlson. I thank them for their time and their guidance.

Furthermore, I would also like to thank my colleagues from Saint Luke's Hospital, especially Dr. Mikhail Kosiborod and Mr. Philip Jones. They gave me a lot of flexibility and support, without which it would not have been possible for me to complete my dissertation while working full time.

Finally, the work in chapter 2 was supported in part by a NIH Clinical and Translational Science Award (UL1TR002366) to the University of Kansas, and KUMC Biostatistics & Data Science Department, as well as The University of Kansas Cancer Center (P30 CA168524). The work in chapter 3 was partially supported by NIH Clinical and Translational Science Award UL1TR002366. The work in chapter 4 was funded through Patient-Centered Outcomes Research Institute (PCORI) award OTO-1402-09413 as well as by The University of Kansas Cancer Center Support Grant (CCSG) awarded by the National Cancer Institute (P30 CA168524). R package brms was used in preparing data and generating STAN code.

Table of Contents

List of Figures

List of Tables

Chapter 1 : Introduction

Although the frequentist paradigm has been the predominant approach to clinical studies in the past several decades, some limitations associated with the frequentist null hypothesis testing that reports dichotomized P values have been recognized in statistic society (1,2). On the other hand, the Bayesian paradigm derives probability estimates of model parameters reflecting the clinical interest and can provide better data interpretation. It has gained popularity in recent years owing to the advancement in powerful computing capacity and the invention of efficient Bayesian statistic software. In this dissertation, we explored three Bayesian methodological extensions, including an adaptive Bayesian design featuring participant reuse for comparative effectiveness clinical trials, an innovative Bayesian dose-response EMAX model for a mixture of normal distributions, and a Bayesian analysis of weight loss for a cluster randomized clinical trial.

In chapter 2, we developed an adaptive Bayesian clinical trial design in the setting of comparative clinical research where multiple treatments are of interest and the accrual rate is slow. One of the biggest challenges in designing clinical trials for rare diseases is the slow accrual rate. This challenge is amplified in comparative effectiveness research where multiple treatments are compared to identify the treatment that works best for improving health. Motivated by real-world clinical practice that allows patients to switch therapies if the desired outcome is not achieved, we proposed a design that reuses participants. In our design, participants are randomized to one study drug as the initial treatment. If the participant responds to the initial treatment, then the participant completes the study and no more treatment will be assigned to the participant. On the other hand, if the participant does not respond to the initial treatment, the participant will be assigned a new treatment from the remaining therapies. This

process is repeated until either the desired treatment outcome is achieved, or all study treatments are given to the participant. With efficiency in mind, we further improve the reusing participants design by employing a Bayesian adaptive design. The basic idea is to utilize response adaptive randomization (RAR) to assign more participants to the most promising arms, by updating the randomization probability using interim analyses. The reusing participants RAR design starts by randomizing participants with equal probability to one of the study treatments. As enrollment continues, interim analyses will be performed according to a pre-specified schedule. The data available at the interim analyses will be used to calculate the posterior probabilities of treatments being the most effective, which will then be used to update the RAR allocation rates for future participants. Extensive simulations were used to compare this design with a conventional adaptive clinical design where each participant is randomized to one treatment only, a non-adaptive design that reuses participants, and a non-adaptive design that does not reuse participants.

In chapter 3, We developed an innovative Bayesian dose-response EMAX mixture model that incorporates finite mixture distributions into the EMAX framework. The model was motivated by a proposal investigating dose effect of DHA supplementation on preterm birth rate. One frequently used dose-response model is the pairwise independent doses model. In this model, no functional relationship is assumed between the dose and effect, and all doses are modeled independently and compared with each other. The independent doses model is often inefficient and results in lower power because of its lack of functional relationship assumption. When the dose-response relationship can be assumed monotonic, an EMAX (MAXimum Effect) model has been shown to provide a good empirical fit for designing and analyzing dose-response data across a wide range of pharmaceutical studies. The EMAX model assumes the dose-

response relationship follows a nonlinear monotonic function with a parameter representing the maximum effect that can be achieved when the dose approaches infinity and another parameter representing the dose that achieves 50% of the maximum effect. One option to evaluate the DHA dose effect on PTB is to apply the EMAX model treating PTB as a dichotomous endpoint. However, studies have shown that dichotomizing continuous endpoints resulted in a loss of information and reduced power (3,4,5). We propose a Bayesian EMAX model that analyzes gestational age as continuous. Schwartz et al. showed that the distribution of gestational age can be described by a mixture of three normal distributions (6). Thus, we developed our EMAX mixture model for a continuous endpoint with a mixture distribution. We compared our model with two models that dichotomize gestational age: the EMAX model (EMAX logistic model) and the independent doses logistic model. Extensive simulations showed that the EMAX Mixture model achieved a higher power for detecting the DHA dose effect on PTB than the other two models and resulted in smaller mean squared errors (MSEs) in estimates of PTB rates. Additionally, the EMAX Mixture model is attractive because it allows for statistically efficient estimates of PTB rates using different gestational age cut-points within the same parsimonious model. For example, we can estimate the rate of early preterm birth (<34 weeks gestation), preterm birth (<37 weeks gestation), and late-term birth (>41 weeks gestation) using the same model.

In chapter 4, we reanalyzed the weight loss data from Rural Engagement in Primary Care for Optimizing Weight Reduction (REPOWER) clinical trial using a Bayesian approach. Repower is a cluster randomized clinical trial conducted to compare three delivery models of Intensive Behavioral Therapy for Obesity (IBT): the fee-for-service individual delivery model, the in-clinic group visits model, and the phone-based group visits model. Participant weight was

measured at baseline, 6, 18, and 24 months by trained staff. Frequentist methods were used to compare the three delivery models in the original analysis. In this dissertation, we first analyzed the percent weight loss over time using a Bayesian three-level hierarchical model to answer questions such as what the probability of obtaining a greater weight loss in the in-clinic group visits arm vs. the individual visits arm is. we also used a four-level hierarchical model with an additional level to assess the group assignment impact on the effect of delivery models on weight loss.

References

1. Greenland S, Senn SJ, Rothman KJ, et al. Statisticaltests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337-350. doi: 10.1007/s10654-0160149-3 20.
2. Wasserstein R, Lazar N. The ASA's statement on P values: context, process, and purpose. Am Stat. 2016; 70(2):129-133. doi: 10.1080/00031305. 2016. 1154108
3. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ (Clinical research ed). 2006; 332:1080.
4. Deyi BA, Kosinski AS, Snapinn SM. Power considerations when a continuous outcome variable is dichotomized. Journal of biopharmaceutical statistics. 1998; 8:337–52.
5. Peacock JL, et al. Dichotomising continuous data while retaining statistical power using a distributional approach. Statistics in medicine. 2012; 31:3089–103.
6. Schwartz, S., Gelfand, A., and Miranda, M. , "Joint Bayesian Analysis of Birthweight and Censored Gestational Age Using Finite Mixture Models," Statistics in Medicine. 2010; 29: 1710–1723.

Chapter 2 Comparative Effectiveness Research using Bayesian Adaptive Designs for Rare Diseases: Response Adaptive Randomization Reusing Participants

2.0 Abstract

Slow accrual rate is a major challenge in clinical trials for rare diseases and is identified as the most frequent reason for clinical trials to fail. This challenge is amplified in comparative effectiveness research where multiple treatments are compared to identify the best treatment. Novel efficient clinical trial designs are in urgent need in these areas. Our proposed response adaptive randomization (RAR) reusing participants trial design mimics the real-world clinical practice that allows patients to switch treatments when desired outcome is not achieved. The proposed design increases efficiency by two strategies: 1) Allowing participants to switch treatments so that each participant can have more than one observation and hence it is possible to control for participant specific variability to increase statistical power; and 2) Utilizing RAR to allocate more participants to the promising arms such that ethical and efficient studies will be achieved. Extensive simulations were conducted and showed that, compared with trials where each participant receives one treatment, the proposed participants reusing RAR design can achieve comparable power with a smaller sample size and a shorter trial duration, especially when the accrual rate is low. The efficiency gain decreases as the accrual rate increases.

2.1 Introduction

One of the biggest challenges in designing clinical trials for rare diseases is slow accrual rate. Recent studies show that slow accrual rate is a significant hurdle in advancing the translation of clinical discoveries (1), and a poor patient accrual is identified as the most frequent reason for clinical trials to be classified as "fail to complete" (2). This challenge is amplified in comparative effectiveness research where multiple treatments are compared to identify the treatment that works best for improving health. Frequently, investigators have to reduce the

number of arms because sufficient patients cannot be enrolled in a reasonable length of study duration. Novel efficient clinical trial designs are in urgent need in these areas.

In conventional parallel randomized clinical trial designs, participants are randomized to one study treatment and each participant contributes one observation regardless of the participant's outcome. However, in real world clinical practice, patients often switch therapies if the desired outcome is not achieved. This motivated our proposal of reusing participants in a clinical trial design. In our design, participants are randomized to one study drug as the initial treatment. If the participant responds to the initial treatment, then the participant completes the study and no more treatment will be assigned to the participant. On the other hand, if the participant does not respond to the initial treatment, the participant will be assigned a new treatment from the remaining therapies. This process is repeated until either the desired treatment outcome is achieved, or all study treatments are given to the participant. The advantage of the proposed design is that it mimics the real-world clinical practice and it can achieve the desired power with fewer participants.

The proposed reusing participants design is an extension to the two-arms crossover trial for absorbing binary endpoint proposed by Nason and Follmann (3). An absorbing binary endpoint is an outcome that cannot be repeated in the second period if it occurs in the first period, such as mortality or pregnancy in infertility studies. In our proposed design, responding to a treatment is analyzed as an absorbing binary endpoint and participants will not switch to a new treatment if the desired outcome is achieved. It is more ethical than the conventional crossover design which requires participants to receive all candidate treatments in sequence (4) and results in participants who receive effective treatments first to cross-over to ineffective treatments.

With efficiency in mind, we further improve the reusing participants design by employing a Bayesian adaptive design. Bayesian adaptive designs have been broadly accepted to be able to increase efficacy, reduce duration, and provide more ethical clinical trials (5). The basic idea is to utilize response adaptive randomization (RAR) to assign more participants to the arms that are most promising, by updating the randomization probability using interim analyses. The reusing participants RAR design starts by randomizing participants with equal probability to one of the study treatments. As enrollment continues, interim analyses will be performed according to a pre-specified schedule. The data available at the interim analyses will be used to calculate the posterior probabilities of treatments being the most effective, which will then be used to update the RAR allocation rates for future participants. It is worth noting that, in order to avoid overly complicating trial operations, the RAR randomization only applies to the initial treatment of each participant. Once the initial treatment is determined for a participant, the order of subsequent treatments will be determined using sampling without replacement from the remaining study treatments. Participants will receive treatments until they achieve the desired outcome or until the they go through all the study treatments.

We will compare this design, called Reuse-RAR, with conventional adaptive clinical design (Conventional-RAR) where each participant is randomized to one treatment only, such as the design described by Gajewski et al. (6). In addition, we will also compare the Reuse-RAR design with a non-adaptive design that reuses participants (Reuse-noRAR) and a non-adaptive design that does not reuse participants (Conventional-noRAR) . The remainder of this article is arranged as follows. In section 2.1, we first describe the motivation study and give an overall summary for each of the four designs in the context of the motivation study. In section 2.2, we describe the statistical models for designs that reuse participants (i.e., Reuse-noRAR and Reuse-

RAR) and designs that do not reuse participants (i.e., Conventional-RAR and Conventional-noRAR) separately. Section 2.3 – 2.9 cover accrual rate patterns, interim analysis schedule, response adaptive randomization, virtual response rate, success criteria and model calibration, carryover effect and period effect, and simulation. They are applied to both Reuse-RAR and Conventional-RAR designs to ensure a fair comparison is made. Extensive simulations are used to compare operating characteristics of the designs including power, duration of study, number of participants required. The results are summarized in section 3. In section 4, we draw conclusion from our analysis and discuss the advantages and limitations of our proposed Reuse-RAR design. Section 5 is discussion and future work.

2.2 Methods

2.2.1 Trial summary

To illustrate the method, we use the setting of Patient Assisted Intervention for Neuropathy: comparison of Treatment in Real Life Situations (PAIN-CONTRoLS) (6,7), a comparative effectiveness clinical trial studying four treatments for cryptogenic sensory polyneuropathy (CSPN). CSPN, also known as idiopathic polyneuropathy, is a diagnosis made when all known causes of neuropathy have been ruled out. Although CSPN accounts for 10 – 30% of all polyneurophy cases (8), very few trials have been conducted to study the treatments of CSPN. There is an urgent need for evidence generating trials to guide physicians treating CSPN patients (7). PAIN-CONTRoLS is one of the first such trials. The primary endpoint is evaluated using visual analog scale pain score (VAS) (9). A subject is considered a responder if the VAS score drops by 50% or more after being on a treatment for 12 weeks. The goal of the study is to identify which drug is the most effective in reducing pain with fewest side effects. Although the actual PAIN-CONTRoLS trial had four arms we consider a "what-if" trial with five

arms and for simplicity we assume a binary endpoint rather than the trinary endpoint used in PAIN-CONTRoLS. Below is a summary for the four designs (Conventional-noRAR, Conventional-RAR, Reuse-noRAR, and Reuse-RAR) in the context of PAIN-CONTRoLS trial.

In the Reuse-RAR design, participants are randomized to one of the five treatments as their initial treatment. At first the study uses equal randomization, which is then updated using RAR after the first interim analysis. The order of the subsequent treatment assignments is determined by sampling without replacement from the four remaining treatments. After 12 weeks, depending on the VAS score measurements, the participants may be given the next treatment in line if the desired effect is not achieved, or be considered as a responder and complete the trial. Each participant can have multiple observations (between one and five).

In the Conventional-RAR design, participants are randomized to one of the five treatments using the same sample randomization scheme as Reuse-RAR, however, no additional treatments will be assigned beyond the first treatment. Each participant can only have one observation.

In both Reuse-RAR design and Conventional-RAR design, interim analyses will be performed according to a pre-specified schedule. At each interim, all current data will be analyzed, and the treatment allocation rates will be updated so that more participants will be allocated to the arm with the maximal effect.

In the Reuse-noRAR design, participants are randomized to one of the five treatments as their initial treatment using equal randomization. The order of the subsequent treatment assignments is determined by sampling without replacement from the four remaining treatments.

No interim analyses will be performed, and the allocation rates will stay the same for the whole study. Like the Reuse-RAR design, each participant can have one to five observations.

In the Conventional-noRAR design, participants are randomized to one of the five treatments using equal allocation rates. Each participant will have one observation. No interim analysis will be performed.

2.2.2 Statistical models

In all four designs, we assume the five treatments are not ordered in any explicit manner.

In Conventional-RAR and Conventional-noRAR design, each participant has exactly one observation. These designs will use an independent logistic model, described in section 2.2.1.

In Reuse-RAR and Reuse-noRAR design, each participant can have more than one observation. The observations from the same participant are correlated due to participant variation (or participant disease severity), which is modeled by including participant as a random effect in a hierarchical logistic model (also known as linear mixed model). A normal, hierarchical prior on the logit scale is used for the participant effect. This approach is similar to that of Nason and Follmann(3), where participant variation was modeled using Beta distribution. Furthermore, we assume that the carryover effect is consistent across participants and treatments. A single carryover factor is used to model the amount of effect that is carried over from previous period. Furthermore, a period effect can also be incorporated to account for the effect of period. Model details are described in 2.2.2.

2.2.2.1 Independent logistic model (for Conventional-RAR design and Conventional-noRAR design)

Let $\boldsymbol{x}_i$ be a 5-element binary vector indicating the treatment participant i received. For example, $x_i = (0,0,1,0,0)$ indicates participant i received the 3rd treatment. Let y_i be the binary outcome variable (0 non-responder, 1 responder). Assuming y_i follows a $Bernoulli(p_i)$ distribution, where

$$logit(p_i) = \boldsymbol{x_i\beta}$$

$\boldsymbol{\beta} = \{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ is a 5-element vector denoting the treatment effect on logit scale. $\theta_j = \frac{exp(\beta_j)}{1+exp(\beta_j)}$ is the probability of being a responder for a participant received treatment j. A vague normal prior, N(0, 5^2) is assigned to each β_j. When transformed back to probability scale using anti-logit function, the vague prior gives a 95% equal-tailed interval of (0.001, 0.999). Hamiltonian Monte Carlo (10,11) is used to obtain the posterior distribution for $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5|\boldsymbol{y}\}$. The best arm is defined as $j_{max} = \underset{j\in(1,2,3,4,5)}{\arg\max}\ (\beta_j)$. The probability of being the best arm for arm j is denoted as $prob(j = j_{max}|\boldsymbol{y})$.

2.2.2.2 Hierarchical logistic model (for Reuse-RAR design and Reuse-noRAR design)

Let T_i be the number of periods for participant i and let $t \in \{1, 2, \dots, T_i\}$ denote the period index. Let $\boldsymbol{x}_{it}$ be a 5-element binary vector indicating the treatment participant i received during period t. For example, $x_{it} = (0,0,1,0,0)$ indicates participant i received the 3rd treatment during period t. y_{it} is the binary outcome variable for participant i during period t (0 for non-responder and 1 for responder), and follows a $Bernoulli(p_{it})$ distribution, where

$$logit(p_{it}) = \boldsymbol{x_{it}\beta} + \epsilon_i$$

$\boldsymbol{\beta} = \{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ is a 5-element vector denoting the treatment effect for the 5 treatments on logit scale. ϵ_i denotes the participant-specific effect (i.e. participant disease severity) on logit scale and it follows a normal distribution: $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. For priors, an independent normal distribution $N(0,5^2)$ is used for β_j and a truncated normal distribution $N(0,3^2)$ is used for σ_ϵ^2.

To accommodate the carryover effect, the model can be expanded as follows,

$$logit(p_{it}) = \begin{cases} \boldsymbol{x_{it}\beta} + \epsilon_i, & t = 1 \\ \boldsymbol{x_{it}\beta} + \pi * \boldsymbol{x_{i(t-1)}\beta} + \epsilon_i, & t > 1 \end{cases}$$

Where π is the carryover factor which models the proportion of treatment effect that persists from one treatment to the next. A prior of $N(0,0.5^2)$ is used for π.

In cases where study period has important impact on treatment outcome, we can further expand the model as follows,

$$logit(p_{it}) = \begin{cases} \boldsymbol{x_{it}\beta} + \epsilon_i, & t = 1 \\ \boldsymbol{x_{it}\beta} + \pi * \boldsymbol{x_{i(t-1)}\beta} + f(t) + \epsilon_i, & t > 1 \end{cases}$$

where $f(t)$ is a function of t, which can be chosen to model a potential period effect. It can be a polynomial function or a function representing flexible cubic splines. For example, a linear function is a reasonable choice for the PAIN-CONTRoLS study due to the small number of periods,

$$\boldsymbol{f}(t) = \gamma_1 t$$

where γ_1 is the regression coefficient. A vague $N(0,5^2)$ prior is used for γ_1.

Hamiltonian Monte Carlo is used to obtain the posterior distribution for model parameters : $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \pi, \gamma_1 \mid \boldsymbol{y_1}, \boldsymbol{y_2}, .. \boldsymbol{y_n}\}$. The best arm is defined as $j_{max} = \underset{j \in (1,2,3,4,5)}{\arg\max}\ (\beta_j)$. The probability of being the best arm for arm j is denoted as $prob(j = j_{max})$.

2.2.3 Accrual rate patterns

We assume the distribution of participant accrual patterns follows a Poisson distribution. In order to investigate the operating characteristics of the trial designs, we run simulations using four different rates: 1.5 participants per week, 3 participants per week, 4.5 participants per week, and 6 participants per week. If a participant is reused in a trial, we assume no waiting time between their last visit and randomization to the next study drug.

2.2.4 Interim analysis schedule

For the Reuse-RAR design, each participant can have multiple observations (between 1 and 5) with each observation corresponding to a treatment the participant received. We will use number of observations initiated instead of number of participants enrolled to describe sample size. The interim analyses will be conducted when 300, 500, and 700 observations are initiated. Only observations with assessable endpoints (being on a treatment for 12 weeks and with an additional 4 weeks of lag for collecting and observing endpoints) will be included in the interim analyses. A final analysis will be conducted when 900 observations are assessable. Conventional-RAR design will have the same interim and final analysis schedule but in terms of participants enrolled. For the Reuse-noRAR and Conventional-noRAR design, no interim analysis will be performed. A final analysis will be conducted after assessable endpoint is available for 900 observations in Reuse-noRAR design and 900 participants in Conventional-noRAR design.

2.2.5 Response Adaptive Randomization (RAR)

For Conventional-RAR and Reuse-RAR, at each interim, the randomization probability needs to be updated to allocate more future participants to the most promising arms. There are many choices of the formula for the randomization probability. For example, one choice is proportional to the posterior probabilities that the arms have maximum effect (i.e., $\Pr(j = j_{max})$). We use the information formula for RAR allocations (6), $V_j = \sqrt{\frac{\Pr(j=j_{max})\, Var(\theta_j)}{n_j+1}}$, where n_j is the number of participants whose initial treatment is drug j and $var(\theta_j)$ is the sample variance of $\theta_j|\boldsymbol{y_j}$, and $\frac{var(\theta_j)}{n_j+1}$ is the expected change in variance (a proxy for information gained). This approach balances the goal of randomizing to the arm with the maximum effect and the design to gain new information by allocating to under explored arms..

2.2.6 Virtual response rate

Virtual response rate is the true efficacy rate used to generate participant response in simulations. We label $\boldsymbol{\theta}^T = (\theta_1^T, \theta_2^T, \theta_3^T, \theta_4^T, \theta_5^T)$ the virtual response rates for the five study treatments. If we assume there is no participant variation, which means participant outcome is solely determined by the treatment the participant received, the sampling distribution of the participant outcome is $y_{it} \sim Bernoulli(\theta_j^T)$, where j represents the treatment participant i received during period t. For the purpose of the study, we investigated three scenarios. The first scenario assumes $\boldsymbol{\theta}^T = (0.2, 0.2, 0.2, 0.2, 0.2)$, where all treatments are equivalent. In comparative effectiveness setting, the scenarios where all treatments are equivalent is the null scenarios. This is the null scenario, denoted by H_0. The second scenario assumes $\boldsymbol{\theta}^T = (0.3, 0.3, 0.3, 0.4, 0.5)$, where treatment 5 is the most effective and treatment 4 is the second effective. We denote it by H_1. The third scenario assumes $\boldsymbol{\theta}^T = (0.3, 0.3, 0.3, 0.5, 0.5)$, where

treatment 5 and treatment 4 are equally effective. We denote it by H_2. H_1 and H_2 are two alternative scenarios.

In the real world, it is not realistic to assume there is no participant variation. The observations from the same participant are usually more alike than those from different participants. We assume, on logit scale, participant variation follows a normal distribution: $\epsilon_i^T \sim normal(0, \sigma_{\epsilon^T}^2)$. And the sampling distribution of participant outcome becomes $y_{it} \sim Bernoulli\left(logit^{-1}\left(logit\left(\theta_j^T\right) + \epsilon_i^T\right)\right)$. We use $\sigma_{\epsilon^T}^2$ = 0.25 in simulations, which can be translate to an ICC (intraclass correlation coefficient) of 0.07.

2.2.7. Success criteria and Model calibration

At the final analysis, an arm may be declared superior if its posterior probability of being the best arm meets a pre-specified success criterion, i.e. $\Pr(j = j_{max}) > \delta$. Type I error is the proportion of simulations that meet the success threshold in null scenarios (12). Power is the proportion of simulations that meet the success threshold in alternative scenarios. Below we will discuss how to prespecify the success thresholds (δ).

In order to make the designs comparable, success thresholds (δ) are chosen to achieve similar type I error rates across designs using simulations in null scenarios. For example, figure 2.1 is the plot for the proportion of success (i.e. Type 1 error) by threshold (δ) based on simulations using the Conventional-noRAR design in the null scenario when $\lambda = 1.5$. As the threshold increases, the proportion of simulations meet the success criterion (i.e. type I error) decreases. When the threshold is 0.829, the type I error rate is roughly 4.9%. Using the same method, we identified δ is to be 0.829, 0.794, 0.832, and 0.827 for Conventional-noRAR, Conventional-RAR, Reuse-noRAR, and Reuse-RAR, respectively. It is worth noting that

Conventional-RAR has a lower cutoff than Conventional-noRAR (0.794 vs. 0.829) and Reuse-RAR has a lower cutoff than Reuse-noRAR (0.827 vs. 0.832). The reason is that, in RAR designs, when one arm has a high response rate in the early stage of the study due to random variation, more participants will be assigned to that arm and the response rate will regress to the actual rate, and hence it will be less likely to observe simulations with extremely high $prob(j = j_{max})$. Along the same lines of reasoning, Conventional-RAR has a much lower cutoff than the Reuse-RAR (0.794 vs. 0.827) because Reuse-RAR adapts much less aggressively than Conventional-RAR in two aspects: (1) Reuse-RAR runs much faster than the Conventional-RAR and has much less time to adapt; (2) RAR only applies to the first treatment of each participant in the Reuse-RAR while it applies to all the observations in Conventional-RAR.

The success rates (Type I error rates) for each scenario under the null hypothesis are given in Table 2.1. All the Type I error rates are controlled at around 5%, with a range between 4.8% and 5.1%.

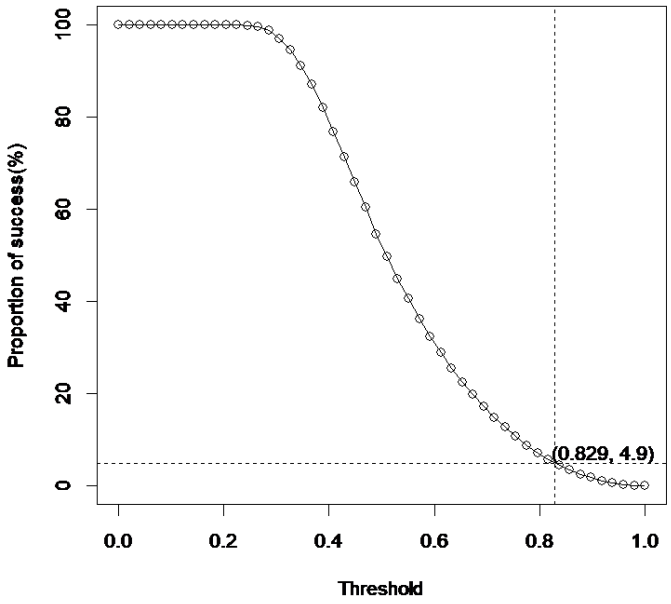

Figure 2.1 Proportion of success (Type I error) by threshold (δ) based on simulations for Conventional-noRAR design in the null scenario when λ=1.5

		Accrual rate			
Design	Threshold	1.5	3	4.5	6
Conventional-noRAR	0.829	4.9%	5.0%	5.0%	5.0%
Conventional-RAR	0.794	4.9%	5.1%	5.0%	4.9%
Reuse-noRAR	0.832	5.0%	5.0%	4.9%	5.1%
Reuse-RAR	0.827	5.1%	4.8%	4.8%	4.8%

Table 2.1 Type I error under H_0

2.2.8. Carryover effect and period effect

For Reuse-noRAR and Reuse-RAR, we assume there is a 20% carryover effect and no period effect. The sampling distribution is

$$y_{it} \sim \begin{cases} Bernoulli\left(logit^{-1}\left(logit\left(\theta_j^T\right) + \epsilon_i^T\right)\right) & when\ t = 1 \\ Bernoulli\left(logit^{-1}\left(logit\left(\theta_j^T + \pi^T * \theta_{j'}^T\right) + \epsilon_i^T\right)\right) & when\ t > 1 \end{cases}$$

where $\theta_{j'}^{T}$ is the virtual response rate of the treatment participant received during period $t-1$; π^{T} = 20%.

2.2.9. Simulations

In total, we investigated 12 scenarios: the combinations of 4 different accrual rates (λ: 1.5, 3, 4.5, and 6), 1 participant variation ($\sigma_{\epsilon^T}^2$: 0.25), and 3 virtual response rates ($\boldsymbol{\theta}^T = (0.2, 0.2, 0.2, 0.2, 0.2)$, $(0.3, 0.3, 0.3, 0.4, 0.5)$, $(0.3, 0.3, 0.3, 0.5, 0.5)$). Each scenarios will be conducted using 4 designs: Conventional-noRAR, Conventional-RAR, Reuse-noRAR, and Reuse-RAR.

For each scenario, we run 10,000 simulations. The maximum 95% margin of error is$1.96\sqrt{0.5 * 0.5/10000} < 0.01$. With a Type I error of 0.05 or power of 0.90, the margin of error is much smaller, $1.96\sqrt{0.05 * 0.95/1000} = 0.004$ and is $1.96\sqrt{0.1 * 0.9/1000} = 0.005$ respectively.

The simulations are implemented in R(13) and Stan(14, 15). R is used to generate participant response data and Stan is used to perform interim and final analyses.

2.3 Results

In this section, we report the simulation results comparing the four designs in terms of the following operating characteristics: power, number of participants enrolled, trial duration, and proportion of observations that received the best treatment. We also explored the performance of the Reuse-RAR design when allowing RAR for subsequent treatments and the impact of participant dropouts on the Reuse-RAR design.

2.3.1 Power

The power for different scenarios under $H_1: \boldsymbol{\theta}^T = (0.3, 0.3, 0.3, 0.4, 0.5)$ and $H_2: \boldsymbol{\theta}^T = (0.3, 0.3, 0.3, 0.5, 0.5)$ are given in Figure 2.2. For both H_1 and H_2, Conventional-noRAR design had the lowest power and Conventional-RAR design had the highest power. The two Reuse designs had a power between that of the two Conventional designs, with Reuse-RAR higher than the Reuse-noRAR. RAR increased power in both Conventional designs and Reuse designs.

Under H_1, when there was a single drug that was better than the other drugs, Conventional-noRAR had a notably lower power than Reuse-noRAR. Given that both designs had exactly 900 observations, the reason to have such a big difference in power was that the independent logistic model used by Conventional-noRAR design did not control for patient specific variation while the hierarchical logistic model used by the Reuse-RAR did. On the other hand, we did not see a big difference between the Conventional-RAR and Reuse-RAR design. Two factors reduced the power advantage of the hierarchical logistic model used by Reuse-RAR design: (1) Reuse-RAR design ran faster than the Conventional-RAR and had less time to adapt; (2) RAR only applied to the first treatment of each participant in the Reuse-RAR while it applied to all the observations in Conventional-RAR. As a result, a smaller proportion of observations were assigned to the better treatments (Figure 2.5) and the efficient gain due to RAR in the Reuse-RAR design was less than in the Conventional-RAR design.

Under H_2, when there are two equally effective treatments, the power is much lower than under H_1 across the designs and scenarios. In section 2.6, we define power as the proportion of simulations that meet the success threshold: $\Pr(j = j_{max}) > \delta$. Treatment 4 and treatment 5 compete against each other under H_2 and the probability of meeting success threshold is much

lower than in H_1. The success criterion we choose is not appropriate when there is no single winner arm. We will discuss some other options in the discussion section.

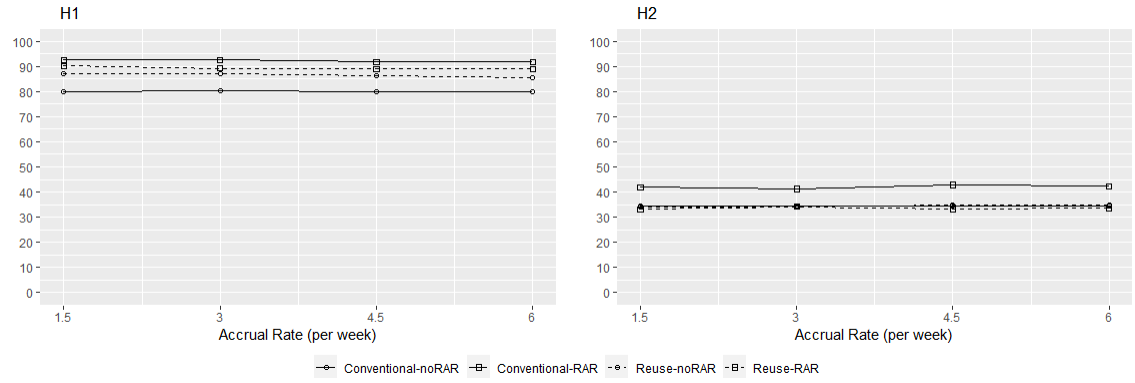

Figure 2.2 Power under H_1 and H_2

2.3.2 Number of participants enrolled

The numbers of participants enrolled in different scenarios are shown in Figure 2.3. Conventional-noRAR and Conventional-RAR design enrolled 900 participants in all scenarios. Reuse-noRAR and Reuse-RAR designs enrolled much fewer participants than the two conventional designs (less than 450) owing to their ability to reuse participants. The reduction in number of participants enrolled in the Reuse designs is the greatest when the accrual rate is low and it decreases gradually as accrual rate increases. Reuse-RAR enrolled slightly more participants than the Reuse-noRAR. This is because Reuse-RAR design assigns more participants to the better drugs as their initial treatment, which in turn deceases the average number of periods per participant. Consequently, with a fixed number of observations (900) for all designs, the number of participants enrolled in the Reuse-RAR is more than the Reuse-noRAR.

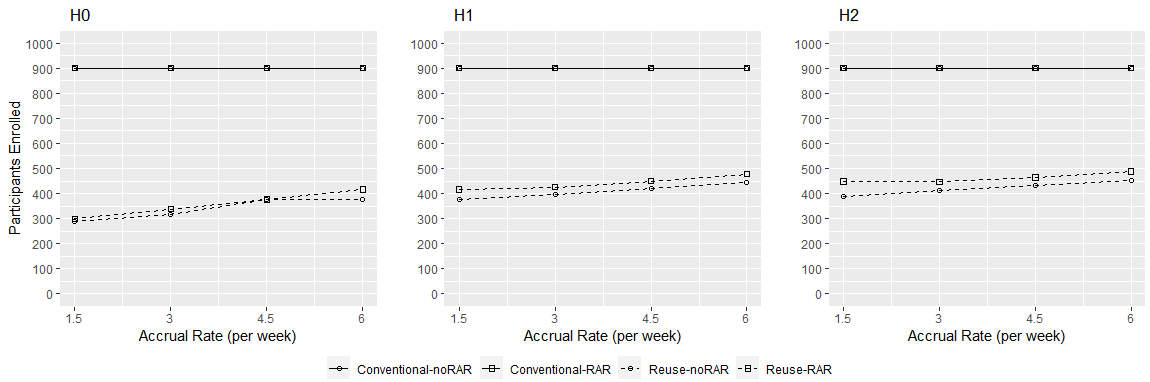

Figure 2.3 Number of participants enrolled

2.3.3 Trial duration:

Trial duration for different scenarios are presented in Figure 2.4. It is directly related to the number of participants enrolled. Conventional-noRAR and conventional-RAR had roughly the same trial duration due to the same number of participants enrolled (900). The two Reuse designs had a much shorter trial duration than the two Conventional designs. The reduction in trial duration was the highest when accrual rate is low. And Reuse-RAR had a slight longer trial duration than the Reuse-noRAR.

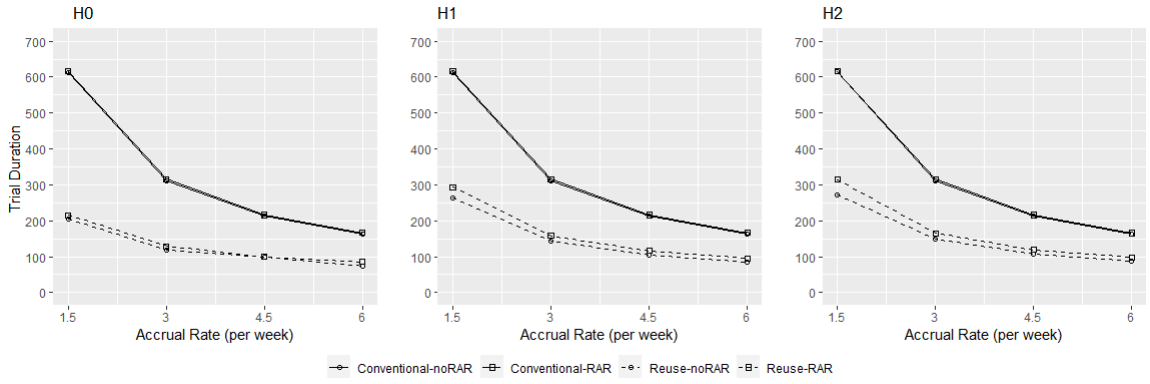

Figure 2.4 Trial duration in weeks

2.3.4 Proportion of observations that received treatment 5

The proportion of observations that received treatment 5 (the best treatment under H_1) in different scenarios are shown in Figure 2.5. Under H_0, 20% of observations received treatment 5 regardless of scenarios and designs. Under H_1 and H_2, the rates were about 20% for both Conventional-noRAR and Reuse-noRAR with reuse-noRAR slightly higher. Conventional-RAR had the highest proportion of observations receiving treatment 5, and it was followed by Reuse-RAR.

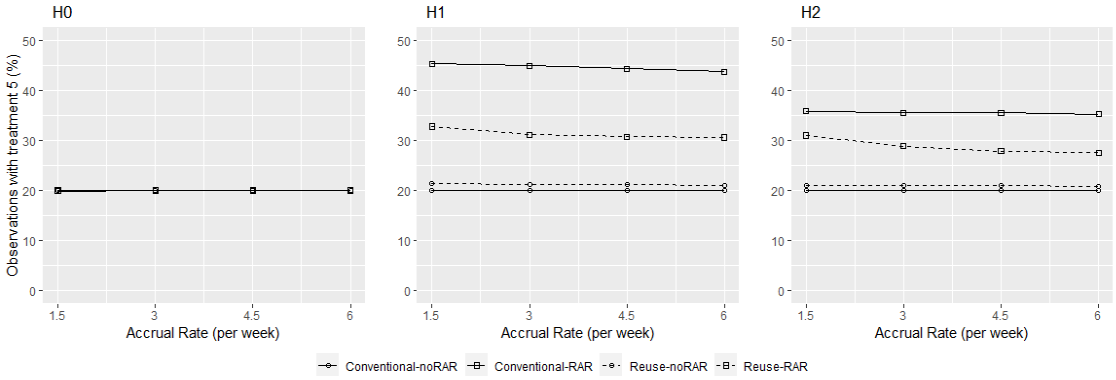

Figure 2.5 Proportion of observations received treatment 5

2.3.5 Compare Reuse-RAR(complete) and Reuse-RAR

In the introduction section, we pointed out that, the RAR randomization only applied to the initial treatment of each participant in order to avoid overly complicating trial conduction. We call this approach Reuse-RAR. We can further improve the efficiency of the Reuse-RAR design by increasing the aggressiveness of adaption to allow RAR randomization for the subsequent treatment assignments. Specifically, instead of using sampling without replacement to determine the order of subsequent treatments, v_j of the remaining treatments were normalized and used to randomly assign treatment for the next period. This approach is called Reuse-

RAR(complete). Simulations were conducted to assess the impact of Reuse-RAR(complete) on operating characteristics. Figure 2.6 compares Reuse-RAR(complete) and Reuse-RAR under H_1 when the accrual rate is 3 and participant variation is 0.25. The Reuse-RAR(complete) assigned slightly more observations to the best arm and increased power slightly. The number of participants enrolled and the trial duration are almost identical.

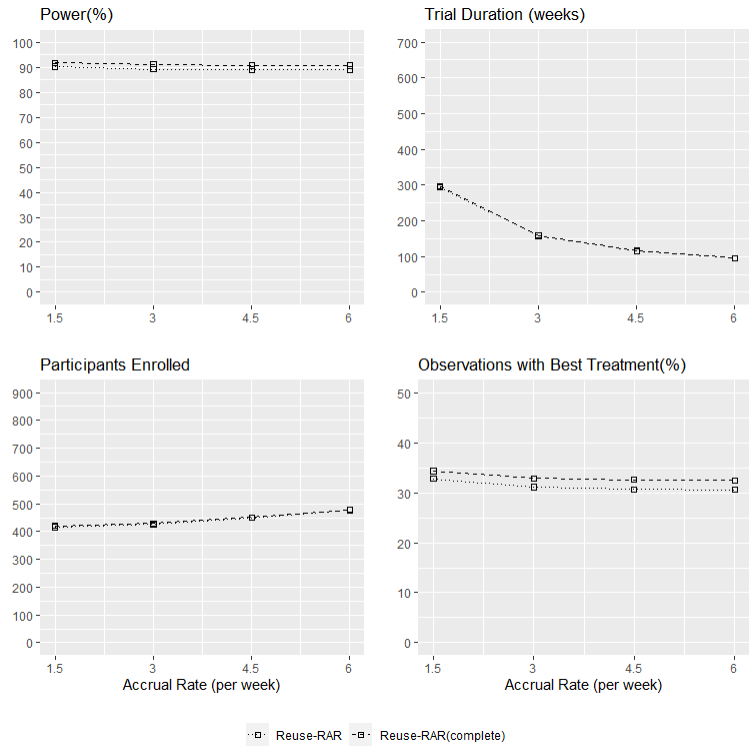

Figure 2.6 Compare Reuse-RAR(complete) and Reuse-RAR(initial)

2.3.6 Participant dropouts

In this section, we explored the impact of participant dropouts on operating characteristics. Figure 2.7 shows simulation result comparing the scenarios with 10% dropouts and the scenarios with no dropouts when accrual rate is 3 and participant variation is 0.25. Overall, in scenarios with a 10% dropout, the number of assessable observations decreased by around 10% and the power decreased by around 4% across all designs. While trial duration and number of participants enrolled were not affected by the dropouts in the Conventional-RAR and Conventional-noRAR design, they were slightly higher in the Reuse-RAR and Reuse-noRAR design when there was a 10% dropouts. Nonetheless, in both scenarios with or without dropouts,

the Reuse-RAR and Reuse-noRAR design were much more efficient in terms of trial duration and number of participants enrolled.

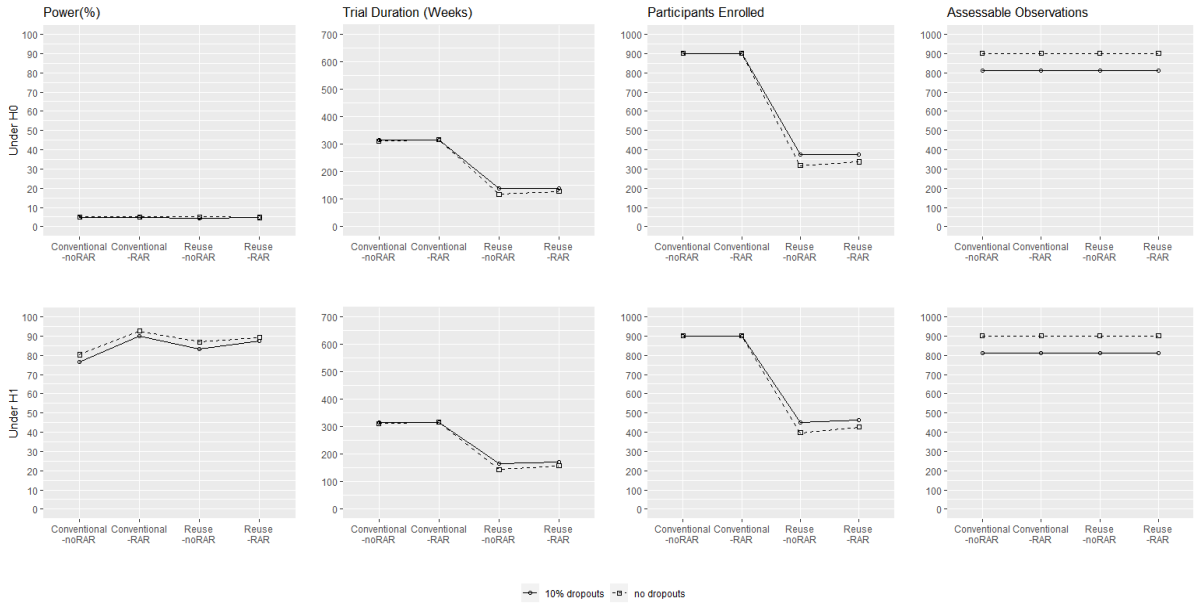

Figure 2.7 Compare scenarios with a 10% dropout with scenarios with no dropouts

2.4 Conclusion

Our simulations showed that, of the four designs, Reuse-RAR is the most efficient design which can achieve a higher power with a shorter trial duration and a smaller number of participants. Conventional-noRAR is the least efficient design. RAR does improve efficiency in both Conventional designs and Reuse designs.

When compared with Reuse-noRAR design, Reuse-RAR has a slightly higher power with a comparable number of participants and trial duration. This efficiency improvement is achieved by assigning more participants to the promising treatments.

When compared with Conventional-RAR design, the Reuse-RAR design can achieve a slightly lower power with a much smaller number of participants and a much shorter trial duration, especially when the accrual rate is low. This efficiency improvement is achieved by reusing participants, and it goes beyond the conventional-RAR design's efficiency gain by assigning more participants to the promising arms. However, when the accrual rate increases, the efficiency gain decreases. This is because the Reuse-RAR runs much faster than the Conventional-RAR design and it has less time for the trial to adapt.

2.5 Discussion

The proposed Reuse-RAR design belongs to the large class of RAR designs. The aggressiveness and timing of adaptation have a significant impact on the RAR performance (1). Reuse-RAR(complete) can further improve the efficiency of the Reuse-RAR design by increasing the aggressiveness of adaption to allow RAR randomization for the subsequent treatment assignments. Our simulations showed that, compared with Reuse-RAR, Reuse-RAR(complete) assigned slightly more observations to the best arm and increased the power slightly. However, the efficiency gain was at the cost of increased trial conduction complexity, which has been one of the major critiques of RAR design. Whether to employ Reuse-RAR(complete) should be decided case by case by balancing the efficiency gain against the increased trial conduction complexity.

The two designs that reuse participants (Reuse-RAR and Reuse-noRAR) require participants to be engaged in the study longer than Conventional-RAR and Conventional-noRAR. In the extreme case where participants do not respond to any treatments, the time required to be engaged in the Reuse-RAR and Reuse-noRAR could be as long as 5 times that of the Conventional-RAR and Conventional-noRAR. As a result, Reuse-RAR and Reuse-noRAR

are more susceptible to dropouts. Simulations in section 3.7 showed that participant dropouts slightly increase the trial duration and number of participants for the designs that reuse participants, but not for the conventional designs. However, simulations also showed that, even in the presence of participant dropouts, Reuse-RAR and Reuse-noRAR performed better than the conventional designs by achieving a similar power with a much shorter trial duration and less participants. Another related concern caused by the long engagement time required in the Reuse designs is that participant characteristics, including disease stage, drug exposure, treatment resistance, etc., can evolve during the course of the treatment. Bias can be introduced because observations at later treatment periods may have more severe conditions or higher drug resistance. To mitigate the bias, we can expand the model by adjusting related participant characteristics. Furthermore, covariate-adjusted adaptive randomization (17), which allows allocation rules to consider both patient response and patient characteristics, can be used to further improve RAR.

In the simulations, we assumed there was a carryover effect that was consistent across different participants and different treatments. This assumption may not be true in general. The model can be modified to better capture the carryover effect according to substantive subject matter knowledge. Another frequently used approach is to include a washout period between two treatments. Including a washout period will increase the trial duration for Reuse-RAR and Reuse-noRAR. The extent to which the washout period will affect the trial duration is determined by the length of the washout period.

In section 3.1, simulations showed that the power was much lower under H_2, when there were two equally effective arms, than under H_1, when there was a single treatment that was better than the rest of the treatments. Power was defined as the proportion of simulations that

meet the success threshold: $\Pr(j = j_{max}) > \delta$. Under H_2, treatment 4 and treatment 5 compete against each other and the likelihood of having an arm to meet the success threshold $\Pr(j = j_{max}) > \delta$ is very low. For scenarios with multiple arms that are equally effective, the success criterion we chose is not appropriate. An option is to assess the probability of being a better arm when compared with other arms. The success criterion can be defined as $\Pr(\theta_j > \theta_{j'}) > \delta'$ for any $j \in (1,2,3,4,5)\ and\ j' \in (1,2,3,4,5)$.

Freidlin et al. (18) pointed out, for studies with only 2 arms, the RAR performs poorly and results in a lower power due to the deviation from the optimal 1:1 randomization. We should use fixed 1:1 randomization in two arms studies, if optimizing power is of primary concern (a caveat would be if one is willing to sacrifice a bit of power for placing participants on the better arm, see Wick et al. (19)). However, the reusing participant scheme is still relevant, and it may result in smaller and shorter clinical trial. The benefit may be small to moderate due to the fact each participant will contribute maximum of 2 observations. More research is needed to evaluate the performance of the reusing participants scheme in two arm studies. The Reuse participants scheme is best suit for studies with multiple arms and with a slow accrual rate.

References

1. Tang c, Sherman SI, Price M, Weng J, Davis SE, Hong DS, Yao JC, Buzdar A, Wilding J, Lee JJ. Clinical Trial Characteristics and Barriers to Participant Accrual: The MD Anderson Cancer Center Experience over 30 years, a Historical Foundation for Trial Improvement. Cancer Therapy: Clinical March 2017
2. Stensland KD, McBride RB, Latif A, Wisnivesky J, Hendricks R, Roper N, et al.Adult cancer clinical trials that fail to complete: an epidemic. J Natl Cancer Inst 2014
3. Nason M, Follman D.Design and Analysis of Crossover Trials for Absorbing Binary Endpoints.Biometrics 2010;7:958-965
4. Wellek S, Blettner M. On the proper use of the crossover design in clinical trials: part 18 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2012;109(15):276–281.
5. Berry SM, Carlin BP, Lee JJ, Muller, P, Bayesian Adaptive Methods for Clinical Trials. New York, NY: CRC Press; 2011.
6. Gajewski BJ, Berry SM, Quintana M, Pasnoor M, Dimachkie M, Herbelin L, Barohn R. Building efficient comparative effectiveness trials through adaptive designs, utility functions, and accrual rate optimization: finding the sweet spot. Stat Med. 2015; 34(7):1134-49.
7. Brown AR, Gajewski BJ, Aaronson LS, Mudaranthakam DP, Hunt SL, Berry SM, Quintana M, Pasnoor M, Dimachkie MM, Jawdat O, Herbelin L, Barohn RJ. A Bayesian comparative effectiveness trial in action: developing a platform for multisite study adaptive randomization. Trials. 2016; 17(1): 428

8. Pasnoor M, Dimachkie MM, Barohn RJ. Cryptogenic sensory polyneuropathy. Neurol Clin. 2013;31:463–76.
9. Burckhardt CS, Jones KD. Adult measures of pain: The McGill Pain Questionnaire (MPQ), Rheumatoid Arthritis Pain Scale (RAPS), Short-Form McGill Pain Questionnaire (SF-MPQ), Verbal Descriptive Scale (VDS), Visual Analog Scale (VAS), and West Haven-Yale Multidisciplinary Pain Inventory (WHYMPI). Arthritis Care Res. 2003;49(S5):S96–S104.
10. Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D. Bayesian Data Analysis. New York, NY: CRC Press; 2014.
11. Betancourt M. A conceptual introduction to Hamiltonian Monte Carlo. Preprint arXiv:1701.02434. Columbia University, New York.
12. U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry. Nov. 2019.
13. R Core Team(2017). R: a Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing
14. Stan Development Team (2017) RStan: the R interface to Stan, version 2.16.1. (Available from http://mc-stan.org.)
15. Stan Modeling Language User's Guide and Reference Manual, Version 2.16.0. Stan Development Team. (Available from http://mc-stan.org.)

16. Jang Y, Zhao W, Durkalsk-Mauldin V. Impact of adaption algorithm, timing, and stopping boundaries on the performance of Bayesian response adaptive randomization in confirmative trials with binary end-point. Contemp Clin Trials 2017; 62: 114-120
17. Rosenberger WF, Vidyashankar AN, Agarwal DK. Covariate-adjusted response-adaptive designs for binary response. Journal of Biopharmaceutical Statistics 2001; 11(4), 227-236
18. Freidlin B, Korn EL. Adpative randomization versus interim monitoring. J clin Oncol 2013; 31(7): 969-970.
19. Wick J, Berry SM, Yeh H, Choi W, Pacheco CM, Daley C, Gajewski BJ, A Novel Evaluation of Optimality for Randomized Controlled Trials. Journal of Biopharmaceutical Statistics 2017; 27 (4), 659-672.

Chapter 3 : Innovative Bayesian EMAX model with a mixture of normal distributions for dose-response in clinical trials

3.0 Abstract

When a dose-response relationship is monotonic, the EMAX model has been shown to provide a good empirical fit for designing and analyzing dose-response data across a wide range of pharmaceutical studies. However, the EMAX model has never been applied to a finite mixture distribution. Motivated by a proposal investigating DHA dose effect on preterm birth (PTB, <37 weeks gestation) rate, we developed an innovative Bayesian EMAX mixture model incorporating the three normal components finite mixture model into the EMAX framework. The proposed Bayesian EMAX mixture model analyzes gestational age as a continuous variable, which allows for statistically efficient estimates of PTB rate using various cut point with the same parsimonious model. For example, we can estimate the rate of early PTB (ePTB, <34 weeks gestation), PTB (<37 weeks gestation), and late-term birth (>41 weeks gestation) using the same model. We compared our proposed EMAX mixture model with an EMAX logistic model and an independent doses logistic model for a dichotomized endpoint using extensive simulations. Across the scenarios under consideration, the EMAX mixture model achieved higher power than the EMAX logistic model and the independent doses logistic model in detecting the effect of DHA supplementation on the PTB rate. The EMAX mixture model also resulted in smaller mean squared errors (MSE) in PTB rate estimates.

3.1 Introduction

Preterm birth (PTB) is defined as birth before 37 weeks gestation. One in ten U.S. pregnancies ends in PTB, yielding nearly half a million preterm infants born each year. PTB is the primary cause of infant mortality, costs the U.S. health system billions of dollars annually, and, for many of the infants who survive, results in continued individual, family, and societal challenges due to associated morbidity and disabilities. Despite the significant investment of the

National Institutes of Health (NIH), the Centers for Disease Control and Prevention (CDC), and foundations such as the March of Dimes toward understanding and preventing PTB, researchers have only recently identified prevention strategies for spontaneous PTB. In a November 2018 Cochrane Review (1), researchers concluded there was strong evidence that consumption of the omega-3 fatty acid docosahexaenoic acid (DHA) could reduce PTB by 11%, and early PTB (ePTB, <34 weeks gestation) by 42%. These results are compelling. However, additional research is necessary to move from an observed effect of DHA to a scalable preventive intervention for PTB. The critical issue is that the DHA dose needed to reduce PTB is unknown. At present, the National Academy of Medicine does not set a Dietary Reference Intake (DRI) for DHA in pregnancy because the amount of DHA required to reduce PTB has not been established. Most prenatal supplements available in the U.S. contain ~0.2g DHA, a much lower dose than provided in most randomized controlled trials (RCTs) included in the Cochrane Review (≥0.6g DHA). A dose-response study is necessary to develop evidence-based policy and advise women about the DHA dose needed to reduce PTB. Our goal is to identify an efficient trial design to evaluate the effect of DHA dose on PTB.

One frequently used dose-response model is the pairwise independent doses model. In this model, no functional relationship is assumed between the dose and effect, and all doses are modeled independently and compared with each other. The independent doses model is often inefficient and results in lower power because of its lack of functional relationship assumption. When the dose-response relationship can be assumed monotonic, an EMAX (MAXimum Effect) model has been shown to provide a good empirical fit for designing and analyzing dose-response data across a wide range of pharmaceutical studies (2). The EMAX model assumes the dose-response relationship follows a nonlinear monotonic function with a parameter representing the

maximum effect that can be achieved when dose approaches to infinity and another parameter representing the dose that achieves 50% of the maximum effect. One option to evaluate the DHA dose effect on PTB is to apply the EMAX model treating PTB as a dichotomous endpoint. However, studies have shown that dichotomizing continuous endpoints results in a loss of information and reduced power (3,4,5). We propose a Bayesian EMAX model that analyzes gestational age as continuous. Schwartz et al. showed that the distribution of gestational age can be described by a mixture of three normal distributions (6). Thus, we developed our EMAX mixture model for a continuous endpoint with a mixture distribution. We compared our model with two models that dichotomize gestational age: the EMAX model (EMAX logistic model) and the independent doses logistic model. Extensive simulations showed that the EMAX Mixture model achieved a much higher power for detecting the DHA dose effect on PTB than the other two models and resulted in much smaller mean squared errors (MSEs) in estimates of PTB rates. Additionally, the EMAX Mixture model is attractive because it allows for statistically efficient estimates of PTB rates using different gestational age cut-points within the same parsimonious model. For example, we can estimate the rate of early preterm birth (<34 weeks gestation), preterm birth (<37 weeks gestation), and late-term birth (>41 weeks gestation) using the same model.

The remainder of the article is organized as follows. In Sections 2.1 and 2.2, we describe the study motivation and cover the three statistical models in detail (EMAX Mixture, EMAX logistic, and independent doses logistic). Section 2.3 describes the simulation scenarios used to assess model operating characteristics. Section 2.4 is model calibration and type I error, and Section 2.5 provides simulation details. The simulation results are summarized in Section 3. In Section 4, we apply the three models to a simulated dataset to illustrate the models' application.

In Section 5, we conclude from our analysis and discuss the advantages and limitations of the EMAX Mixture model and future work.

3.2 Methods

3.2.1 Study summary

Our research was motivated by a proposal whose primary aim was to evaluate the dose-response relationship for DHA supplementation on PTB by leveraging the data from six NICHD supported randomized clinical trials (RCTs) of DHA supplementation in pregnancy conducted between 2006 and 2020 (R21 HD058269, R21 HD059019; R01 HD084586; R01 HD086001, R01 HD047315, R01 HD083292). The trials combined enrolled over 2000 U.S. women with a singleton pregnancy in four metropolitan areas (Kansas City, Chicago, Cincinnati, and Columbus). Six DHA doses were used across the trials: 0g (n=350), 0.2g (n=700), 0.45g (n=175), 0.6g (n=180), 0.8g (n=150) and 1g (n=550).

As a secondary aim, DHA supplementation was hypothesized to have a bigger effect on ePTB and/or PTB in participants with a lower phospholipid DHA level at enrollment. By dividing the participants into two groups according to their phospholipid DHA (as a percent of total fatty acids) at enrollment (Low: phospholipid DHA <6%; High: phospholipid DHA $\geq$ 6%), the proposal wanted to determine if phospholipid DHA status at enrollment influences the effect of DHA supplement on ePTB and/or PTB.

Our goal was to identify an efficient trial design to evaluate the primary and secondary aims of the proposal.

3.2.2 Statistical models

Let T_{di} denote the gestational age for participant i in arm d, where d represents the DHA supplement dose and can take values of $0g, 0.2g, 0.45g, 0.6g, 0.8g,$ and $1g$. The number of

participants in each arm, denoted by n_d, is 350, 700, 175, 180, 150, and 550 for the 6 doses, respectively. Let y_d be the number of ePTBs in dose d, which can be determined by $y_d = \sum_{i=1}^{n_d} I(T_{di} < 34)$, where $I(x < a) = \begin{cases} 1 & if\ x < a \\ 0 & if\ x \geq a \end{cases}$. This section will describe statistical models for the independent doses logistic model, EMAX logistic model, and the EMAX Mixture model.

1.2.2.1Independent doses logistic model

Let p_d be the probability of an ePTB in dose d. The number of ePTBs in dose d follows a binomial distribution, $y_d \sim Bin(n_d, p_d)$, and it is modeled independently for each dose. A normal distribution $N(0,5^2)$ is used as a vague prior for $\theta_d = \log\left(\frac{p_d}{1-p_d}\right)$. When transformed back to probability scale using an anti-logit function, the prior yields a 95% equal-tailed interval of (0.001, 0.999). Hamiltonian Monte Carlo (Betancourt; Gelman et al., 2014) is used to obtain the posterior distribution of θ_d. The posterior probability p_d can be calculated using $p_d = \frac{\exp(\theta_d)}{1+\exp(\theta_d)}$. The posterior probability of dose d being better than the control arm, $Pr(p_d < p_0|data)$, can be estimated as the proportion of Monte Carlo draws satisfying $p_d < p_0$. The trial success is achieved when$\max\left(Pr(p_d < p_0|data)\right) > \delta_{ind.1}$. The threshold $\delta_{ind.1}$ is chosen by simulations to ensure a 5% type I error rate.

To determine whether phospholipid DHA at enrollment influences the effect of DHA supplementation, we model the high phospholipid DHA cohort and low phospholipid DHA cohort using the same model described above, but separately. We denote the odds ratio between the arm with the highest dose (d =1g) and the control arm ($d = 0g$) using $O = \frac{p_1}{p_0}$. The posterior probability of having a bigger DHA effect in the low phospholipid DHA cohort than in the high phospholipid DHA cohort is $\Pr(O_L < O_H|data)$. It can be calculated as the proportion of Monte

Carlo draws satisfying $O_L < O_H$. Trial success is achieved when $\Pr(O_L < O_H|data) > \delta_{ind.2}$. The threshold $\delta_{ind.2}$ is chosen by simulations to ensure a 5% type I error rate.

1.2.2.2 EMAX logistic model

As in the independent doses logistic model, $y_d \sim Bin(n_d, p_d)$. Instead of modeling θ_d independently for each dose d, the EMAX function is used to model the relationship between θ_d and d: $\theta_d = a_1 + \frac{a_2 * d}{d + a_3}$.

- a_1 is a constant offset. When $d = 0$, a_1 determines θ_0 solely, which in turn determines the ePTB rate in the control arm.
- a_2 is a scalar coefficient reflecting the dose effect. It is the theoretical maximum effect above the constant offset that can be achieved. As dose tends to infinity the theoretical maximum efficacy on the logit scale is $a_1 + a_2$, thus the model is called the EMAX model.
- a_3 is a positive scalar representing the effective dose strength that achieves 50% of the theoretical maximal effect above the constant offset. For an effective dose of d=a_3 the efficacy on logit scale is $a_1 + \frac{a_2}{2}$.

A non-informative prior distribution $N(0,4)$ is used for $a_1, a_2, and\ a_4$. For a_3, a half-normal prior $N(0,1)$ is used so that a_3 can take positive values only. Hamiltonian Monte Carlo (Betancourt; Gelman et al., 2014) is used to obtain the posterior distribution of a_1, a_2, and a_3 and the posterior probability p_d can be calculated using $p_d = \frac{\exp(\theta_d)}{1+\exp(\theta_d)}$, where $\theta_d = a_1 + \frac{a_2 * d}{d + a_3}$. It is easy to prove that when a_2 is negative, θ_d decreases as d increases. Therefore, the success of a trial is defined as having a posterior probability of $a_2 < 0$ greater than a cutoff value, $\Pr(a_2 < 0) > \delta_{EMAX.1}$. The threshold $\delta_{EMAX.1}$ is chosen by simulations to ensure a 5% type I error rate.

To determine whether phospholipid DHA at enrollment influences the effect of DHA supplementation on PTB, we model the high and low phospholipid DHA cohorts using the same model described above separately. Let a_{2H} and a_{2L} denote EMAX parameters a_2 in the high and low phospholipid DHA cohorts, respectively. The success of the trial is defined as having a posterior probability of $a_{2L} < a_{2H}$ greater than a cutoff, $\Pr(a_{2L} < a_{2H}|data) > \delta_{EMAX.2}$, where $\Pr(a_{2L} < a_{2H}|data)$ can be estimated using the proportion of Monte Carlo draws satisfying $a_{2L} < a_{2H}$, and the threshold $\delta_{EMAX.2}$ is chosen by simulations to ensure a 5% type I error rate.

1.2.2.3 EMAX Mixture model

In the finite mixture model developed by Schwartz et al.(6), gestational age T_{di} follows a finite mixture model with three normal components that describe the mixture of high-, medium-, and low-risk groups: $N_1 = N(33.29,13.23)$, $N_2 = N(38.26, 2.48)$, and $N_3 = N(39.59, 0.960)$. The probability distribution function of T_{di} is $f(T_{di}|\Delta_{1\mathrm{d}}, \Delta_{2\mathrm{d}}, \Delta_{3\mathrm{d}}) =$

$\Delta_{1\mathrm{d}}\phi(T_{di}|33.29, 13.23) + \Delta_{2\mathrm{d}}\phi(T_{di}|38.26,2.48) + \Delta_{3\mathrm{d}}\phi(T_{di}|39.59,0.96)$, where $\phi(T|\mu, \sigma^2)$ is the normal probability density function with mean μ and variance σ^2, and $\Delta_{1\mathrm{d}}, \Delta_{2\mathrm{d}}$, and $\Delta_{3\mathrm{d}}$ are the mixture weights for arm d, which can take values between 0 and 1 and with $\Delta_{1\mathrm{d}} + \Delta_{2\mathrm{d}} + \Delta_{3\mathrm{d}} = 1$. The three components represent the high, medium, and low-risk groups for ePTB and can model different populations by adjusting the mixture weights. The model was derived from the North Carolina Detailed Birth Record (NCDBR) registry, with 336,129 records included in the final analysis. It is representative and has generalizability. It has been used successfully in other studies of PTB (12,13,14).

Based on Schwartz's finite mixture model, we propose a dose-response model that applies the EMAX function to finite mixture distributions. We call it the EMAX Mixture model. Let $\theta_{1\mathrm{d}} = \log(\frac{\Delta_{1d}}{\Delta_{3d}})$ represent the odds ratio of the mixture weights for the high- versus the low-

risk groups' normal components. Similarly, $\theta_{2d} = \log\left(\frac{\Delta_{2d}}{\Delta_{3d}}\right)$ is the odds ratio for the medium- versus low-risk groups' normal components. The EMAX function is employed to model the relationship between the odds ratio comparing high- and low-risk groups, θ_{1d}, and dose d. This relationship is given by $\theta_{1d} = a_1 + \frac{a_2 d}{d+a_3}$. Without losing the model generalizability, we assume the odds ratio comparing medium- to low-risk groups, θ_{2d}, stays constant for all doses: $\theta_{2d} = a_4$.

- a_1 and a_4 are the constant offsets. They determine the three mixture weights when the effective dose strength is 0: $\Delta_{10} = \frac{\exp(a_1)}{1+\exp(a_1)+\exp(a_4)}$, $\Delta_{20} = \frac{\exp(a_4)}{1+\exp(a_1)+\exp(a_4)}$, and $\Delta_{30} = \frac{1}{1+\exp(a_1)+\exp(a_4)}$.
- a_2 is the scalar coefficient reflecting the dose effect. When it is negative, as the dose increases the mixture weight of the 1st component (high risk) decreases and the mixing weights of the 2nd (median risk) and 3rd (low risk) components increase. a_2 determines the theoretical maximum effect (the minimum weight of the 1st component) above the constant offset that can be achieved. When the effective dose strength is not 0: $\Delta_{1d} = \frac{\exp\left(a_1+\frac{a_2 d}{d_i+a_3}\right)}{1+\exp\left(a_1+\frac{a_2 d}{d+a_3}\right)+\exp(a_4)}$, $\Delta_{2d} = \frac{\exp(a_4)}{1+\exp\left(a_1+\frac{a_2 d}{d+a_3}\right)+\exp(a_4)}$, and $\Delta_{3d} = \frac{1}{1+\exp\left(a_1+\frac{a_2 d}{d_i+d}\right)+\exp(a_4)}$.
- a_3 is a positive scalar representing the effective dose strength that achieves 50% of the theoretical maximal effect.

We use a vague prior $N(0,4)$ for a_1, a_2, and a_4, and a half-normal $N(0,1)$ for a_3 to restrict it to be positive. Hamiltonian Monte Carlo (Betancourt; Gelman et al., 2014) is used to obtain the posterior distribution of $a_1, a_2, a_3,$ and a_4. The posterior distribution for mixture weights $\Delta_{1d}, \Delta_{2d}, \Delta_{3d}$ can be calculated using the formulas given above. The posterior probability of

having ePTB (<34 weeks gestation age) for dose d can be calculated using $p_d = \int_0^{34} f(t|\Delta_{1d}, \Delta_{2d}, \Delta_{3d})\, dt$. By changing the upper integration bound, we can calculate the posterior probability of PTB rates at different cutoffs. For example, the posterior probability of having PTB (<37 weeks gestation age) is $\int_0^{37} f(t|\Delta_{1d}, \Delta_{2d}, \Delta_{3d})\, dt$. Allowing for statistically efficient PTB rate estimates using various cut points with the same parsimonious model makes the EMAX Mixture model attractive.

The success of a trial is defined as having a posterior probability of $a_2 < 0$ greater than a threshold, $\Pr(a_2 < 0) > \delta_{EMAX\_Mix.1}$. The posterior probability of $a_2 < 0$ can be estimated using the proportion of Monte Carlo draws with $a_2 < 0$. $\delta_{EMAX_Mix.1}$is chosen using simulations to ensure a 5% type I error rate.

The high and low baseline phospholipid DHA cohorts are modeled separately using the same model described above to investigate whether phospholipid DHA at enrollment iinfluences the effect of DHA supplementation. Let a_{2H} and a_{2L} denote the EMAX parameters a_2 in the high and low phospholipid DHA cohorts, respectively. The success of the trial is defined as $\Pr(a_{2L} < a_{2H}|data) > \delta_{EMAX\_Mix.2}$. The posterior probability of $a_{2L} < a_{2H}$ can be estimated using the proportion of Monte Carlo draws with $a_{2L} < a_{2H}$. $\delta_{EMAX_Mix.2}$is chosen using simulations to ensure a 5% type I error rate.

3.2.3 Simulation scenarios

Two sets of simulations were performed to compare the operating characteristics of the three models in consideration. The first set of simulations evaluates the dose-response relationship for the effect of DHA supplement. The second set evaluates whether phospholipid DHA at enrollment impacts the effect of DHA supplement.

Four virtual scenarios (Table 3.1) with realistic ePTB rates derived from an existing clinical trial were used to evaluate the dose-response relationship between DHA and ePTB. The "expected" scenario represents the most likely response we believe based on the result from Kansas University DHA Outcome Study (KUDOS) (7). The "optimistic" and "pessimistic" scenarios reflect the 97.5^{th} and 2.5^{th} percentiles of the expected response. Lastly, the improbable scenario that serves as our null hypothesis is labeled "no effect" in Table 3.1. In this scenario, the assumed rates of ePTB are the same across different doses. Therefore, the extent to which this scenario is "successful" actually reflects the Type I error rate.

	Dose (g/day)					
Scenario	**0 (n=350)**	**0.2 (n=700)**	**0.45 (n=175)**	**0.6 (n=180)**	**0.8 (n=150)**	**1 (n=550)**
1 (optimistic)	6.27%	4.90%	3.91%	3.52%	3.13%	2.85%
2 (expected)	3.34%	2.60%	2.00%	1.74%	1.47%	1.27%
3 (pessimistic)	1.56%	1.17%	0.87%	0.75%	0.63%	0.54%
4 (no effect)	3.34%	3.34%	3.34%	3.34%	3.34%	3.34%

Table 3.1 Virtual scenarios (rate of ePTB) for evaluating dose-response relationship for the effect of DHA.

Simulation scenarios investigating whether phospholipid DHA at enrollment impact DHA supplement's effect are given in Table 3.2. In the "optimistic" scenario, the high phospholipid DHA group has a very low but constant ePTB rate of 1.56% across different doses. The low phospholipid DHA group has decreasing ePTB rates that range from 11.01% when $d = 0g$ to 4.16% when $d = 1g$. The average ePTB rates of the high and low groups are equal to the "optimistic" scenario in Table 3.1 (6.27%, 4.9%, 3.91%, 3.52%, 3.31%, and 2.85% for dose of 0g, 0.2g, 0.45g, 0.6g, 0.8g, and 1g, respectively). The "no effect" scenario represents the null

hypothesis where both high and low groups have ePTB rates equal to the "optimistic" scenario in Table 3.1.

		Dose (g/day)					
Scenario	**DHA**	**0 (n=350)**	**0.2 (n=700)**	**0.45 (n=175)**	**0.6 (n=180)**	**0.8 (n=150)**	**1 (n=550)**
1 (optimistic)	High	1.56%	1.56%	1.56%	1.56%	1.56%	1.56%
	Low	11.01%	8.19%	6.20%	5.43%	4.69%	4.16%
2 (no effect)	High	6.27%	4.90%	3.91%	3.52%	3.13%	2.85%
	Low	6.27%	4.90%	3.91%	3.52%	3.13%	2.85%

Table 3.2 Scenarios (rate of ePTB) for investigating whether DHA status at enrollment impacts the effect of DHA supplementation.

3.2.4 Model calibration

According to the U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research (CDER), and Center for Biologics Evaluation and Research (CBER), the type I error rate can be estimated by the proportion of simulations that meet the success threshold in null scenarios (8), and power can be estimated by the proportion of simulations that meet the success threshold in alternative scenarios. To make designs comparable, success thresholds are chosen to achieve similar type I error rates across designs using simulations. This process is called model calibration. For example, Figure 3.1 is the plot of the proportion of successes (type 1 error rate) by threshold values ($\delta_{EMAX\_Mix.1}$) based on simulations using the EMAX Mixture model under the null scenario. As the threshold increases, the proportion of simulations meeting the success criterion decreases. When the threshold is 0.845, the type I error rate is roughly 5%. The more simulations we run for each scenario, the more precise the type I error rate can be. Using the same method, we identified $\delta_{EMAX\_Mix.1} = 0.845$ and $\delta_{EMAX_{Mix},2} = 0.74$ for the EMAX Mixture model, $\delta_{EMAX.1} = 0.955$ and

$\delta_{EMAX,2} = 0.92$ for the EMAX logistic model, and $\delta_{ind.1} = 0.992$ and $\delta_{Ind,2} = 0.97$ for the independent doses logistic model. With these chosen thresholds, the null scenarios' success rates were controlled under 5% in all models.

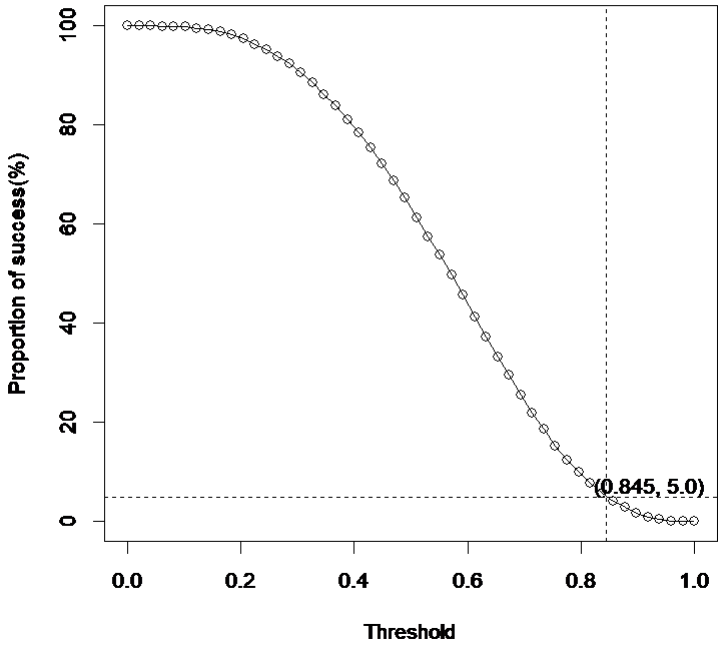

Figure 3.1 Type I error rate (Proportion of success) by threshold (δ_(EMAX_Mix.1)) based on simulations for the EMAX Mixture model in the null scenario

3.2.5 Simulations

For the independent doses logistic model and the EMAX logistic model, we simulated the number of participants with ePTB (y_d) using binomial distributions with $n_d \in \{350, 700, 175, 180, 150\}$ and p_d given in Table 3.1 and Table 3.2.

For the EMAX mixture model, we first used a trial and error method to find values of (a_1, a_2, a_3, a_4) that would result in the early preterm birth rates specified in Table 3.1 and Table

3.2. These values are given in Table 3.3 and Table 3.4, respectively. We then calculated (Δ_{1d}, Δ_{2d}, Δ_{3d}) for each dose using the formulas: $\Delta_{1\mathrm{d}} = \frac{\exp\left(a_1+\frac{a_2 d}{d+a_3}\right)}{1+\exp\left(a_1+\frac{a_2 d}{d+a_3}\right)+\exp(a_4)}$, $\Delta_{2\mathrm{d}} = \frac{\exp(a_4)}{1+\exp\left(a_1+\frac{a_2 d}{d+a_3}\right)+\exp(a_4)}$, and $\Delta_{3\mathrm{d}} = \frac{1}{1+\exp\left(a_1+\frac{a_2 d}{d+a_3}\right)+\exp(a_4)}$. And finally, we used the normal mixture distributions $f(T_{di}|\Delta_{1\mathrm{d}}, \Delta_{2\mathrm{d}}, \Delta_{3\mathrm{d}})$ to generate gestational ages T_{di}.

Scenario	a_1	a_2	a_3	a_4
1 (optimistic)	-2.00	-1.85	1.15	-2.16
2 (expected)	-2.64	-3.6	2.5	-1.72
3 (pessimistic)	-3.53	-3.3	2	-2.52
4 (no effect)	-2.64	0	NA	-1.72

Table 3.3 Parameters used to simulate gestation ages for scenarios in Table 1.

Scenario	**DHA**	a_1	a_2	a_3	a_4
1 (optimistic)	High	-3.53	0	NA	-2.52
	Low	-1.30	-2.5	1.15	-1.85
2 (no effect)	High	-2.00	-1.85	1.15	-2.16
	Low	-2.00	-1.85	1.15	-2.16

Table 3.4 Parameters used to simulate gestation ages for scenarios in Table 2.3.

For each model and each scenario in Table 3.1 and Table 3.2, we ran 10,000 simulations. The maximum 95% margin of error for a binomial is $1.96\sqrt{0.5 * 0.5/10{,}000} < 0.01$, so for a type I error rate of 0.05 and power of 0.90, the margin of error is $1.96\sqrt{0.05 * 0.95/1000} = 0.004$ and is $1.96\sqrt{0.1 * 0.9/1000} = 0.005$, respectively.

The simulations were implemented in R (9 and 10) and Stan (11). R was used to generate gestation age data, and Stan was used to perform analyses.

3.3 Simulation Results

In this section, we report the simulation results comparing the three models under different scenarios described in Section 2.3. We assessed two critical aspects of model performance: statistical power in detecting the effect of DHA supplementation on the PTB rate and mean squared errors (MSE) and bias in PTB rate estimates.

3.3.1 Power

Table 3.5 shows the simulation results for power (proportions of success simulations) for the optimistic, expected, and pessimistic scenarios in Table 3.1, where the goal was to evaluate the dose-response relationship. The EMAX Mixture model had the highest power and independent doses logistic model had the lowest power across all scenarios. In the order of EMAX Mixture, EMAX logistic, and independent doses logistic, power was 99.98%, 84.89%, and 59.86% in the optimistic scenario; 99.79%, 73.35%, and 48.43% in the expected scenario; and 96.76%, 48.92%, and 24.75% in the pessimistic scenario.

Table 3.6 shows the simulation results for power for the optimistic scenario in Table 3.2, where the aim was to investigate whether DHA level at enrollment impacted the effect of DHA supplement. The EMAX Mixture model had the highest power of 95.4%, and the EMAX logistic model of 35.1% followed it. The independent doses logistic model had the lowest power of 27.6%.

Compared with the independent doses logistic model, the EMAX Mixture and EMAX logistic models are more efficient because they take advantage of the monotonic dose-response relationship by using the EMAX function. Compared with the EMAX logistic model, the EMAX Mixture model is more efficient because it treats gestational age as a continuous variable, while

the EMAX logistic model uses a dichotomized gestational age variable. Studies have shown that dichotomizing continuous endpoints results in a loss of information and reduced power (3,4,5).

Scenario	EMAX Mixture	EMAX logistic	Independent logistic
1 (optimistic)	99.98%	84.89%	59.86%
2 (expected)	99.79%	73.35%	48.43%
3 (pessimistic)	96.76%	48.92%	24.75%

Table 3.5 Power for the effective scenarios in Table 1 where the goal was to evaluate the dose-response relationship for effect of DHA supplement on ePTB

Scenario	EMAX Mixture	EMAX logistic	Independent logistic
1 (optimistic)	95.4%	35.1%	27.6%

Table 3.6 Power for the effective scenarios in Table 2 where the aim was to investigate whether DHA level at enrollment had an impact on the effect of DHA supplement on PTBs

3.3.2 MSE and bias

As described in Section 2.2, the posterior distribution of the probability of ePTB, p_d, can be obtained using Monte Carlo simulations. Let $\hat{p}_d|data$ denote the posterior mean of p_d, the expected posterior probability of ePTB can be obtained as the average of $\hat{p}_d|data$ across simulations, $E(\hat{p}_d|data) = \frac{\sum_{k=1}^{S}\hat{p}_{d_k}|data}{S}$, where S is the number of simulations. The sample variance, $\hat{V}_d$, can be calculated as $\hat{V}_d = \frac{\sum_{k=1}^{S}\left(\hat{p}_{d_k} - E(\hat{p}_d)\right)^2}{S-1}$. The bias is the difference between the expected posterior probability $E(\hat{p}_d)$ and the true probability p_d^T, $bias = E(\hat{p}_d) - p_d^T$. The mean squared error is $MSE = bias^2 + \hat{V}_d$.

Table 3.7 shows the simulation results for MSE$\times 10^5$ of $E(\hat{p}_d)$. Across all scenarios and doses, the EMAX Mixture model had the lowest MSE, and the independent doses logistic model

had the highest MSE. When averaged across different doses, MSE$\times 10^5$ for the three models (EMAX Mixture, EMAX logistic, and independent doses logistic) were 1.1, 4.0, and 16.3, respectively, in the Optimistic scenario; 0.8, 2.1, and 8.2, respectively, in the Expected scenario; 0.3, 0.9, 3.6, respectively, in the Pessimistic scenario; 0.8, 2.8, and 14.1, respectively, in the no effect scenario.

Table 3.8 shows the simulation result for bias$\times 10^3$ of $E(\hat{p}_d)$. In most cases, independent doses logistic model had the lowest bias. EMAX Mixture and the EMAX logistic had a comparable amount of bias. Nevertheless, the differences were very small in comparison with sample variance.

Scenario	**Model**	**Dose(g/day)**						**Average**
		0	**0.2**	**0.45**	**0.6**	**0.8**	**1**	
Optimistic	EMAX Mixture	2.3	0.9	0.6	0.7	1.0	1.3	1.1
	EMAX logistic	11.0	3.0	2.0	2.2	2.6	3.0	4.0
	Independent logistic	18.2	6.8	25.7	20.0	21.8	5.1	16.3
Expected	EMAX Mixture	2.0	0.7	0.4	0.4	0.5	0.8	0.8
	EMAX logistic	6.0	1.6	1.0	1.0	1.3	1.7	2.1
	Independent logistic	10.1	3.7	13.2	9.6	10.2	2.3	8.2
Pessimistic	EMAX Mixture	0.9	0.3	0.2	0.1	0.2	0.3	0.3
	EMAX logistic	2.7	0.6	0.4	0.4	0.5	0.6	0.9
	Independent logistic	4.7	1.7	5.7	4.3	4.4	1.0	3.6
No effect	EMAX Mixture	1.6	0.6	0.5	0.6	0.7	0.9	0.8
	EMAX logistic	5.3	1.9	1.7	2.0	2.6	3.3	2.8
	Independent logistic	9.9	4.8	21.9	18.7	23.5	5.8	14.1

Table 3.7 MSE $\times 10^5$ of the expected estimated posterior ePTB rate $E(\hat{p}_d|data)$

Scenario	Model	Dose(g/day)						Average
		0	0.2	0.45	0.6	0.8	1	
Optimistic	EMAX Mixture	0.4	-1.3	0.0	0.8	1.7	2.4	0.7
	EMAX logistic	1.1	-1.3	-0.2	0.5	1.3	1.9	0.5
	Independent logistic	-0.1	0.0	0.1	-0.1	-0.1	0.0	0.0
Expected	EMAX Mixture	1.9	-1.3	-0.5	0.3	1.3	2.1	0.6
	EMAX logistic	1.6	-1.4	-0.5	0.4	1.4	2.2	0.6
	Independent logistic	-0.1	0.0	0.1	-0.1	-0.1	-0.1	0.0
Pessimistic	EMAX Mixture	1.4	-0.7	-0.2	0.2	0.7	1.1	0.4
	EMAX logistic	1.2	-0.6	0.1	0.5	1.0	1.5	0.6
	Independent logistic	0.0	0.0	-0.1	0.0	0.0	0.0	0.0
No effect	EMAX Mixture	1.6	0.4	0.0	-0.2	-0.2	-0.3	0.2
	EMAX logistic	2.0	0.5	-0.1	-0.2	-0.2	-0.1	0.3
	Independent logistic	-0.1	0.1	0.0	0.1	-0.2	0.0	0.0

Table 3.8 Bias$\times 10^3$ of expected estimated posterior ePTB rate $E(\hat{p}_d|data)$

3.4 Application to a simulated data set

To illustrate the three models' application, we simulated a dataset using the expected scenario in Table 3.1. We then applied the three models to the simulated dataset and reported the analysis results.

3.4.1 Generating the simulated dataset

According to Section 2.5, to simulate a cohort with the true ePTB rates in the expected scenario in Table 3.1, a_1 =-2.64, a_2 = -3.6, a_3= 2.5, and a_4 = -1.72. The mixture weights for each dose were calculated using the formulas in Section 2.5 and they are given in Table 3.9. Gestational ages were then simulated using the normal mixture distributions. The descriptive statistics of the simulated data are given in Table 3.10. For dose 0 g/day, 0.2 g/day, 0.45 g/day, 0.6 g/day, 0.8 g/day, and 1g/day, we simulated 325, 690, 150, 175, 140, and 550 gestational ages, respectively. The mean gestational ages were 39.0, 39.1, 39.2, 39.1, 39.1, 39.2 weeks,

respectively. The proportions of ePTB (<34 weeks) were 3.7%, 3.2%, 2.0%, 1.7%, 1.4%, and 1.5%, respectively. The proportions of PTB (<37 weeks) were 8.0%, 7.8%, 5.3%, 6.9%, 7.1%, and 6.0%, respectively.

Mixture weights	**Dose(g/day)**					
	0	**0.2**	**0.45**	**0.6**	**0.8**	**1**
Δ_{1d}	0.0570	0.0442	0.0337	0.0292	0.0246	0.0212
Δ_{2d}	0.1430	0.1449	0.1465	0.1472	0.1479	0.1484
Δ_{3d}	0.8000	0.8108	0.8197	0.8236	0.8275	0.8304

Table 3.9 Mixture weights used to simulate dataset.

Dose (g/day)	**n**	**mean**	**sd**	**min**	**Q1**	**median**	**Q3**	**max**	**ePTB (%)**	**PTB (%)**
0	325	39.0	1.8	30.3	38.6	39.3	40.0	42.0	3.7%	8.0%
0.2	690	39.1	2.0	25.0	38.6	39.4	40.1	42.5	3.2%	7.8%
0.45	150	39.2	2.0	22.5	38.6	39.3	40.2	43.4	2.00%	5.3%
0.6	175	39.1	2.1	23.8	38.6	39.4	40.1	42.5	1.7%	6.9%
0.8	140	39.1	1.6	31.7	38.4	39.3	40.2	42.7	1.4%	7.1%
1	550	39.3	1.7	23.9	38.8	39.4	40.1	42.2	1.5%	6.0%

Table 3.10 Descriptive statistics of the simulated dataset.

3.4.2 Analysis of the simulated data

The simulated gestational ages were analyzed as a continuous variable using the EMAX Mixture model. The numbers of ePTBs were analyzed as a binomial variable using the EMAX logistic model and independent doses logistic model. The STAN code for the three models can be found in the appendix.

The posterior probabilities of ePTB and their credible intervals are reported in Figure 3.2. The credible intervals of all three models covered the true values (box in the plot). The

independent doses logistic model had the widest credible intervals. The EMAX mixture model and the EMAX logistic model had similar lengths of credible intervals.

Both the EMAX Mixture model and EMAX logistic model had a $\Pr(a_2 < 0)$ greater than their corresponding cutoffs and can be claimed as successful: EMAX Mixture, $\Pr(a_2 < 0)$ = 0.994 > 0.845; EMAX logistic, $\Pr(a_2 < 0)$ = 0.989 > 0.955. The independent doses logistic model had a $\max\left(pr(p_d < p_0 | data)\right)$ = 0.984, which is less than the cutoff of 0.992. The trial was not a success when the independent doses logistic model was used.

As mentioned in Section 2.2, one advantage of the EMAX Mixture model is that it can estimate the posterior probability of different cut points. For example, the posterior probability of PTB (<37weeks) was 8.66%, 7.34%, 6.59%, 6.31%, 6.05%, and 5.87%, respectively. If we were to estimate the probabilities of PTBs (<37weeks) using the EMAX logistic model and independent doses logistic model, we would have to conduct another set of analyses using the numbers of gestational ages < 37 weeks.

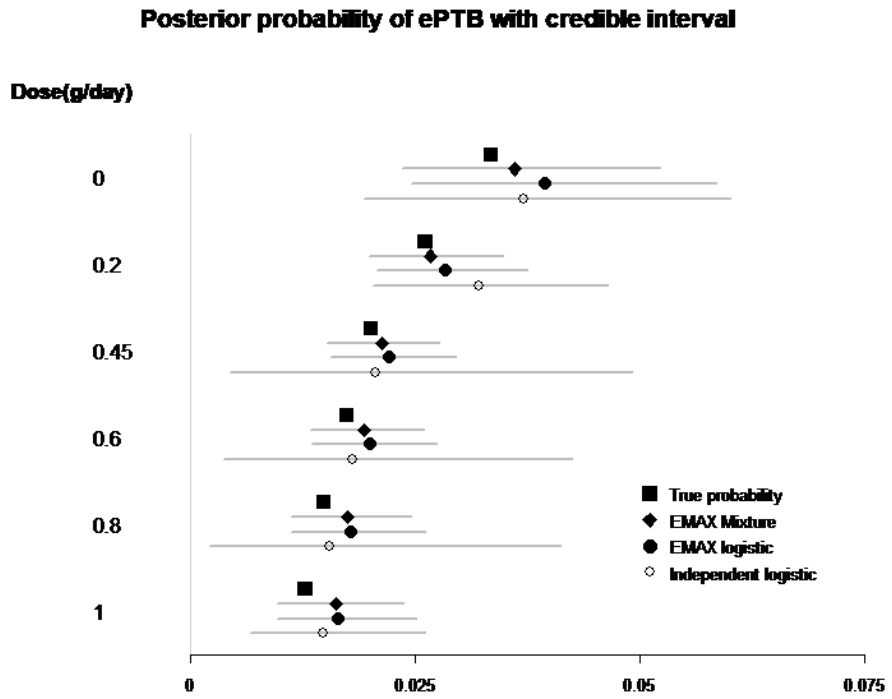

Figure 3.2 Analysis result for the simulated dataset: the posterior probability of ePTB.

3.5 Conclusion and discussion

The EMAX model has never been applied to finite mixture distributions. The Bayesian EMAX Mixture model we proposed applies the EMAX model to a three normal components finite mixture distribution developed for gestational age by Schwartz et al. We compared the EMAX Mixture model with the EMAX logistic model and the independent doses logistic model using extensive simulations. Across different scenarios, the EMAX Mixture model achieved significantly higher power in detecting DHA effect on ePTB and resulted in much smaller MSE in the posterior expected estimate of ePTB rate. The EMAX Mixture model had comparable bias to the EMAX logistic model, but was slightly worse than the independent doses logistic model.

Another attractive feature of the EMAX Mixture model is that it allows for statistically efficient estimates of PTB rates using various cut points with the same parsimonious model. For

example, we can estimate the rate of early preterm birth (<34 weeks gestation), preterm birth (<37 weeks gestation), and late-term birth (>41 weeks gestation) using the same model. In future work, when we conduct analyses on the data collected in the 6 RCTs, it will be valuable to report these estimates.

Though the EMAX Mixture model was motivated by the three normal finite mixture model used for gestational age, it can have a much wider range of applications. It can be modified to accommodate almost all kinds of mixture distributions. For example, if there are two, instead of three, normal components in the mixture distribution, the EMAX Mixture model can be easily adapted by removing θ_{2d} from the model and the mixture weight can be written as: $\Delta_{1\mathrm{d}} = \frac{\exp\left(a_1+\frac{a_2 d}{d_i+a_3}\right)}{1+\exp\left(a_1+\frac{a_2 d}{d+a_3}\right)}$, and $\Delta_{2\mathrm{d}} = \frac{1}{1+\exp\left(a_1+\frac{a_2 d}{d+a_3}\right)}$. Additionally, the EMAX Mixture model can also be applied to non-normal finite mixture distributions by modifying the density function $f(T_{di}|\Delta_{1\mathrm{d}}, \Delta_{2\mathrm{d}}, \Delta_{3\mathrm{d}})$ accordingly.

One limitation of our study is that we assumed the mean and variance of the three normal distributions for gestational age determined by Schwartz et al. from NCDBR fit the new data well. Although these parameters had been used successfully in the past (12,13,14), it is possible but unlikely that the data from the 6 RCTs under consideration are very different from the NCDBR registry. In that case, one possible solution is to allow the model to estimate the mean and variance of the three normal distributions. The model will be more complicated and may have convergence issues. This will be explored in our future work.

References

1. Middleton P, Gomersall JC, Gould JF, Shepherd E, Olsen SF, Makrides M. Omega-3 fatty acid addition during pregnancy. Cochrane Database Syst Rev. 2018; 11 (11)
2. Thomas N, Sweeney K, Somayaji V. Meta-analysis of clinical dose-response in a large drug development portfolio. Stat Biopharm Res. 2014;6(4):302-317.
3. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ (Clinical research ed). 2006; 332:1080.
4. Deyi BA, Kosinski AS, Snapinn SM. Power considerations when a continuous outcome variable is dichotomized. Journal of biopharmaceutical statistics. 1998; 8:337–52.
5. Peacock JL, et al. Dichotomising continuous data while retaining statistical power using a distributional approach. Statistics in medicine. 2012; 31:3089–103.
6. Schwartz, S., Gelfand, A., and Miranda, M. , "Joint Bayesian Analysis of Birthweight and Censored Gestational Age Using Finite Mixture Models," Statistics in Medicine. 2010; 29: 1710–1723.
7. Carlson SE, et al. DHA supplementation and pregnancy outcomes. The American journal of clinical nutrition. 2013; 97:808–15.
8. U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry. Nov. 2019.
9. R Core Team(2017). R: a Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing

10. Stan Development Team (2017) RStan: the R interface to Stan, version 2.16.1. (Available from http://mc-stan.org.)
11. Stan Development Team: Stan Modeling Language User's Guide and Reference Manual, Version 2.16.0. (Available from http://mc-stan.org.)
12. Lei Y, et al. Comparison of Dichotomized and Distributional Approaches in Rare Event Clinical Trial Design: a Fixed Bayesian Design. Journal of applied statistics. 2017; 44(8): 1466–1478
13. Yelland LN, et al. Predicting the effect of maternal docosahexaenoic acid (DHA) supplementation to reduce early preterm birth in Australia and the United States using results of within country randomized controlled trials. Prostaglandins Leukot Essent Fatty Acids. 2016 Sep; 112: 44–49.
14. Gajewski, BJ, Reese, CS, Colombo, J, & Carlson, S (2016), "Commensurate Priors on a Finite Mixture Model for Incorporating Repository Data in Clinical Trials," Statistics in Biopharmaceutical Research, 8(2), 151-160. (PMCID: PMC4915595).

Chapter 4 : On the use of Bayesian models in weight loss clinical trials: a demonstration with a re-analysis of the REPOWER study

4.1 Introduction

Obesity is a chronic condition affecting an increasing number of Americans. The prevalence increased from 14% in 1980(1) to 35% in 2010(2). And in 2017-2018, it increased to 42% (3). Obesity is a serious health risk and is associated with a wide range of morbidities (4). The Centers for Medicare and Medicaid Services (CMS) approved to cover Intensive Behavioral Therapy for Obesity (IBT) with up to 22 individual 15-minute face-to-face visits over a 12-month period in 2011(5). The CMS employs a fee-for-service delivery model which has been challenged and questioned. A variety of care delivery models have arisen in addition to the traditional face-to-face office visit. The Rural Engagement in Primary Care for Optimizing Weight Reduction (REPOWER) (6) is a cluster randomized clinical trial comparing the fee-for-service individual delivery model to two alternatives: in-clinic group visits and phone-based group visits. Participant weight was measured at baseline, 6, 18, and 24 months by trained staff. The primary outcome was weight change at 24 months. Secondary outcomes included the proportion of participants who achieved 5% and 10% weight loss at 24 months.

In the original analyses, frequentist methods were used: A linear mixed-effect multilevel model for examining the percent weight loss over time and two separate generalized linear models for comparing the percentage of participants achieving 5% and 10% thresholds. Although frequentist paradigm has been the predominant approach to clinical studies in the past several decades, some limitations associated with the frequentist null hypothesis testing that reports dichotomized P values have been recognized in statistic society (7,8). On the other hand, the Bayesian paradigm derives probability estimates of model parameters reflecting the clinical interest and can provide better data interpretation. It has gained popularity in recent years owing to the advancement in powerful computing capacity and the invention of efficient Bayesian

statistic software. In this article, we reanalyzed the percent weight loss over time in REPOWER using a Bayesian hierarchical model to address some limitations of the frequentist approach.

Moreover, the original analyses took into consideration the clustering of sites but ignored the clustering of group assignment in the two group-based arms. Because group assignment is only relevant to the in-person group visits and phone-based group visits arms, it is hard to assessing the impact of group assignment on the effect of delivery models. Bayesian approach can easily handle complex problems using the same statistical framework. In this article, we used a four-level hierarchical model with an additional level to assess the group assignment impact on the effect of delivery models on weight loss.

4.2. Methods

4.2.1 Study design and data structure

REPOWER is a cluster randomized clinical trial with thirty six primary practices from three affiliations (KUMC, UNMC, and Marshfield clinic) randomly assigned to one of the three study arms in equal numbers: 1) in-clinic individual visits in which the participants received 15-minute face-to-face individual counseling sections; 2) in-clinic group visits in which the participants received group visits held at practices with a median of 14 participants per group; 3) Phone-based group visits in which participants received lifestyle intervention delivered remotely via audio-only conference calls with a median of 14 participants per group. 1407 participants enrolled in total. Weight was measure at baseline, 6, 18, and 24 months by trained staff. The primary outcome was change in weight (kg) at 24 months. Secondary outcomes included percent weight loss and the proportion of participants who achieved 5% and 10% weight loss at 24 months. The detailed information about the trial conduction has been published by Befort et al (6). In this article, we will first analyze the percent weight loss using a three-level Bayesian

hierarchical model to compare the effect of different intervention delivery models on weight loss. A second Bayesian hierarchical model includes an additional level, group assignment, for the two group-based delivery methods to assess whether group assignment impacted the result.

4.2.2 Model one: three level Bayesian hierarchical model for percent weight loss

Let y_{ijt} be the percent weight loss for participant j from site i at time t. x_1 and x_2 are the arm indicators: (0,0) for in-clinic individual visits, (1,0) for in-clinic group visits, and (0,1) for phone-based group visits. t_{18} and t_{24} are the time indicators: (0,0) for month 6, (1,0) for month 18, and (0,1) for month24. We also include arm and time interactions so that delivery model effect can be evaluated at each time point. To be consistent with the original model, we include affiliation as covariates (denoted by x_3 and x_4). The three-level Bayesian hierarchical model can be represented as follows.

$$y_{ijt} = \alpha_{0ij} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 t_{18} + \beta_4 t_{24} + \beta_5 x_1 * t_{18} + \beta_6 x_1 * t_{24} + \beta_7 x_2 * t_{18} + \beta_8 x_2 * t_{24} + \beta_9 x_3 + \beta_{10} x_4 + \epsilon_{ijt}$$

$$\alpha_{0ij} = \alpha_{0i0} + \gamma_j$$

$$\alpha_{0i0} = \alpha_{000} + \eta_i$$

Where a_{000} is the intercept; $\epsilon_{ijt} \sim N(0, \sigma^2)$ is within patient residual error; $\gamma_j \sim N(0, \sigma_\gamma^2)$ is patient level variation; $\eta_i \sim N(0, \sigma_\eta^2)$ is site level variation. To make apples to apples comparison with the frequentist analyses, noninformative priors are used: flat priors for a_{000} and βs; truncated $N^+(0, 10^2)$ for σ, σ_r, and σ_η so that only positive values are allowed.

Let $\Delta_1 = a_{000} + \beta_4 + \frac{1}{3}\beta_9 + \frac{1}{3}\beta_{10}$, $\Delta_2 = a_{000} + \beta_1 + \beta_4 + \beta_7 + \frac{1}{3}\beta_9 + \frac{1}{3}\beta_{10}$, and $\Delta_3 = a_{000} + \beta_2 + \beta_4 + \beta_8 + \frac{1}{3}\beta_9 + \frac{1}{3}\beta_{10}$. They represent the expected percent weight loss at 24 months for in-clinic individual visits arm, in-clinic group visits arm, and phone-based group

visits arm, respectively. Their posterior distributions can be obtained from the MCMC samples of a_{000} and $\beta's$. The estimate of the expected percent weight loss for the three treatment arms can be obtained using their mean of MCMC samples. The absolute difference in weight loss when compared to the in-clinic individual visits arm, can be assessed using the posterior distribution of $\delta_2 = \Delta_2 - \Delta_1 = \beta_1$ for in-clinic group visits arm and $\delta_3 = \Delta_3 - \Delta_1 = \beta_2 + \beta_8$ for phone-based group visits arm. The probability of having a bigger weight loss compared with the in-clinic individual arm can be calculated as $\int_0^\infty \Pr(\delta_2|data)\, d\delta_2$ for in-clinic group arm and $\int_0^\infty \Pr(\delta_3|data)\, d\delta_3$. Additionally, the posterior distribution of rate of achieving 5% weight loss at 24 months for each arm can be obtained by $\int_5^\infty \phi\left(x|\Delta_1, \sigma^2 + \sigma_r^2 + \sigma_\eta^2\right) dx$, $\int_5^\infty \phi\left(x|\Delta_2, \sigma^2 + \sigma_r^2 + \sigma_\eta^2\right) dx$, and $\int_5^\infty \phi\left(x|\Delta_3, \sigma^2 + \sigma_r^2 + \sigma_\eta^2\right) dx$, where $\phi(T|\mu, \sigma^2)$ is the normal probability density function with mean μ and variance σ^2. The posterior distribution of achieving 10% weight loss at 24 months can be obtained by simply change the lower integration bound to 10.

4.2.3 Model two: Bayesian hierarchical model for percent weight loss with group assignment

In the in-clinic group visits arm and the phone-based group visits arm, the participants received the interventions in groups. We want to exam the impact of group assignment on the effect of group-based delivery models. Because group assignment is relevant only to the in-person group visits and phone-based group visits arms, it is challenging to be incorporated into the linear mixed model framework and it was not tackled in the original analyses. In model two, we will use a four-level hierarchical Bayesian model to assess the impact of intervention group.

For participants in In-clinic individual visits arm, the model is the same as in Model one. For participants in In-clinic group visits and Phone-based group visits, let k index the intervention group. The four-level Bayesian hierarchical model can be represented as follows.

$$y_{ijkt} = \alpha_{0ijk} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 t_{18} + \beta_4 t_{24} + \beta_5 x_1 * t_{18} + \beta_6 x_2 * t_{18} + \beta_7 * t_{24} + \beta_8 x_2 * t_{24} + \beta_9 x_3 + \beta_{10} x_4 + \epsilon_{ijkt}$$

$$\alpha_{0ijk} = \alpha_{0i0k} + \gamma_j$$

$$\alpha_{0i0k} = \alpha_{0i00} + \theta_k$$

$$\alpha_{0i00} = \alpha_{0000} + \eta_i$$

where $\theta_k \sim N(0,\ \sigma_\theta^2)$ denotes the variation due to intervention group. Truncated normal distribution $+N(0, 10^2)$ is used for σ_θ^2. To answer whether intervention group impacts on the effect of delivery methods, we use two model selection methods, leave-one-out cross-validation (loo-cv) and widely available information criterion (WAIC) (9), to determine whether Model two improves model fit.

4.2.4 Posterior distribution computation and software

Hamiltonian Monte Carlo (10) was performed in Stan (11) to obtain the posterior distributions for parameters of interest. Figure representations of posterior distributions were computed from gaussian kernel density estimates, which provided a smoothed version of the sampled histograms. R package *Rstan* was used to as the interface to call Stan code(12). All the other analyses and plots were conducted in R.

4.3 Results

4.3.1 Model convergence assessment and predictive checking

For both models we ran two parallel MCMC chains with starting points randomly generated from the prior distributions. For each chain, we allowed 3000 iterations for the sampler to converge and another 3000 for sampling the posterior distributions. Convergence were checked visually utilizing caterpillar plots and $\hat{R}$ (13).

4.3.2 Model result

4.3.2.1 Model I

Table 4.1 summarizes the model parameters using posterior means and 95% credible intervals (CrI) based on their MCMC samples of posterior distributions. Because non-informative priors were used, the means and 95% credible intervals were very close to the result from the original linear mixed-effect multilevel model.

		Mean	Standard Deviation	2.50%	97.50%
Intercept	a_{000}	-6.25	0.53	-7.27	-5.2
In-clinic group	β_1	-2.65	0.67	-3.96	-1.32
Phone-based group	β_2	-1.95	0.66	-3.28	-0.64
18 months	β_3	2.26	0.27	1.72	2.78
24 months	β_4	3.05	0.27	2.53	3.61
In-clinic group*18 months	β_5	0.4	0.4	-0.36	1.16
Phone-based group*18 months	β_6	0.1	0.39	-0.65	0.86
In-clinic group:24 months	β_7	0.82	0.39	0.05	1.59
Phone-based group*24 months	β_8	0.5	0.39	-0.27	1.27
Affiliation: Marshfield Clinic	β_9	-0.05	0.6	-1.25	1.15
Affiliation: UNMC	β_{10}	2.17	0.79	0.64	3.75
Sigma	σ	3.91	0.06	3.8	4.03
Site level variation	σ_η	0.97	0.37	0.18	1.68
Patient level variation	σ_r	6.66	0.15	6.38	6.97

Table 4.1 Posterior mean and 95% credible interval for model parameters in Model 1.

Figure 4.1A displays the posterior distribution of the expected weight loss at 24 months for the three arms (i.e. Δ_1, Δ_2, and Δ_3). The posterior mean and credible interval were 2.5%[95% CrI: 1.5, 3.5] for the in-clinic individual visits, 3.9%[95% CrI: 2.9, 4.9] for the phone-based group visits, and 4.3[95% CrI: 3.4, 5.3] for the in-clinic group visits. Because the non-informative priors we used, the estimates were almost identical to the result reported in the original analysis: 2.5%[95%: 1.4, 3.5] for the in-clinic individual visits, 3.8[95% CI: 2.8,4.9] for the phone-based group visits, and 4.3[95% CI: 3.3, 5.3] for the in-clinic group arm.

Figure 4.1B displays the posterior distribution of the absolute difference in the expected percent weight loss at 24 months for the in-clinic group visits and the phone-based group visits when compared with the in-clinic individual visits. The posterior mean and 95% credible interval were 1.4%[95% CrI: 0.1, 2.8] for the phone-based group visits and 1.8%[95% CrI: 0.5,3.2] for the in-clinic group visits. The shaded areas (AUCs to the right of zero) represent the probability of having a greater weight loss in the phone-based group visits arm (98.1%) and in the in-clinic group visits arm (99.6%) than in the in-clinic individual visits arm. The original analysis reported there was a significant difference between the in-clinic group visits vs. the in-clinic individual visits (1.8%[95% CI: 0.4, 3.2; p value: 0.01]), but not in the phone-based visits (1.3[95% CI: -0.1, 2.8; p value: 0.06]) because the p value was slightly bigger than 0.05.

Figure 4.2A and Figure 4.3A display the posterior distribution for the probability of achieving 5% and 10% weight loss at 24 months. In the order of in-clinic individual visits, phone-based group visits, and in-clinic group visits, the posterior mean and the 95% credible interval were 37.4%[95% CrI: 33.7, 42.3], 44.6%[95% CrI: 40.0, 49.5], and 46.6%[95% CrI: 41.7, 51.7] for achieving 5% weight loss; 16.8%[95% CrI: 13.7, 20.4], 21.9%[95% CrI: 18.1, 25.8], and 23.4%[95% CrI: 19.7, 27.6] for achieving 10% weight loss. In the original analyses, two separate generalized mixed models were used to assess the proportions of 5% and 10% weight loss. The estimated proportion of 5% weight loss were 36.0% [95% CI:30.2, 42.3], 41.4%[95% CI: 37.9, 50.6], and 44.1% [95% CI: 35.2, 47.8]. The estimated proportion of 10% weight loss were 17.1% [95% CI: 13.3, 21.8], 22.3% [95% CI: 17.9, 27.6], and 22.6% [95% CI: 18.1, 27.9]. While the Bayesian point estimates for proportions of achieving 10% and 5% weight loss were close to original result, the interval widths were narrower in the Bayesian model because it leveraged the continuous model.

Figure 4.2B and Figure 4.3B display the absolute differences in the probability of achieving 5% and 10% weight loss for the phone-based group visits and the in-clinic group visits when compared with the in-clinic individual visits at 24 months. In the order of phone-based group visits and in-clinic group visits, the posterior mean and 95% credible interval were 7.1%[95% CrI: 0.3, 13.4] and 9.2% [95% CrI: 2.5, 15.9] for achieving 5% weight loss, and 5.0%[95% CrI: 0.2, 9.7] and 6.6% [95% CrI: 1.8, 11.5] for achieving 10% weight loss. The shaded areas (AUCs to the right of zero) represent the probability of having a higher rate of achieving 5% and 10% weight loss in the in-clinic group visits arm and in the phone-based group visits arm. For both 5% and 10% weight loss, the probabilities were 99.6% for in-clinic group arm and 98.4% for the phone-based group arm and they are consistent with the probability of having a greater weight loss than the in-clinic individual visits arm as shown in Figure 4.1B. In the original analyses, the odds ratios of achieving >5% and >10% weight loss were reported for the in-clinic group visits and the phone-based group visits: 1.4 [95% CI: 1.0, 2.0; p value: 0.07] and 1.3 [95% CI: 0.9, 1.8; p value: 0.22] for 5% weight loss, and 1.4 [95% CI: 0.9, 2.1; p value: 0.09] and 1.4 [95% CI: 0.9, 2.1; p value: 0.11]. The authors concluded there was no significant difference in rate of achieving 5% and 10% weight loss for in-clinic group vs. in-clinic individual and phone-based group vs. in-clinic individual.

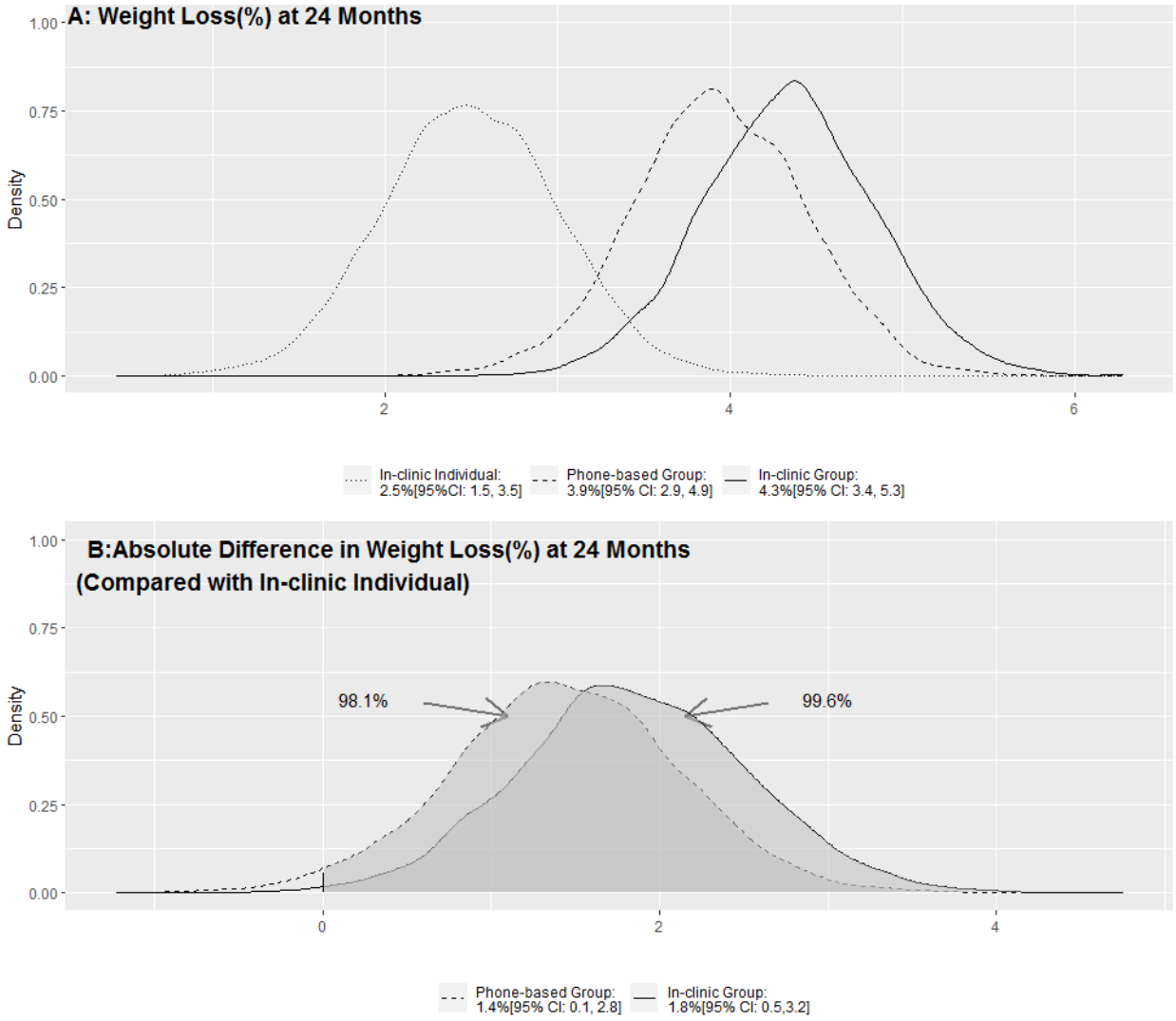

Figure 4.1 Posterior distributions of the expected weight loss(A) and the absolute difference in weight loss(B) at 24 months

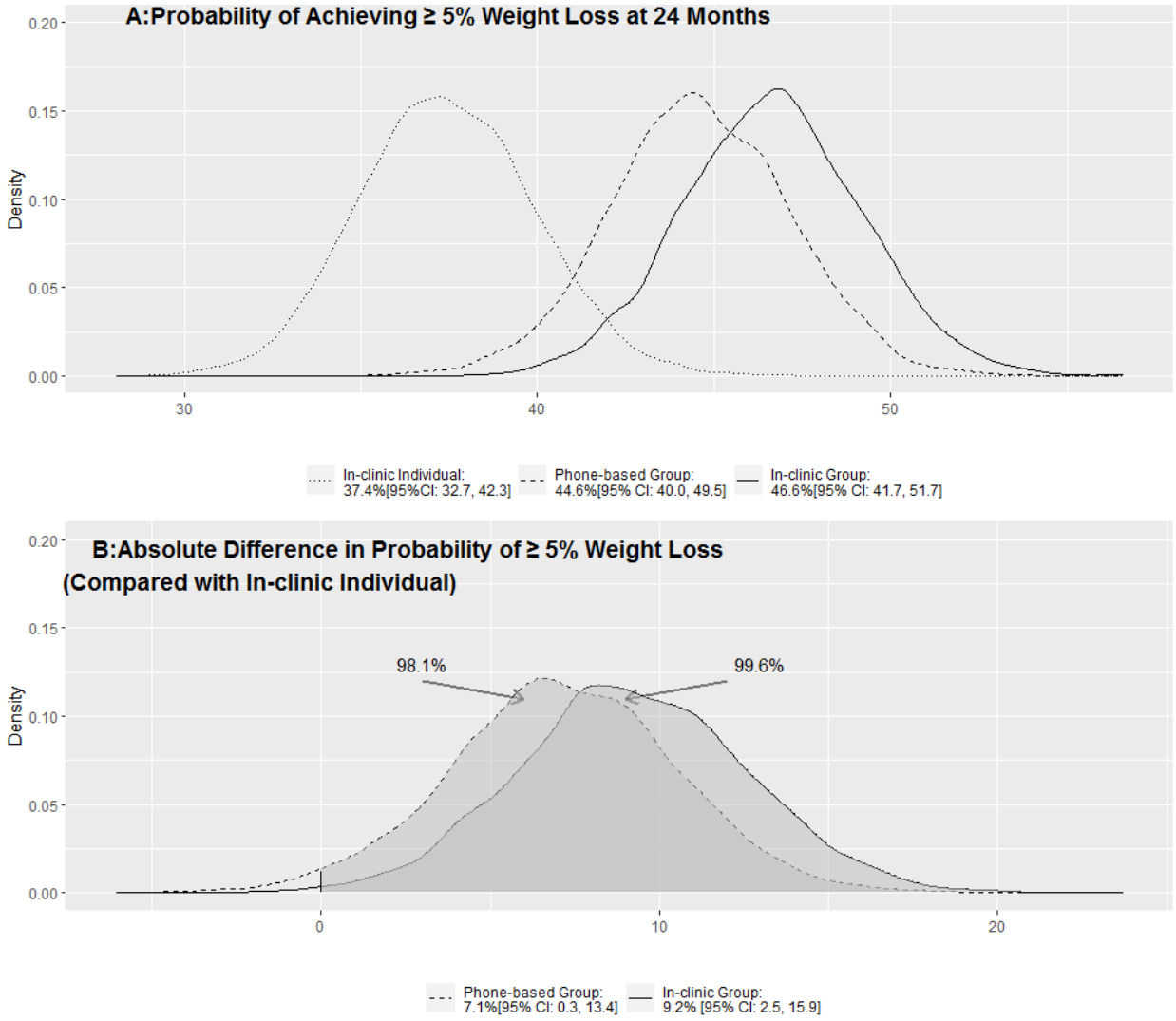

Figure 4.2 Posterior distributions of the probability of achieving 5% weight loss(A) and Posterior distributions of the absolute difference in the probability of achieving 5% weight loss when compared with in-clinic individual visits(B).

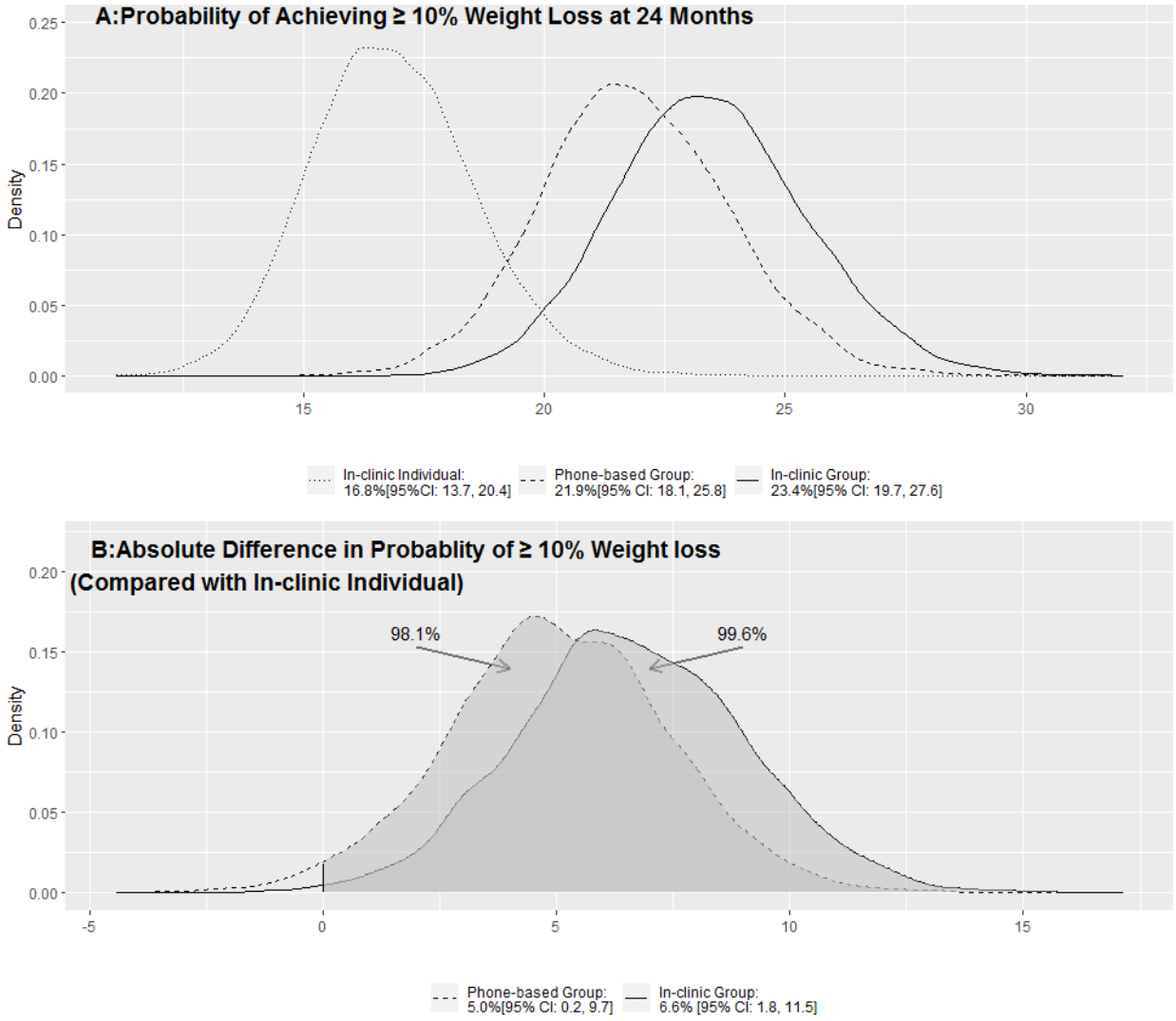

Figure 4.3 Posterior distributions of the probability of achieving 10% weight loss(A) and Posterior distributions of the absolute difference in the probability of achieving 10% weight loss when compared with in-clinic individual visits(B).

4.3.2.2 Model 2

Table 4.2 shows the posterior means and 95% credible intervals for model parameters in Model 2 based on their MCMC samples of the posterior distributions. The values were very close to Model 1 for the parameters in common. The mean and 95% credible interval for σ_θ were 0.59[95% CrI: 0.03, 1.44]. To assess whether including group assignment as an additional hierarchical level will improve model fit, we used leave-one-out cross-validation (loo-cv) and

widely available information criterion (WAIC) implemented in the *loo* R package (14). Both looic and WAIC were slightly smaller in Model 1: 22014 vs. 22017 and 21812 vs. 21819, respectively. The differences were small in comparison with their standard error: -3(6.8) and -7(4.1). We concluded that Model 2 didn't improve model fit. Therefore, including group assignment in the model didn't impact the effect of the intervention delivery models.

		Mean	Standard Deviation	2.50%	97.50%
Intercept	a_{000}	-6.28	1.31	-8.94	-3.65
In-clinic group	β_1	-2.64	1.37	-5.36	0.14
Phone-based group	β_2	-1.91	1.39	-4.62	0.93
18 months	β_3	2.26	0.28	1.71	2.79
24 months	β_4	3.05	0.28	2.5	3.59
In-clinic group*18 months	β_5	-0.04	0.49	-1	0.93
Phone-based group*18 months	β_6	2.18	0.68	0.86	3.5
In-clinic group:24 months	β_7	0.42	0.39	-0.34	1.17
Phone-based group*24 months	β_8	0.11	0.4	-0.66	0.9
Affiliation: Marshfield Clinic	β_9	0.84	0.39	0.06	1.6
Affiliation: UNMC	β_{10}	0.52	0.39	-0.25	1.29
Sigma	σ	3.92	0.06	3.81	4.03
Site level variation	σ_η	1.24	0.43	0.25	2.02
Group level variation	σ_θ	0.59	0.39	0.03	1.44
Patient level variation	σ_r	6.62	0.15	6.33	6.92

Table 4.2 Posterior mean and 95% credible interval for model parameters in Model 2.

	Looic (se)	WAIC (se)
Model 1	22013.5(126.9)	21812.0(122.4)
Model 2	22016.5(127.9)	21819.0(123.2)
Model 1 – Model 2	-3(6.8)	-7(4.1)

Table 4.3 Leave-one-out cross validation(loo-cv) and widely available information criterion(WAIC) for Model 1 and Model 2.

4.4 Conclusion and discussion

In the original analyses, the authors concluded that, the in-clinic group visits, compared with the in-clinic individual visits, resulted in a significantly greater weight loss at 24 months (difference: 1.8%[95% CI: 0.4, 3.2; p value: 0.01]), while the difference between phone group versus in-clinic individual was not significantly different (1.3%[95% CI: -0.1, 2.8; p value: 0.06]). The proportions of achieving 5% and 10% weight loss were not significantly different between the in-clinic group vs. in-clinic individual visits(for 5% weight loss: OR=1.4 [95% CI: 1.0, 2.0], p=0.07; for 10% weight loss: OR=1.4 [95% CI: 0.9, 2.1], p=0.09), nor between the phone-based group visits vs the in-clinic individual visits (for 5% weight loss: OR=1.3 [95% CI: 0.9, 1.8], p=0.22; for 10% weight loss: (OR=1.4[95% CI: 0.9, 2.1], p=0.11). By contrast, the Bayesian analyses estimated 99.6% probability that in-clinic group visits, compared with in-clinic individual visits, resulted in a bigger percent weight loss (1.8%[95% CI: 0.5,3.2]), a bigger proportion of achieving 5% threshold (9.2% [95% CI: 2.5, 15.9]), and bigger proportion of achieving 10% threshold (6.6% [95% CI: 1.8, 11.5]). For phone-based group visits, there was a 98.1% probability resulted in a bigger percent weight loss (1.4%[95% CI: 0.1, 2.8]), a bigger proportion of achieving 5% threshold (7.1%[95% CI: 0.3, 13.4]), and bigger proportion of achieving 10% threshold (5.0%[95% CI: 0.2, 9.7]). In Model 2, we concluded that group assignment did not impact model fit and did not impact the effect of intervention delivery methods significantly.

Frequentist analyses base the inference on P-values and confidence intervals (CI). P value is the probability of observing data as extreme or more extreme if the study were to repeat many times under the null hypothesis (no treatment effect). The interpretation of P values is awkward and rarely reflects the questions clinicians interested in. Furthermore, the P values are often

dichotomized using the cut point of 0.05 for decision making. Thus, the original analyses concluded, when compared to in-clinic individual visit, there was a significantly greater weight loss at 24 months for in-clinic group visits (P value: 0.01), but not for phone-based group visits (P value: 0.06) and there were no significant differences for in-clinic group visits vs. in-clinic individual visits and phone-based group visits vs. in-clinic individual group in probability of achieving 5% weight loss(P values: 0.07 and 0.09) and in probability of achieving 10% weight loss(P values: 0.22 and 0.11). The inconsistency in the conclusion for different endpoints demonstrated the drawback of reporting P values in frequentist analyses. Bayesian analyses, on the other hand, enable a more direct interpretation by providing posterior distribution for parameters of interest. They can answer the clinician's questions directly, such as "what is the probability that the participants in one arm will achieve more weight loss than the participants in another arm? ". The Bayesian analysis provided consistent estimates of probabilities of resulting in more weight loss in-clinic group visits vs. in-clinic individual visits (99.6%) and phone-based visits vs. in-clinic individual visits (98.4%) for all three endpoints: percent weight loss, probability of achieving 5% weight loss, and probability of achieving 10% weight loss.

Although the result of Model 2 was not significant, it showed another advantage of Bayesian approach that it can easily handle complex problems using the same statistical framework. In our analysis, we just simply expand the 3-level hierarchical model to a 4-level hierarchical model by including group assignment.

References

1. Flegal KM, Carroll MD, Kuczmarski RJ, Johnson CL. Overweight and obesity in the United States: prevalence and trends, 1960-1994. Int J Obes Relat Metab Disord 1998;22(1):39-47.
2. Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of obesity in the United States, 2009-2010. National Center for Health Statistics Data Brief No 82. 2012.
3. Centers for disease control and prevention. Prevalence of Obesity and Severe Obesity Among Adults: United States, 2017–2018. https://www.cdc.gov/nchs/products/databriefs/db360.htm. (accessed Feb 2021).
4. Hammond RA, Levine R. The economic impact of obesity in the United States. Diabetes, metabolic syndrome and obesity : targets and therapy 2010;3:285-295.
5. Centers for Medicare and Medicaid Service. Decision memo for intensive behavioral therapy for obesity. 2011. http://www.cms.gov/medicare-coverage-database/details/nca decision-memo (accessed Feb 2021).
6. Befort CA, VanWormer JJ, Desouza C, Ellerbeck EF, Gajewski B, Kimminau KS, Greiner KA, Perri MG, Brown AR, Pathak RD, Huang T, Eiland L, Drincic A. Effect of Behavioral Therapy With In-Clinic or Telephone Group Visits vs In-Clinic Individual Visits on Weight Loss Among Patients With Obesity in Rural Clinical Practice. JAMA 2021; 325(4), 363-372.
7. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31(4):337-350. doi: 10.1007/s10654-0160149-3 20.
8. Wasserstein R, Lazar N. The ASA's statement on P values: context, process, and purpose. Am Stat 2016; 70(2):129-133.

9. Watanabe S. A widely applicable Bayesian information criterion. Journal of Machine Learning Research 2013; 14:867–897
10. Betancourt M. A conceptual introduction to Hamiltonian Monte Carlo. Preprint arXiv:1701.02434. Columbia University, New York.
11. Stan Development Team. Stan Modeling Language User's Guide and Reference Manual, Version 2.16.0. (Available from http://mc-stan.org.)
12. Stan Development Team. RStan: the R interface to Stan, version 2.16.1. (Available from http://mc-stan.org.)
13. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner P. Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC. arXiv 2019; arXiv:1903.08008
14. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 2017; 27(5), 1413–1432

Chapter 5 : Summary and Future Directions

In chapter two, we developed an adaptive Bayesian clinical trial design in the setting of comparative clinical research where multiple treatments were of interest and the accrual rate was slow. Extensive simulations were conducted to compare our proposed Reuse-RAR design with a conventional adaptive clinical design where each participant was randomized to one treatment only, a non-adaptive design that reused participants, and a non-adaptive design that did not reuse participants. The simulation result showed that, of the four designs, Reuse-RAR was the most efficient design which achieved a higher power with a shorter trial duration and a smaller number of participants. Conventional-noRAR was the least efficient design. RAR improved efficiency in both Conventional designs and Reuse designs. One limitation of the Reuse designs is that they require participants to be engaged in the study longer than Conventional designs. In the extreme case where participants did not respond to any treatments, the time required to be engaged in a Reuse design could be as long as five times that of Conventional designs. Therefore, Reuse-RAR and Reuse-noRAR were more susceptible to dropouts. Another limitation is that we assumed there was a carryover effect that was consistent across different participants and different treatments. This assumption may not be true in general. However, the model can be modified to better capture the carryover effect according to substantive subject matter knowledge.

In chapter three, we developed an innovative Bayesian dose-response EMAX mixture model that incorporated finite mixture distributions into the EMAX framework. The EMAX model has never been applied to finite mixture distributions. The Bayesian EMAX Mixture model we proposed applied the EMAX model to a three normal components finite mixture distribution developed for gestational age by Schwartz et al. We compared the EMAX Mixture model with the EMAX logistic model and the independent doses logistic model using extensive

simulations. Across different scenarios, the EMAX Mixture model achieved significantly higher power in detecting the DHA effect on ePTB rate and resulted in much smaller MSE in the posterior expected estimate of ePTB rate. The EMAX Mixture model had a comparable bias to the EMAX logistic model, but was slightly worse than the independent doses logistic model. Another attractive feature of the EMAX Mixture model is that it allows for statistically efficient estimates of PTB rates using various cut points with the same parsimonious model. In future work, when we conduct analyses on the data collected in the six RCTs, it will be valuable to report these estimates.

In chapter 4, we reanalyzed the weight loss data from REPOWER using a Bayesian approach. We first analyzed the percent weight loss over time using a Bayesian three-level hierarchical model. The Bayesian analysis provided consistent estimates of probabilities of greater weight loss in-clinic group visits vs. in-clinic individual visits (99.6%) and phone-based visits vs. in-clinic individual visits (98.4%) for all three endpoints: percent weight loss, probability of achieving 5% weight loss, and probability of achieving 10% weight loss. In contrast, the original analyses concluded a significantly greater weight loss at 24 months for in-clinic group visits vs. in-clinic individual visit (P value: 0.01) but not for phone-based group visits vs. in-clinic individual visits(P value: 0.06) and no significant differences between in-clinic group visits vs. in-clinic individual visits and phone-based group visits vs. in-clinic individual group in the probability of achieving 5% weight loss(P values: 0.07 and 0.09) and in the probability of achieving 10% weight loss(P values: 0.22 and 0.11). The inconsistency in the conclusion for different endpoints demonstrated the drawback of reporting P values in frequentist analyses. Additionally, we also used a four-level hierarchical Bayesian model to assess the group assignment impact on the effect of delivery models on weight loss, which was not explored in

the original analysis. Although the result of Model 2 was not significant, it showed another advantage of the Bayesian approach that it can easily handle complex problems using the same statistical framework.

Appendices

Appendix A: Stan code for chapter 2

Stan code for Conventional-RAR design:

```
data {
              int<lower=0> N;
              int<lower=0> K;
              matrix[N,K] X;
              int<lower=0, upper=1> Success[N];
 }
 parameters {
            vector[K] b;
 }
 model {
              b ~ normal(0,5) ;
              Success ~ bernoulli_logit(X*b);
 }
 generated quantities{
           vector[K] p =exp(b)./(1+exp(b));
 }
```

Stan code for Reuse-RAR and Reuse-noRAR design:

```
 data {

            int<lower=0> N;  //number of observations

            int<lower=0> J;  //number of participants

            int<lower=0> K;  //number of treatments

            matrix[N,K] X;   //treatment indicators

            matrix[N,K] X1;  //prior treatment indicators

            int<lower=0> ptid[N];

            int<lower=0, upper=1> Success[N];

 }

 parameters {

            vector[K] b;

            vector[J] theta;

            real<lower=0> sigma;

            real<lower=0> pi1;
```

```
}
model {
        theta~normal(0,1);
        sigma~normal(0,3);
        b~normal(0,5);
        pi1~normal(0,0.5);
        Success ~ bernoulli_logit(X*b+X1*b*pi1+theta[ptid]*sigma);
}
generated quantities{
       vector[K] p =exp(b)./(1+exp(b));
}
```

Appendix B: Stan code for chapter 3

Stan code for EMAX Mixture model

```
data {
  int<lower=1> K; // number of mixture components
  int<lower=1> N; // number of data points
  real y[N]; // observations
  real<lower=0> dose[N];//treatment
  ordered[K] mu;
  vector<lower=0>[K] sigma;
}

parameters{
  real a11;
  real a12;
  real<lower=0> a13;
  real a21;
 }

model {
  vector[K] theta;
  real beta1;
  real beta2;
  vector[K] lps;
  a11 ~ normal (0, 2);
  a12 ~ normal (0, 2);
  a13 ~ normal (0, 1);
  a21 ~ normal (0, 2);

  for (n in 1:N) {
    beta1 = exp(a11+a12*dose[n]/(a13+dose[n]));
    beta2 = exp(a21);
    theta[1] = beta1/(1+beta1+beta2);
    theta[2] = beta2/(1+beta1+beta2);
    theta[3] = 1/(1+beta1+beta2);
    lps=log(theta);
    for (k in 1:K)
       lps[k] += normal_lpdf(y[n] | mu[k], sigma[k]);
    target += log_sum_exp(lps);
    }
}
```

Stan code for EMAX logistic model

```
data {
  int<lower=1> N; // number of data points
  int<lower=0, upper=1> y[N]; // observations
  real<lower=0> dose[N];//treatment
}
parameters{
  real a11;
  real a12;
  real<lower=0> a13;
 }

model {
  real theta;
  a11 ~ normal (0, 2);
  a12 ~ normal (0, 2);
  a13 ~ normal (0.5, 1);
  for (n in 1:N) {
    theta = a11+a12*dose[n]/(a13+dose[n]);
    y[n]~ bernoulli_logit(theta);
  }
}
```

Stan code for independent doses logistic model

```
data {
  int<lower=1> N; // number of data points
  int<lower=0, upper=1> y[N]; // observations
  matrix[N,6] dose;//6 treatments
}
parameters{
  vector[6] beta;
 }
model {
  beta ~ normal(0,5);
  y~ bernoulli_logit(dose * beta);
 }
```

Appendix C: Stan code for chapter 4

Stan code for model 1

```
data {
  int<lower=1> N;  // total number of observations
  vector[N] Y;  // response variable
  int<lower=1> K;  // number of fixed effects
  matrix[N, K] X;  // design matrix

  int<lower=1> N_1;  // number of sites
  int<lower=1> M_1;  // number of site level coefficients
  int<lower=1> J_1[N];  // site indicators
  vector[N] Z_1_1; //site level predictor values

  int<lower=1> N_2;  // number of patients
  int<lower=1> M_2;  // number of patient level coefficients
  int<lower=1> J_2[N];  // patient indication
  vector[N] Z_2_1; //patient level predictor values


  //predictive data
  int<lower=1> N_tilde;
  matrix[N_tilde, (K-1)] X_tilde;
}
transformed data {
  int Kc = K - 1;
  matrix[N, Kc] Xc;  // centered version of X without an intercept
  vector[Kc] means_X;  // column means of X before centering
  for (i in 2:K) {
    means_X[i - 1] = mean(X[, i]);
    Xc[, i - 1] = X[, i] - means_X[i - 1];
  }
}
parameters {
  vector[Kc] b;  // fixed effects
  real Intercept;  // temporary intercept for centered predictors
  real<lower=0> sigma;  // residual SD
  vector<lower=0>[M_1] sd_1;  //site level standard deviations
  vector[N_1] z_1[M_1];  // standardized site level  effects
  vector<lower=0>[M_2] sd_2;  // participant standard deviation
  vector[N_2] z_2[M_2];  // standardized participant effects
}
transformed parameters {
  vector[N_1] r_1_1;  // actual site level effects
  vector[N_2] r_2_1;  // actual participant level effects
```

```
  vector[N] mu = Intercept + rep_vector(0.0, N);
  r_1_1 = (sd_1[1] * (z_1[1]));
  r_2_1 = (sd_2[1] * (z_2[1]));
  for (n in 1:N)   mu[n] += r_1_1[J_1[n]] * Z_1_1[n] + r_2_1[J_2[n]] * Z_2_1[n];
}
model {
  // likelihood including all constants
  target += normal_id_glm_lpdf(Y | Xc, mu, b, sigma);
  // priors including all constants
 // target += normal_lpdf(Intercept | 0, 30);
  target += normal_lpdf(sigma | 0, 10);
  target += normal_lpdf(sd_1 | 0,10);
  target += std_normal_lpdf(z_1[1]);
  target += normal_lpdf(sd_2 | 0,10);
  target += std_normal_lpdf(z_2[1]);
}
generated quantities {
  // actual population-level intercept
  real b_Intercept = Intercept - dot_product(means_X, b);
  vector[N_tilde] y_tilde ;
  vector[N_tilde] y_pred ;
  vector [N_tilde] prob_5;
  vector [N_tilde] prob_10;
  vector [N] log_lik;
    for (i in 1:N) {
      log_lik[i] =normal_lpdf(Y[i] |mu[i] + Xc[i]*b, sigma);
    }
    for (i in 1:N_tilde) {
      y_tilde[i]=normal_rng(b_Intercept+X_tilde[i]*b,
sqrt(sd_1[1]^2+sd_2[1]^2+sigma^2));
      y_pred[i]=b_Intercept+X_tilde[i]*b;
      prob_5[i] = 1-normal_cdf(5, b_Intercept+X_tilde[i]*b,
sqrt(sd_1[1]^2+sd_2[1]^2+sigma^2));
      prob_10[i] = 1-normal_cdf(10, b_Intercept+X_tilde[i]*b,
sqrt(sd_1[1]^2+sd_2[1]^2+sigma^2));
    }
}
```

Stan code for model 2

```
data {
  int<lower=1> N;  // total number of observations
  vector[N] Y;  // response variable
  int<lower=1> K;  // number of fixed effects
  matrix[N, K] X;  //  design matrix
```

```
  int<lower=1> N_1;  // number of sites
  int<lower=1> M_1;  // number of site level coefficients
  int<lower=1> J_1[N];  //site indicators
  vector[N] Z_1_1; //site level predictor values

  int<lower=1> N_2;  // number of participants
  int<lower=1> M_2;  // number of participant level coefficients
  int<lower=1> J_2[N];  // participant indicator
  vector[N] Z_2_1;//Participant level predictor values

  int<lower=1> N_3;  // number of intervention groups
  int<lower=1> M_3;  // number of intervention group level coefficients
  int<lower=1> J_3[N];  // intervention groups indicator
  vector[N] Z_3_1; //intervention groups level predictor values

  //predictive data
  int<lower=1> N_tilde;
  matrix[N_tilde, (K-1)] X_tilde;
}
transformed data {
  int Kc = K - 1;
  matrix[N, Kc] Xc;  // centered version of X without an intercept
  vector[Kc] means_X;  // column means of X before centering
  for (i in 2:K) {
    means_X[i - 1] = mean(X[, i]);
    Xc[, i - 1] = X[, i] - means_X[i - 1];
  }
}
parameters {
  vector[Kc] b;  // fixed effects
  real Intercept;  // intercept for centered predictors
  real<lower=0> sigma;  // residual SD
  vector<lower=0>[M_1] sd_1;  // site level standard deviations
  vector[N_1] z_1[M_1];  // standardized site level effects

  vector<lower=0>[M_2] sd_2;  // participant level standard deviations
  vector[N_2] z_2[M_2];  // standardized participant level effects

  vector<lower=0>[M_3] sd_3;  // intervention groups level standard deviations
  vector[N_3] z_3[M_3];  // standardized intervention groups level effects
}
transformed parameters {
  vector[N_1] r_1_1;  // actual site-level effects
  vector[N_2] r_2_1;  // actual participant-level effects
  vector[N_3] r_3_1;  // actual intervention group-level effects
  vector[N] mu = Intercept + rep_vector(0.0, N);
```

```
  r_1_1 = (sd_1[1] * (z_1[1]));
  r_2_1 = (sd_2[1] * (z_2[1]));
  r_3_1 = (sd_3[1] * (z_3[1]));
    for (n in 1:N) {
      if (X[n, 2] == 0 && X[n,3]==0) mu[n] += r_1_1[J_1[n]] * Z_1_1[n]  + r_3_1[J_3[n]]
* Z_3_1[n];
      else                        mu[n] += r_1_1[J_1[n]] * Z_1_1[n] + r_2_1[J_2[n]] * Z_2_1[n]
+ r_3_1[J_3[n]] * Z_3_1[n];
    }
 }
model {
   // priors including all constants
  target += normal_lpdf(b|0,10);
  target += normal_lpdf(Intercept | -5,5);
  target += normal_lpdf(sigma | 0,5);
  target += normal_lpdf(sd_1 | 0,5);
  target += std_normal_lpdf(z_1[1]);
  target += normal_lpdf(sd_2 | 0,5);
  target += std_normal_lpdf(z_2[1]);
  target += normal_lpdf(sd_3 | 0,5);
  target += std_normal_lpdf(z_3[1]);
}
generated quantities {
  // actual population-level intercept
  real b_Intercept = Intercept - dot_product(means_X, b);

  vector[N_tilde] y_tilde;
  vector [N_tilde] prob_5;
  vector [N_tilde] prob_10;
  vector [N] log_lik;
   for (i in 1:N) {
      log_lik[i] =normal_lpdf(Y[i] |mu[i]+Xc[i]*b, sigma);
    }

    for (i in 1:N_tilde) {
      y_tilde[i]=normal_rng(b_Intercept+X_tilde[i]*b,
sqrt(sd_1[1]^2+sd_2[1]^2+sigma^2));
      prob_5[i] = normal_cdf(-5, b_Intercept+X_tilde[i]*b,
sqrt(sd_1[1]^2+sd_2[1]^2+sigma^2));
      prob_10[i] = normal_cdf(-10, b_Intercept+X_tilde[i]*b,
sqrt(sd_1[1]^2+sd_2[1]^2+sigma^2));
    }
}
```