Improving Clinical Trials Through Enrichment and Historical Controls

By

Chuanwu Zhang

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.

Heather D. Gibbs, Ph.D.

Jianghua (Wendy) He, Ph.D.

Matthew S. Mayo, Ph.D.

Jo A. Wick, Ph.D.

Date Defended: 29 June 2020

The dissertation committee for Chuanwu Zhang certifies that this is
the approved version of the following dissertation:

Improving Clinical Trials Through Enrichment and Historical Controls

Chair: Byron J. Gajewski, Ph.D

Graduate Director: Dr. Jo A. Wick, Ph.D

Date Approved: 29 June 2020

Abstract

In this dissertation, Bayesian adaptive design used to identify subgroup treatment effect is firstly explored. We investigate three Bayesian adaptive models for subgroup treatment effect identification: pairwise independent, hierarchical, and cluster hierarchical achieved via Dirichlet Process (DP). The impact of interim analysis and longitudinal data modeling on the personalized medicine study design is also explored. Interim analysis is considered since they can accelerate personalized medicine studies in cases where early stopping rules for success or futility are met. We apply integrated two-component prediction method (ITP) for longitudinal data simulation, and simple linear regression for longitudinal data imputation to optimize the study design. The designs' performance in terms of power for the subgroup treatment effects and overall treatment effect, sample size, and study duration are investigated via simulation. We found that the hierarchical model with interim analysis and longitudinal modelling is an optimal approach to identifying subgroup treatment effects, and the cluster hierarchical model with interim analysis and longitudinal imputation is an excellent alternative approach in cases where sufficient information is not available for specifying the related priors.

We then investigate several Bayesian designs incorporating historical control borrowing: power prior via overlapping area, commensurate prior, and some other methods. The impact of historical data type and different types of the threshold used in Bayesian decision rule are also explored. The designs' performance in terms of power as a function of treatment effect, sample size, and posterior summary are investigated via simulation. It was found that it is a good consideration to apply the power prior adaptive design with power parameter determination via overlapping area of posterior distribution under certain values of true response rates of concurrent control, historical control, and treatment effect. Study design with commensurate prior is an admissible choice as well, however, appropriate priors need to be specified.

Lastly, we use logistic regression and classification and regression tree (CART) models to identify the risk factors of early preterm birth (ePTB) from maternal perspective based on birth data from Center for Disease Control (CDC) and National Center for Health Statistics (NCHS)' 2014 Natality public file. It revealed that the subgroup with a preterm birth history and a race designation as Black had the highest risk for ePTB. Those findings can provide valuable information for a future enrichment trial design. Moreover, both models can be applied to identify risk factors for other studies.

Acknowledgements

First of all, I would like to express my sincere appreciation and gratitude to my supervisor, Dr. Byron. J. Gajewski. I am so lucky to have the opportunity to learn from and work with Dr. Gajewski through which I would like him to be my advisor. Dr. Gajewski always cares about my research interest and growth, and provides as much resource as he can to develop myself. Definitely, I cannot complete my dissertation without his always advice, encouragement and support. On the other hand, Dr. Gajewski also sets an excellent example that I would like to learn and follow to guide myself how to study, work, get along with others and balance of work and life. To be his student is the asset of my whole life.

I would like to also extend my appreciation and gratitude to the committee members, Dr. Matthew S. Mayo’s great support for my research and funding, Dr. Jianghua (Wendy) He’s excellent comments for my dissertation and tips of being social, Dr. Jo A. Wick’s always support for my Ph.D study and insightful advice for my dissertation, and Dr. Heather D. Gibbs’s providing the great opportunity to collaborate with her project. I will benefit myself from all those valuable experiences for the whole life.

I would like to thank the funding and related institutes to support my research and graduate research assistant (GRA), they are the University of Kansas Cancer Center (P30 CA168524), Biostatistics and Informatics Shared Resource (BISR) of KUMC Biostatistics & Data Science Department, NIH Clinical and Translational Science Award (UL1TR002366). I would like to thank Rickey E. Carter, Bradley H. Pollock, and Lai Wei, whose comments during discussions within the DACTPeRM Working Group helped motivate the research in Chapter Two. We also appreciate the comment and feedback from the editor and two reviewers.

I am greatly thankful to KU Center for Research Computing (CRC), Dr. Devin Koestler and Dr. Dong Pei for their support and guide to speed up the simulation work, to Mr. Dinesh Pal Mudaranthakam and his team's support of technical and software, and to all the faculties and staff for their guide and support in my study and research, and to KUMC (Dykes) library for their permission to incorporate my published work into my dissertation.

I want to thank my fellow colleagues: Lauren Clark, Lili Garrard, Fengming Tang, Zhiwen Wang, Alex Karanevich, Xueyi Chen, Duccan Rotich, Yi Zhong, Jiawei Duan, Guanglin Zhang, Bo Zhang, Xiaqing Huang and Lisa Neums. Their help for my study and the friendship are never faded.

I would like to express my sincere appreciation and gratitude to my master's supervisor, Dr. Xuanjuan Guo who always let me follow my heart when I struggled between my dream and the reality; to my middle school teacher, Mr. Xuling He who helped me establish the confidence that I can do the math and statistics; to my primary school teacher, Mrs Xiaowen Lv who always praised, encouraged and challenged me when I was a boy.

I am thankful to my dear friends Jack New, Lisa Li, Lily Chen, Yubing Yao, Lingling Li, Xiaowen Na (In China), Cong Sui (In UK) and Deshuang Li(In China) who are always there.

I wish to thank my parents for their always support and encouragement over the years. My parents make every effort on my education. I also want to thank my sister and brother in law for their support, devotion and spending more time with my parents when I have been in US over the years.

Lastly, all praises to Buddha and my master – Khenpo Sodargye. I appreciate all the fellows from Bodhi Institute of Compassion and Wisdom in US and Fo Guang Shan Kansas Buddhist Center for all their help and support to make a home away from home.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

The clinical trial is a mandatory process for the development of new medicine. The safety and efficacy of the new medicine must be proved in order to be approved by the health authority before marketing. However, majority of the clinical trials are “negative” (e.g., p value>.05), and it has been estimated that 85% ($200 billions) of the funding spent on the medical research each year is “a waste of money” (Macleod, Michie, et al. 2014). It is necessary to explore some creative studies designs to lower the cost and improve benefit of the clinical trial from statistical perspective. Food and Drug Administration (FDA) has also released some guidance to encourage to research the innovative clinical trial designs reference (Fda. 2012, Fda. 2019). In this dissertation, the related personalized medicine clinical trial and the trials that incorporates historical control are explored.

1.1 Personalized Medicine

In Chapter Two, the design and analysis of clinical trial for personalized medicine is explored. Personalized medicine clinical trials are designed to test for a treatment effect in a particular subgroup (Alosh, Huque, et al. 2017, Zhang, Mayo, et al. 2018). The subgroup factor is patient-specific characteristics, such as biomarkers, demographics, and disease sub-categories.

Recently, researchers have proposed both frequentist and Bayesian approaches to identifying subgroup treatment effect. developed a frequentist non-parametric recursive partitioning method for the analysis of subgroup treatment effects was developed by some researchers (Lipkovich, Dmitrienko, et al. 2011). The random forests of interaction trees (RFIT), was proposed by Su et al.(Su, Peña, et al. 2018) to estimate subgroup treatment effects. Foster et al.(Foster, Taylor, et al. 2011) created the virtual twins method to identify the subgroup treatment effects. Bayesian adaptive designs can also be applied to identify the treatment effect for a particular subgroup (Gajewski, Berry, et al. 2016). Bayesian adaptive designs have a straightforward interpretation and thus are friendly to scientific researchers with little statistical background. Additionally, the Food and Drug Administration (FDA) recently released guidelines that encourage the use of prespecified interim analysis in personalized medicine adaptive designs to evaluate subgroup factors and modify the subpopulation enrollment accordingly (Fda. 2012).

The focus of this research is a prospective study design where different subgroup treatment effects have already been noted but must be investigated in a confirmatory environment among competing treatments that are used in practice (e.g. comparative effectiveness). Thus, this research aims to identify the best treatment by subgroup, avoiding the term “futility”, as one treatment’s futility is another’s success. We investigate three Bayesian adaptive models for subgroup treatment effect identification: pairwise independent, hierarchical, and cluster hierarchical achieved via Dirichlet Process (DP). The impact of interim analysis on the personalized medicine study design is also explored. In our research, interim analyses are specified at a fixed number of subjects enrolled; stopping rules for success are based on posterior probability criteria set for individual subgroups. It should be noted that our research does not adjust the randomization ratio after interim analysis. Longitudinal modelling imputation for missing data is also explored to improve the study design. We apply integrated two-component prediction method (ITP) for longitudinal data simulation, and simple linear regression for longitudinal data imputation to optimize the study design. The designs’ performance in terms of power for the subgroup treatment effects and overall treatment effect, sample size, and study duration are investigated via simulation.

1.2 Historical Control

In Chapter Three, the Bayesian designs incorporated historical controls are explored. Generally, the historical control may come from real world data (RWD, such as medical chart (Clarke and Loudon 2011, Salman, Beller, et al. 2014), patient registry (Gliklich, Dreyer, et al. 2014, Richesson 2011), natural history (NH) trial (Groft 2010)) and completed clinical trials (Bhuyan, Chen, et al. 2015). The historical control is beneficial to patients, especially for those studies aim of rare diseases treatment or unethical to provide placebo to the patients. The FDA has released guidance to regulate how to design a trial that borrows historical information (Fda. 2019), which encourage researches to borrow the historical information. It is good for pharmaceutical companies since they have large amount of related control arm before a trial conducted (Liu 2018), and more resources can be used for the treatment arm.

From statistical perspective, historical control application has some desired properties, such as increase in power, decrease the in size (Liu 2018), minimize the patient burden (Lim, Walley, et al. 2018), etc. The important thought of historical control borrowing is how to connect the historical data to concurrent data. There are several structures of the connection (Spiegelhalter, Abrams, et al. 2004): full equal, discounted equal, biased, similar (i.e., exchangeable), and functional dependent. Then the related methods were derived and applied accordingly. In Chapter Three, we mainly explore the commensurate prior and power prior; with a novel estimation approach in the latter.

The connection between the historical and concurrent control of commensurate prior is the conditional distribution of parameter of concurrent data given the historical data (Gamalo-Siebers, Savic, et al. 2017). The conditional distribution is served as the prior and incorporated with the concurrent data to have the posterior estimation of control parameter. Commensurate prior is essentially a hierarchical model as well. However, it assumes that the historical response rate is non-systematically biased from the current response rate (Lim, Walley, et al. 2018).

There are some explorations of power prior borrowing the historical data (Gravestock and Held 2018, Hobbs, Carlin, et al. 2011, Liu 2018). The degree of power prior borrowing is controlled by the power parameter of power prior. The borrowing changes from “full borrowing” to “no borrowing” as the power parameter goes from 1 to 0. The limitation of power prior is to specify an appropriate power parameter. Some researchers proposed an estimated power parameter to adjust the limitation. Specifically, the power parameter follows a distribution rather than fixed (Neelon and O' Malley 2010). However, this adjustment tends to heavily discount historical data and does not efficiently borrow the historical data unless a very informative prior used for the power parameter (Lim, Walley, et al. 2018).

In Chapter three, we researched the performance of several study designs incorporating historical control via different Bayesian borrowing methods – power prior, commensurate prior and some reference borrowing method. The performance is compared by the simulating trials. The impact of historical data type and different types of the threshold used in Bayesian decision rule are also explored. The designs’ performance in terms of power as a function of treatment effect, sample size, and posterior summary are investigated via simulation.

1.3 Subgroup Identification

It is necessary to identify the subgroup factors, and then explore related statistical methodology accordingly. In Chapter Four, we mainly introduce how to use logistics regression and classification and regression tree (CART) to identify the risk factor of early preterm birth (ePTB) from maternal perspective based on birth data from Center for Disease Control (CDC) and National Center for Health Statistics (NCHS)' 2014 Natality public data file.

The multivariate logistic regression model was applied to estimate odds ratios (OR) and the corresponding 95% confidence intervals (CI) to investigate the association of ePTB with the potential risk factors. All predictors entered the model and they were selected via backward elimination. The predicted probabilities were calculated for the validation cohort based on the model obtained from the training cohort. The calibration plot was generated to compare the average predicted probabilities and the average observed probabilities via the validation cohort. The c-index was calculated to identify the model discriminatory capacity in terms of the training and validation cohorts.

CART model is a useful complement to a logistic regression model because the CART model can identify unknown interactions among the risk factors of ePTB. The most discriminating predictor is selected to partition the data to minimize the subgroup variance of the dependent variable (e.g. ePTB) (Lemon, Roy, et al. 2003). This step is executed repeatedly to the following partitions until the sample size of each subgroup (i.e., a terminal node) is at or below a pre-specified level. Then, a maximum tree was constructed and standard pruning strategies were applied to arrive at a parsimonious tree with a low misclassification rate and a high discriminatory capacity (Breiman, Friedman, et al. 1984). The final CART model can be visualized as an upside-down tree with the parent node of the tree containing the entire sample. The training cohort was used to generate an appropriate CART tree, and the validation cohort was utilized to evaluate the CART tree via the C-index and the misclassification rate. More details regarding the methods and how to apply them to analyze the ePTB data is introduced in Chapter Four.

Chapter 2: Designing and Analyzing Clinical Trials for Personalized Medicine via Bayesian Models

Other Contributors for this Chapter: Matthew S. Mayo, Jo A. Wick, Byron J. Gajewski

Abstract

Patients with different characteristics (e.g., biomarkers, risk factors) may have different responses to the same medicine. Personalized medicine clinical studies that are designed to identify patient subgroup treatment efficacies can benefit patients and save medical resources. However, subgroup treatment effect identification complicates the study design in consideration of desired operating characteristics.

We investigate three Bayesian adaptive models for subgroup treatment effect identification: pairwise independent, hierarchical, and cluster hierarchical achieved via Dirichlet Process (DP). The impact of interim analysis and longitudinal data modeling on the personalized medicine study design is also explored. Interim analysis is considered since they can accelerate personalized medicine studies in cases where early stopping rules for success or futility are met. We apply integrated two-component prediction method (ITP) for longitudinal data simulation, and simple linear regression for longitudinal data imputation to optimize the study design. The designs' performance in terms of power for the subgroup treatment effects and overall treatment effect, sample size, and study duration are investigated via simulation.

We found that the hierarchical model with interim analysis and longitudinal modelling is an optimal approach to identifying subgroup treatment effects, and the cluster hierarchical model with interim analysis and longitudinal imputation is an excellent alternative approach in cases where sufficient information is not available for specifying the related priors. These findings can be applied to future personalized medicine studies with discrete or time-to-event endpoints.

Key words: Bayesian (cluster) hierarchical model, Dirichlet process, Interim analysis, Longitudinal modeling, Integrated two component prediction

2.1 Introduction

Personalized medicine is defined as the tailoring of treatment to patients based on their characteristics, needs, and preferences during medical care . Therefore, personalized medicine clinical trials are designed to test for a treatment effect in patient subgroups (Alosh, Huque, et al. 2017, Zhang, Mayo, et al. 2018). In general, these subgroups are defined using "personalized" or patient-specific characteristics such as biomarkers, demographics, and disease sub-categories. Personalized randomized clinical trials (RCTs) can be categorized as prospective, prospective-concurrent, prospective-retrospective, or retrospective based on the availability of the data relative to the design of the study (Ruberg and Shen 2015). Personalized RCTs are sufficiently powered to test for a treatment effect while controlling both the overall Type I error and the subgroup false positive rates (Alosh, Huque, et al. 2017). However, personalized RCTs that optimize time and resource use without sacrificing statistical rigor are both essential and unexplored.

Recently, researchers have proposed both frequentist and Bayesian approaches to identifying subgroup treatment effect. Lipkovich et al.(Lipkovich, Dmitrienko, et al. 2011) developed a frequentist non-parametric recursive partitioning method for the analysis of subgroup treatment effects. Another non-parametric method, random forests of interaction trees (RFIT), was proposed by Su et al.(Su, Peña, et al. 2018) to estimate subgroup treatment effects. Additionally, Foster et al.(Foster, Taylor, et al. 2011) created the virtual twins method, and Altstein et al.(Altstein, Li, et al. 2011) suggested a new computational method for parameter estimation of an accelerated failure time (AFT) model with subgroups identified by a latent variable. Alosh et al. also introduced the solutions to solve the issues of chance findings, low power of interaction statistical tests for the treatment-by-subgroup interaction, etc. when executing the subgroup analysis from frequentist perspective (Alosh, Huque, et al. 2017).

Compared to the frequentist approaches, Bayesian adaptive designs have potential benefits for prospective personalized RCTs since they naturally extend from simple (Almirall, Compton, et al. 2012) to more complex but efficient models (Bayman, Chaloner, et al. 2010), have higher power for a given type I error rate, and facilitate decision making in advance via interim analysis (Gajewski, Berry, et al. 2015, Wang and Hung 2013). Bayesian adaptive designs also provide the probability that a treatment is best for a particular subgroup (Gajewski, Berry, et al. 2016), which has a straightforward interpretation and thus is friendly to scientific researchers with little statistical background. Additionally, the Food and Drug Administration (FDA) recently released guidelines that encourage the use of prespecified interim analysis in personalized medicine adaptive designs to evaluate subgroup factors and modify the subpopulation enrollment accordingly (Fda. 2018, Fda. 2012). Finally, Bayesian adaptive designs can illustrate the effectiveness of a treatment in subpopulations or the overall population with higher power when compared to a fixed design of the same size (Berry, Broglio, et al. 2013).

The focus of this research is a prospective study design where different subgroup treatment effects have already been noted but must be investigated in a confirmatory environment. A study design in terms of Bayesian models, longitudinal data, and interim analysis is involved (Alosh, Fritsch, et al. 2015, Alosh, Huque, et al. 2017, Dmitrienko, Muysers, et al. 2016). Research has been done for trials whose purpose is to identify a single subgroup (Morita, Yamamoto, et al. 2014), which may be useful for seamless phase II to III designs (Magnusson and Turnbull 2013, Rufibach, Chen, et al. 2016). In addition, Hobbs et al.(Hobbs and Landin 2018) have proposed an innovative sequential basket trial design formulated with Bayesian monitoring rules based on multisource exchangeability and hierarchical modeling.

Some studies (Mehta and Gao 2011, Simon and Simon 2013, Wassmer and Dragalin 2015) refer to RCTs for adaptive personalized medicine. Personalized medicine designs adjust enrollment of subjects for specific subgroups at interims to maximize power and/or shorten study duration (Fda. 2018). It should be noted that our research does not adjust the randomization ratio after interim analysis. Additionally, this research is motivated by comparative effectiveness and thus aims to identify the best treatment by subgroup, avoiding the term "futility", as one treatment's futility is another's success.

One of the trending issues in RCTs for personalized medicine is the handling of multiplicity across subgroups. A well-calibrated RCT will have a Type I error rate of 5% (based on two-sided test) or 2.5% (based on one-sided test), and this frequentist calibration is also crucial for Bayesian RCTs (Grieve 2016, Jenkins, Stone, et al. 2011). Much effort in group sequential designs (Rosenblum, Luber, et al. 2016) is spent controlling the familywise Type I error rate because of the multiple points of testing due to both the number of subgroups and the number of interim analyses. Random effects linear models for identification of subgroup treatment effects with longitudinal data have also been presented (Facts 2018), but little research exists on Bayesian models for longitudinal data with subgroup treatment effects identification. A more effective modeling approach is to borrow strength across the subgroups via a Bayesian hierarchical model. Berry et al. (Berry, Broglio, et al. 2013) concluded that this type of modeling provides a better chance at identifying efficacy or futility than the models that promote independence across subgroups. Gamalo-Siebers et al. (Gamalo-Siebers, Tiwari, et al. 2016) pointed out that in some instances, hierarchical models suffer from "over-shrinkage" and a Dirichlet Process (DP) prior is a possible alternative to the lighter-tailed alternatives. Hierarchical models and DP priors are also candidate models in this research.

This research is the result of a National Center for Advancing Translational Sciences (NCATS) national working group with the name of Designing and Analyzing Clinical Trials for Personalized Medicine (DACTPerM), brought together to explore the properties of several statistical models to be applied to academic medical RCTs for personalized medicine. The exploration is done by simulating trials in which several treatments are tested simultaneously (e.g., two drugs tested in different sub-populations). Interim analyses are specified at a fixed number of subjects enrolled; stopping rules for success are based on posterior probability criteria set for individual subgroups. Longitudinal modelling imputation for missing data is also explored to improve the study design.

In Section 2.2, we introduce the motivating study, Patient Assisted Intervention for Neuropathy: Comparison of Treatment in Real Life Situations (PAIN-CONTRoLS) (Barohn, Gajewski, et al. 2018), and several models for RCTs in personalized medicine are described as well. In Section 2.3, operating characteristics for the different possible designs are presented and compared. We demonstrate the models' simulation-based performance. We conclude with discussion and conclusions in Section 2.4.

2.2 Method

2.2.1 Motivating Study

The objective of the PAIN-CONTRoLS study was to identify the most effective medicine for reducing pain and improving the quality of life in patients with Cryptogenic Sensory Polyneuropathy (CSPN). The study investigates four candidate medicines: nortriptyline, duloxetine, pregabalin, and mexiletine. The study found that both nortriptyline and duloxetine had the highest posterior probability of being the best treatment among the four candidates. However, an exploratory analysis found that nortriptyline and duloxetine had results that varied by subject characteristics such as gender, age, and race. Therefore, we wish to design a future prospective trial that verifies this subgroup hypothesis via an innovative and efficient Bayesian model. The primary endpoint, pain, is an approximately continuous measure of risk reduction in pain (scale 0-10) at 12-weeks relative to that at randomization. Specifically, it is equal to $\frac{P_0 - P_{12}}{P_0}$, where P_0 is pain score at randomization and P_{12} is the one at 12 weeks.

2.2.2 Model Specification

Selecting a model for personalized medicine RCTs is important for optimizing operating characteristics. Generally, it is unlikely that one model can be recommended for all RCTs. The strategy for model selection is to pick the candidate model with the most desirable operating characteristics calculated via simulation. It is also a good strategy to build the candidate models from simple to complex. A pairwise independent subgroup model (i.e., a model for one subgroup is independent of those for the other subgroups) is a straightforward one to begin with. We also consider the hierarchical and cluster hierarchical model since these models adapt depending on the variation of the treatment effect across subgroups.

Generally, we assume the endpoints for all subjects from both treatment arms (A or B), i.e., both Arm A and B are active arms which means our research is based on effectiveness comparison, are normally distributed with identical standard deviations but different means. Specifically, observations from arm A are denoted:

$$Y_{1g}^{(A)}, Y_{2g}^{(A)}, Y_{3g}^{(A)}, Y_{4g}^{(A)} \ldots. \;\; Y_{N_g^{(A)}g}^{(A)} \sim N\left(\gamma_g, \sigma^2\right);$$

and for arm B:

$$Y_{1g}^{(B)}, Y_{2g}^{(B)}, Y_{3g}^{(B)}, Y_{4g}^{(B)} \ldots. \;\; Y_{N_g^{(B)}g}^{(B)} \sim N\left(\gamma_g + \; \theta_g, \sigma^2\right)$$

where g is the index indicating the subgroup and $g \in \{1, 2, 3, \ldots g_n\}$. $N_g^{(A)}$ and $N_g^{(B)}$ represent the sample size of subgroup g for treatment arm A and B, respectively. The common standard deviation is given by σ and the means for arm A and B are γ_g and $\gamma_g + \theta_g$, respectively. Thus, θ_g represents the treatment difference for subgroup g.

Pairwise Independent Model. In a pairwise independence model, separate priors are used for each treatment arm such that each γ_g and θ_g have normal prior distributions,

$$\gamma_g \sim N\left(\mu_g^{(A)}, \tau_g^{(A),2}\right), \theta_g \sim N\left(\mu_g^{(B)}, \tau_g^{(B),2}\right), \text{ and } \sigma^2 \sim IG\left(\frac{\sigma_n}{2}, \; \frac{\sigma_\mu^2 \sigma_n}{2}\right).$$

We assume $\tau_g^{(A),2}$ is equal to $\tau_g^{(B),2}$, and σ_μ and σ_n are the central and weight parameters of the inverse gamma distribution. We use weakly informative priors whose information was obtained from the example study and inflate the related prior variance values to diminish the effect that priors play in the following simulations. The complete conditional distributions of treatment difference (θ_g) and treatment effect from arm A (γ_g), given data and all other parameters, are both normal. Specifically,

$$\theta_g | Y_{1g}^{(B)} .. Y_{N_g^{(B)}g}^{(B)}, \gamma_g, \sigma^2, \mu_g^{(B)}, \tau_g^{(B),2} \sim N \left(\frac{\tau_g^{(B),2} N_g^{(B)} \left(\bar{Y}_g^{(B)} - \gamma_g \right) + \sigma^2 \mu_g^{(B)}}{N_g^{(B)} \tau_g^{(B),2} + \sigma^2}, \frac{\tau_g^{(B),2} \sigma^2}{N_g^{(B)} \tau_g^{(B),2} + \sigma^2} \right) \text{(2.1)},$$

$$\gamma_g | Y_{1g}^{(A)} \ldots, Y_{1g}^{(B)} \ldots, \theta_g, \sigma^2, \mu_g^{(A)}, \tau_g^{(A),2} \sim N \left(\frac{\tau_g^{(A),2} \left(N_g^{(A)} \bar{Y}_g^{(A)} + N_g^{(B)} \bar{Y}_g^{(B)} - N_g^{(B)} \theta_g \right) + \sigma^2 \mu_g^{(A)}}{\left(N_g^{(A)} + N_g^{(B)} \right) \tau_g^{(A),2} + \sigma^2}, \frac{\tau_g^{(A),2} \sigma^2}{\left(N_g^{(A)} + N_g^{(B)} \right) \tau_g^{(A),2} + \sigma^2} \right) \text{(2.2)}$$

Hierarchical Model. The hierarchical model's borrowing strength across subgroups is achieved through shared prior distributions for each treatment. Consequently, $\mu_\gamma^{(A)}$, $\mu_\gamma^{(B)}$ and $\tau_\gamma^{(A),2}, \tau_\gamma^{(B),2}$ are considered random parameters from a set of shared distributions. For treatment arm A (γ_g):

$$\gamma_g \sim N \left(\mu_\gamma^{(A)}, \tau_\gamma^{(A),2} \right), \mu_\gamma^{(A)} \sim N \left(\mu_0, \sigma_0^2 \right), \tau_\gamma^{(A),2} \sim IG \left(\frac{\tau_n}{2}, \ \frac{\tau_\mu^2 \tau_n}{2} \right);$$

and for the difference between treatment arms in subgroup g (θ_g):

$$\theta_g \sim N \left(\mu_\gamma^{(B)}, \tau_\gamma^{(B),2} \right), \mu_\gamma^{(B)} \sim N \left(\mu_0, \sigma_0^2 \right), \tau_\gamma^{(B),2} \sim IG \left(\frac{\tau_n}{2}, \ \frac{\tau_\mu^2 \tau_n}{2} \right).$$

Here, $\mu_\gamma^{(A)}$ and $\mu_\gamma^{(B)}$are independent and identically distributed, as are $\tau_\gamma^{(A),2}$ and $\tau_\gamma^{(B),2}$. We specify the values of the hyperparameters μ_0, σ_0^2, τ_n and τ_μ^2 when simulation is executed. The expressions of the completely conditional distributions of the treatment difference (θ_g) and the treatment effect from arm A (γ_g) given data and all other parameters are identical to (2.1) and (2.2) from the pairwise independent model. However, the complete conditional distributions of $\mu_\gamma^{(A)}$ and $\mu_\gamma^{(B)}$ given data and all other parameters are given by

$$\mu_\gamma^{(B)}|\theta_1..\theta_{g_n},\tau_\gamma^{(B),2},\mu_0,\sigma_0^2 \sim N\left(\frac{\sigma_0^2\sum_{g=1}^{g_n}\theta_g+\tau_\gamma^{(B),2}\mu_0}{g_n\sigma_0^2+\tau_\gamma^{(B),2}},\frac{\tau_\gamma^{(B),2}\sigma_0^2}{g_n\sigma_0^2+\tau_\gamma^{(B),2}}\right) \text{(2.3);}$$

$$\mu_\gamma^{(A)}|\gamma_1..\gamma_{g_n},\tau_\gamma^{(A),2},\mu_0,\sigma_0^2 \sim N\left(\frac{\sigma_0^2\sum_{g=1}^{g_n}\gamma_g+\tau_\gamma^{(A),2}\mu_0}{g_n\sigma_0^2+\tau_\gamma^{(A),2}},\frac{\tau_\gamma^{(A),2}\sigma_0^2}{g_n\sigma_0^2+\tau_\gamma^{(A),2}}\right) \text{(2.4).}$$

Cluster Hierarchical Model. The cluster hierarchical model is a non-parametric Bayesian method that uses a Dirichlet process with scale parameter, α, and base distribution, G_0. Specifically, a random distribution, G, is drawn from the base distribution, G_0. The scale parameter α determines the discreteness of the random distribution G, and it varies from a single discrete point mass to the base distribution G_0 as α goes from zero to infinity. The random distribution G is considered a combination of clusters, and the data from one subgroup are drawn from some certain cluster. In the DACTPerM study, for subject *i* in subgroup g from cluster w_c, the subject's response is given by

$$Y_{ig}|\, w_c \sim F(w_c)$$

$$w_c \sim G$$

$$G \sim DP\ (\alpha, G_0);$$

where $G_0 = N\ (\mu_0, \sigma_0^2)$, and μ_0 and σ_0^2 are identical to those from the hierarchical model presented previously. In addition, $F(w_c) = N = \left(\left(\mu_\gamma^{(A)} + \mu_\gamma^{(B)}\right)|_{w_c},\ \tau_\gamma^2|_{w_c}\right)$, and $\mu_\gamma^{(A)}$ and $\mu_\gamma^{(B)}$ have the same interpretation as those from the hierarchical model. Here, τ_γ^2 is shared across arms A and B. Detailed specifications regarding the three models and derivations can be found in the appendix 2.1.

2.2.3 Study Design Considerations

The study design is assessed by the properties and performance of candidate models under varying assumptions and conditions prior to study execution. However, when simulating a clinical trial, apart from the analysis model and its parameters, a variety of functional factors must be considered to obtain reliable results. Those factors include, but are not limited to, the number of interim analyses, visit information, treatment allocation ratios, and accrual and drop-out rates. We define all the functional input as functional parameters, and those directly related to the response models, longitudinal modeling, and imputation as model parameters.

Design Input - Models for treatment. As discussed in Section 2.2.2, three candidate models are considered for the statistical analysis plan and protocol: a pairwise independent model, a hierarchical model, and a cluster hierarchical model. All priors are specified based on the PAIN-CONTRoLS study.

Design Input - Interim analysis and early evaluation criteria. Interim analysis is important for the execution of an adaptive clinical trial, as it provides the means by which the design uses accumulating data to adapt. In this simulation, scenarios that include and exclude interim analysis are considered to assess their impact on operating characteristics. If interim analysis is included, all related early evaluation criteria are specified simultaneously for all subgroups. Specifically, the early success definition is that the posterior probability of one arm better than the other one is greater than the criterion (i.e. threshold) since both arms are active. I.e., the early success definition is that $P(\theta_g > 0 \,\middle|\, Data) >$ criterion for all g, which indicates Arm B is successful; or, $P(\theta_g < 0 \,\middle|\, Data) >$ criterion for all g, which indicates Arm A is successful. This study will stop for early success when it meets the early success definition.

Design Input - Final evaluation criteria. The final success criteria, like the early success criteria, are a function of the posterior probability one treatment arm being better than the other. Moreover, the final evaluation threshold values differ since we would like to control the overall type I error equal to 5%. Specifically, the final success definition is that the posterior probability of one arm better than the other one is greater than the threshold for some subgroup. I.e., the final success definition is that $P(\theta_g > 0 \mid Data) >$ criterion for some g, which indicates Arm B is successful; or, $P(\theta_g < 0 \mid Data) >$ criterion for some g, which indicates Arm A is successful.

To sum up, if no interim analysis is involved in the study design, the final success definition is that $P\left(\theta_g > 0 \middle| Data\right) >$ criterion for some g; or, $P\left(\theta_g < 0 \middle| Data\right) >$ criterion for some g. The type I error is controlled via the formula (2.5) below:

$$\Pr\left[P\left(\theta_g > 0 \middle| Data\right) > \text{criterion for some } g \text{ at final analysis} \middle| H_0\right] +$$

$$\Pr\ [P(\theta_g < 0 \mid Data) > \text{criterion for some } g \text{ at final analysis} | H_0]\ (2.5),$$

where H_0 is correspondent to *no effect* scenario (introduced in Section 2.2.3 - Simulation Description), and it means there is no treatment differences between Arm A and B for all subgroups. Criterion is adjusted to meet the type I error equal to 0.05 for each study design. The power is obtained via the formula (2.6) below:

$$Pr\left[P\left(\theta_g > 0 \middle| Data\right) > \text{criterion for some } g \text{ at final analysis} \middle| H_1\right] +$$

$$Pr[P(\theta_g < 0 \mid Data) > \text{criterion for some } g \text{ at final analysis} |\ H_1]\ (2.6),$$

where H_1 is correspondent to alternative scenarios (introduced in Section 2.2.3 - Simulation Description), and it means there is treatment differences between Arm A and B for some/all subgroups. Given one study design, the thresholds for alternative scenarios are identical to those from *no effect* scenario.

If interim analysis is involved in the study design, the early success definition is $P(\theta_g > 0 \,|\, Data) >$ criterion for all g; or, $P(\theta_g < 0 \,|\, Data) >$ criterion for all g. The final success definition is $P(\theta_g > 0 \,|\, Data) >$ criterion for some g; or, $P(\theta_g < 0 \,|\, Data) >$ criterion for some g. The type I error is controlled via the formula (2.7) below:

$$\Pr\,[P(\theta_g > 0 \,|\, Data) > \text{criterion for all } g \text{ at interim analysis}|H_0] +$$

$$\Pr\,[P(\theta_g < 0 \,|\, Data) > \text{criterion for all } g \text{ at interim analysis}|H_0] +$$

$$\Pr\left[P\left(\theta_g > 0 \middle| Data\right) > \text{criterion for some } g \text{ at final analysis}\middle|H_0\right] +$$

$$\Pr\,[P(\theta_g < 0 \,|\, Data) > \text{criterion for some } g \text{ at final analysis}|H_0] \text{ (2.7)},$$

The power is obtained via the formula (2.8) below:

$$\Pr\,[P(\theta_g > 0 \,|\, Data) > \text{criterion for all } g \text{ at interim analysis}|H_1] +$$

$$\Pr\,[P(\theta_g < 0 \,|\, Data) > \text{criterion for all } g \text{ at interim analysis}|H_1] +$$

$$\Pr\left[P\left(\theta_g > 0 \middle| Data\right) > \text{criterion for some } g \text{ at final analysis}\middle|H_1\right] +$$

$$\Pr\,[P(\theta_g < 0 \,|\, Data) > \text{criterion for some } g \text{ at final analysis}|H_1] \text{ (2.8).}$$

The meanings of H_0 and H_1 are identical to those introduced under the study designs without interim analysis involved.

Given a specific study design involved in interim analysis, the thresholds of interim and final analyses are different, and they are twisted based on the proportions of type I error spending on interim and final analyses. Boolean logic "and" for each subgroup criterion is applied at the interim analysis, and "or" is applied at the final analysis. Moreover, we would like to control type I error less than 0.005 spending on interim analysis. These strategies will result in a longer study and provide more information for the researcher to draw the conclusion. The specific criteria value for interim and final analyses are provided in section 2.2.3 - Simulation Description (Table 2-5). Still, one the thresholds of interim and final analyses are identified under the *no effect* scenario, they will be identically applied to the alternative scenarios.

Design Input - Rates of accrual and drop out. The accrual rate is an essential characteristic of a clinical trial since it determines trial duration. In adaptive designs, the accrual rate is even more important because the length of time between subject accrual and ascertainment of response determines the role of longitudinal data modeling in optimizing outputs. The accrual rate, together with drop-out rates, determine how many subjects are retained in the study. These rates for the simulation are based on the PAIN-CONTRoLS study.

Virtual endpoints. The null scenario (*no effect*) is used to calibrate the study design to a Type I error rate of approximately 5%. This is done via an iterative process that updates early and final evaluation criteria until the Type I error rate approaches but does not exceed 5%. Several alternative hypothesis scenarios that use the same input parameters but have varying response values are investigated.

Integrated two component prediction (ITP) is used for virtual endpoint simulation when longitudinal modeling is incorporated into the design. ITP allows endpoints to follow an exponential model over time with a subject-specific random effect to scale the visit values to the visit-specific specification of subgroup responses. Additionally, ITP does not affect the distribution of the final endpoint (Facts 2018). Three elements—the mean final endpoint, the of inter-subject ‘noise,’ and the noise at the current visit—along with the exponential function’s visit time and shape parameters determine the longitudinal data simulation at each visit (Facts 2018). Complete ITP specifications are in Appendix 2.2.

Design Input - Imputation via longitudinal modeling. Longitudinal modeling is also applied for data imputation, and it is useful whether the trial is fixed or adaptive. Longitudinal modeling can be used in a fixed trial to impute endpoint values for patients that have dropped out of the study. Moreover, in an adaptive design, it can be used for imputing endpoints that have not yet been observed for an interim analysis, allowing the study to maximize the use of data to more efficiently adapt.

Simple linear regression (SLR) for Bayesian multiple imputation is used to model the relationship between responses observed at each pre-final visit and the unobserved (future) final visit. Specifically, for the future final response of subject *i* in subgroup *g* and treatment arm *j*,

$$Y_{i,g}^{(j)} \mid y_{it,g}^{(j)} \sim N\left(\alpha_t + \beta_t y_{it,g}^{(j)}, \lambda_t^2\right),$$

$$\alpha_t \sim N\left(\alpha_\mu, \alpha_\sigma^2\right), \beta_t \sim N\left(\beta_\mu, \beta_\sigma^2\right), \lambda_t^2 \sim IG\left(\frac{\lambda_n}{2}, \frac{\lambda_\mu^2 \lambda_n}{2}\right),$$

where α_t and β_t are the intercept and slope at visit time t, and $y_{it,g}^{(j)}$ is the observed response for the subject *i* at visit time *t*. The model priors are specified identically across all visits (see Section 2.2.3 - Simulation Description).

The subjects' pending endpoints at interim analysis or missing ones at final analysis are imputed by the predicted distribution generated from multiple imputation via the SLR model. The imputed value from the predicted distribution captures both the uncertainty in the estimate of the parameters of the SLR model and the uncertainty of the prediction of the endpoint given particular parameter values (Facts 2018).

Design Input - Allocation. Unequal allocation may be applied in some studies where sample size or randomization ratio adjustments are performed. Here, a 1:1 randomization ratio of subjects to the two treatment arms is fixed within each subgroup.

Design Output - Subgroup power. Power can also be calculated in Bayesian studies via simulation. Subgroup power is defined as the probability that a subgroup meets the success criteria under the assumption that the subgroup responses from the two treatment arms are different.

Design Output - Overall power (study success). Simulations track the proportion of studies that show early success and final success based on the evaluation criteria (See Section 2.2.3- Simulation Description). Overall power is calculated via the summation of both proportions, i.e., early and late success proportions. Both subgroup and overall power provide important model performance information and thus make the model assessment comprehensive.

Design Output - Sample size. Sample size is another key characteristic since it directly relates to the cost of running a trial. Thus, a study design that results in a lower sample size but similar power to a competing design is desirable. Compared to a fixed trial, an adaptive design can result in smaller sample sizes due to early stopping criteria.

Design Output - Trial duration. The trial duration is highly dependent upon accrual and sample size goals. It serves as a complimentary operating characteristic that the sponsor may consider when calculating trial cost prior to study execution.

Simulation Description. The simulation is executed for each study design in terms of an analysis model, interim analysis, and longitudinal modeling. Three analysis models are considered: pairwise independent, hierarchical, and cluster hierarchical. For each model, interim analysis and longitudinal modeling are either included or not. As Table 2-1 below indicates, the simulation is composed of three factors; there are twelve different study designs for the simulations.

Table 2-1 Levels of the three factors for study design

Factor 1: Model	Factor 2: interim analysis involvement	Factor 3: Longitudinal modeling involvement
Pairwise independent Hierarchical Cluster hierarchical	Yes No	Yes No

To assess the designs comprehensively, we propose several alternative hypothesis scenarios that mimic the most frequent responses that can occur in real cases, and each scenario assumes a different response profile under two treatment arms. The specific scenarios include *moderate and homogeneous effect, small and homogeneous effect, spread, opposite,* and *one nugget*. Moreover, Arm B is assumed to have the effect for all the scenarios for the convenience of related formula and distribution specification. Supposing Arm A has the effect, the design outputs will be symmetric, as the related ones in which Arm B has the effect. Tables 2-2 and 2-3 present the specific virtual scenarios for four or eight patient subgroups. We assume the virtual response, a continuous measure of pain reduction, is normally distributed, in which higher values indicate better response to treatment. A common standard deviation (0.3) is specified for each subgroup of the two arms across all the scenarios, and this value is derived from the motivated example.

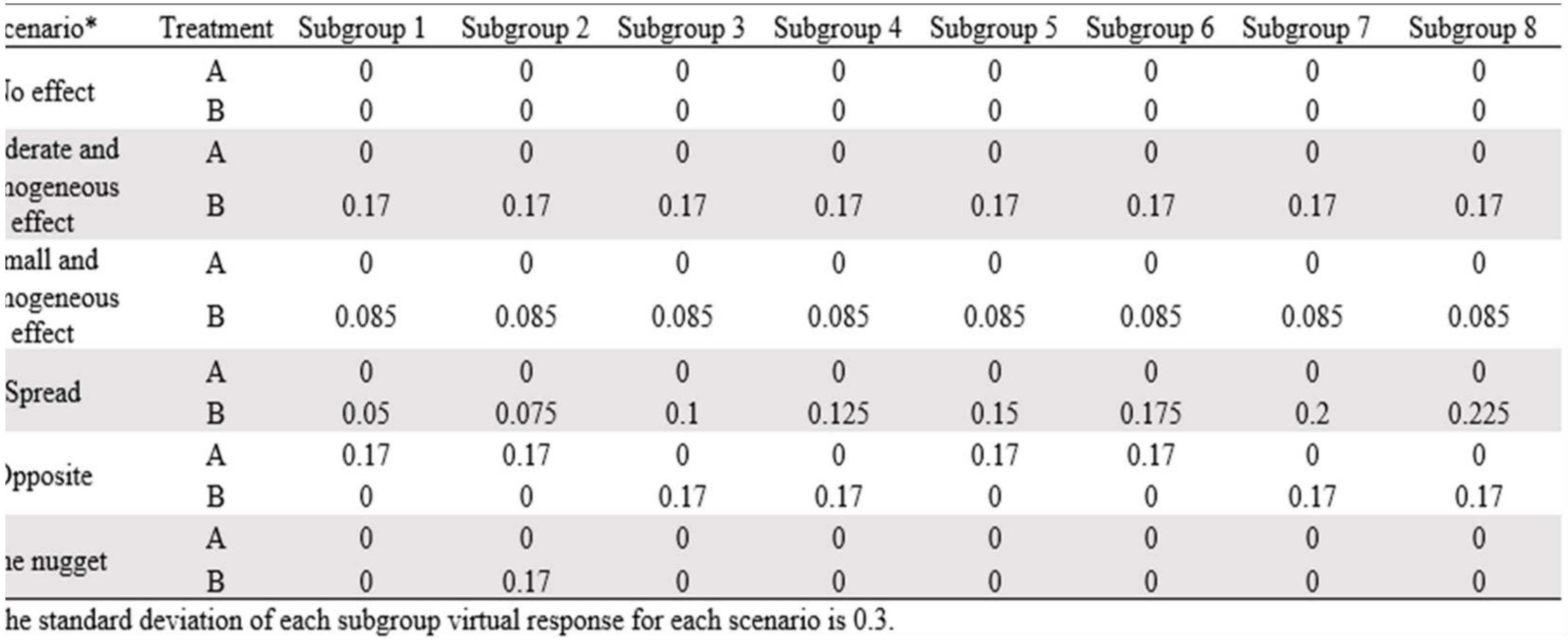
Table 2-2 Four subgroup virtual response under six virtual treatment effect scenarios

Scenario*	Treatment	Subgroup 1	Subgroup 2	Subgroup 3	Subgroup 4
No effect	A	0	0	0	0
	B	0	0	0	0
Moderate and homogeneous effect	A	0	0	0	0
	B	0.17	0.17	0.17	0.17
Small and homogeneous effect	A	0	0	0	0
	B	0.085	0.085	0.085	0.085
Spread	A	0	0	0	0
	B	0.05	0.1	0.2	0.25
Opposite	A	0.17	0.17	0	0
	B	0	0	0.17	0.17
One nugget	A	0	0	0	0
	B	0	0.17	0	0

*: The standard deviation of each subgroup virtual response for each scenario is 0.3.

Weakly informative priors that reflect the PAIN-CONTRoLS study are applied. In the cluster hierarchical model, a larger DP scale parameter will result in the random distribution being close to the base distribution, whereas a smaller DP scale parameter will result in a more discrete (point mass) random distribution. To differentiate it from the hierarchical model, the DP scale parameter is set to 2. All subgroups are assumed to have identical priors for the coefficient and intercept of SLR within each treatment arm. Though the prior mean values of the coefficient and intercept were obtained from PAIN-CONTRoLS, the prior standard deviation values of the coefficient and intercept were increased to 0.4 and 0.1 from 0.04 and 0.01, respectively, to reduce the impact of the motivating study data on simulation results. Table 2- 4 presents the specific values for all priors involved in the simulation.

Table 2-3 Eight subgroup virtual responses under virtual treatment effect scenarios



cenario*	Treatment	Subgroup 1	Subgroup 2	Subgroup 3	Subgroup 4	Subgroup 5	Subgroup 6	Subgroup 7	Subgroup 8
Io effect	A	0	0	0	0	0	0	0	0
	B	0	0	0	0	0	0	0	0
derate and ıogeneous effect	A	0	0	0	0	0	0	0	0
	B	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17
mall and ıogeneous effect	A	0	0	0	0	0	0	0	0
	B	0.085	0.085	0.085	0.085	0.085	0.085	0.085	0.085
Spread	A	0	0	0	0	0	0	0	0
	B	0.05	0.075	0.1	0.125	0.15	0.175	0.2	0.225
)pposite	A	0.17	0.17	0	0	0.17	0.17	0	0
	B	0	0	0.17	0.17	0	0	0.17	0.17
ıe nugget	A	0	0	0	0	0	0	0	0
	B	0	0.17	0	0	0	0	0	0

he standard deviation of each subgroup virtual response for each scenario is 0.3.

Table 2-4 Prior specification for analysis models, longitudinal data imputation and longitudinal data simulation

Model	Parameter	Value
Pairwise independent*	prior mean of treatment arm A response for each subgroup ($\mu_g^{(A)}$)	0
	prior standard deviation of treatment arm A response for each subgroup ($\tau_g^{(A)}$)	0.3
	prior mean of treatment arm B response for each subgroup ($\mu_g^{(B)}$)	0
	prior standard deviation of treatment arm B response for each subgroup ($\tau_g^{(B)}$)	0.3
Hierarchical	hyperprior mean of all response prior mean (μ_0)	0
	hyperprior standard deviation of all response prior mean (σ_0)	0.1
	central parameter of inverse gamma distribution as which all response prior variance is distributed (τ_μ)	0.1
	weight parameter of inverse gamma distribution as which all response prior variance is distributed (τ_n)	2
Cluster hierarchical[&]	DP scale (α)	2
SLR [@] - imputation for longitudinal data	prior mean of coefficient (β_μ)	0.8
	prior standard deviation of coefficient (β_σ)	0.4
	prior mean of intercept (α_μ)	0
	prior standard deviation of intercept (α_σ)	0.1
	central parameter of inverse gamma distribution as which imputed response variance is distributed (λ_μ)	0.18
	weight parameter of inverse gamma distribution as which imputed response variance is distributed (λ_n)	200/400
ITP [$] - simulation for longitudinal data	fraction of the final treatment arm A response variance for each subgroup ($\omega_g^{(A)}$)	0.8
	fraction of the final treatment arm B response variance for each subgroup ($\omega_g^{(B)}$)	0.8
	shape parameter of the exponential component for treatment arm A of each subgroup ($k_g^{(A)}$)	-10
	shape parameter of the exponential component for treatment arm B of each subgroup ($k_g^{(B)}$)	-10
Public parameter	central parameter of inverse gamma distribution as which all response variance is distributed (σ_μ)	1
	weight parameter of inverse gamma distribution as which all response variance is distributed (σ_n)	1

*: We assume all subgroups from both treatment arms have identical prior means and standard derivation.

[&]: Cluster hierarchical model maintains all hyperpriors from the hierarchical model and has additional DP scale parameter.

[@]: All subgroups from both treatment arms are assumed to have same priors for the coefficient and intercept of SLR except for λ_n, and λ_n is specified as 200 for four subgroups and 400 for eight subgroups.

[$]: Both fraction and shape parameter are specified identically within each subgroup from both treatment arms.

Early and final success criteria are designed to identify subgroup effects and study success. Boolean logic “and” is applied for the subgroup criterion at the interim analysis, and “or” is applied at study completion (specific criteria stated in Section 2.2.3 - Design Inputs). This results in a longer study and more conservative analysis. The concrete values for the early and final evaluation criteria are shown in Table 2- 5. Operating characteristics such as power, sample size, and study duration under other effective virtual treatment scenarios with identical evaluation criteria from related *no effect* scenario are obtained accordingly via the simulations.

Table 2-6 presents the functional parameter values for the simulation, which are derived from PAIN-CONTRoLS. Subgroup sample sizes are set to 100, and the final sample size is determined via simulation with the consideration of Type I error and power. Study duration is specified as 12 weeks, and interim analysis will be executed once half the total number of subjects are enrolled. The study assumes three visits, with a 4-week lapse between consecutive visits. Each study design is simulated 10000 times.

2.3. Results

Subgroup power. For the designs with four subgroups without interim analysis or longitudinal modeling (Figure 2-1), the hierarchical model performs best in all the scenarios. The cluster hierarchical model performs similarly with mildly less power compared to the hierarchical model in the scenarios of *opposite* and *one nugget*. Similar findings are identified from Figure 2-2 which presents the designs with four subgroups without interim analysis and with longitudinal modeling. Each of the three models is with mildly higher power compared to that in each scenario under the design without interim analysis and without longitudinal modeling.

Subgroup Number	Study design factor			Early evaluation criteria*		Final evaluation criteria&	
	Model	Interim analysis involvement	longitudinal modelling involvement	Posterior probability for each subgroup	Boolean logic	Posterior probability for each subgroup	Boolean logic
4	Pairwise independent	No	No	----	----	0.9916	OR
4	Hierarchical	No	No	----	----	0.9805	OR
4	Cluster hierarchical	No	No	----	----	0.9822	OR
4	Pairwise independent	No	Yes	----	----	0.991	OR
4	Hierarchical	No	Yes	----	----	0.9777	OR
4	Cluster hierarchical	No	Yes	----	----	0.98	OR
4	Pairwise independent	Yes	No	0.9	AND	0.9932	OR
4	Hierarchical	Yes	No	0.9	AND	0.9818	OR
4	Cluster hierarchical	Yes	No	0.9	AND	0.9822	OR
4	Pairwise independent	Yes	Yes	0.9	AND	0.992	OR
4	Hierarchical	Yes	Yes	0.9	AND	0.979	OR
4	Cluster hierarchical	Yes	Yes	0.9	AND	0.9818	OR
8	Pairwise independent	No	No	----	----	0.9963	OR
8	Hierarchical	No	No	----	----	0.983	OR
8	Cluster hierarchical	No	No	----	----	0.9869	OR
8	Pairwise independent	No	Yes	----	----	0.9955	OR
8	Hierarchical	No	Yes	----	----	0.9818	OR
8	Cluster hierarchical	No	Yes	----	----	0.9865	OR
8	Pairwise independent	Yes	No	0.9	AND	0.9963	OR
8	Hierarchical	Yes	No	0.9	AND	0.9818	OR
8	Cluster hierarchical	Yes	No	0.9	AND	0.9851	OR
8	Pairwise independent	Yes	Yes	0.9	AND	0.9955	OR
8	Hierarchical	Yes	Yes	0.9	AND	0.981	OR
8	Cluster hierarchical	Yes	Yes	0.9	AND	0.985	OR

*: The study will be identified as early success if all the subgroups meet the criteria. i.e. the posterior probability of (Arm B > Arm A, or Arm A > Arm B) for each subgroup meets the related criteria list the table above.

&: The study will be identified as final success if any of the subgroups meet the criterion. i.e. the posterior probability of (Arm B > Arm A, or Arm A > Arm B) for any of the subgroups meet the related criterion list the table above.

Table 2-6. Values of input functional parameters for study design

Functional factor	Value
Sample size per subgroup	100
Study duration	12 Weeks
Interim analysis execution time*	200 and 400 subjects enrolled for 4 and 8 subgroups
Visit times and duration between two consecutive visits*	3 visits; 4 weeks between visits
Allocation ratio of two arms within each subgroup	1:1
Accrual rate	4 /week
Drop-out rate	10 %

*: Interim analysis execution time, specific visit times and duration between two consecutive visits are only involved when the study designs are with interim analysis and/or longitudinal data modeling.

From Figure 2-3, which presents the subgroup power at the designs of three models with interim analysis and without longitudinal modeling, it can be observed that the three models' performance order is identical to that from Figure 2-1. Each model is with a little less power compared to that in each scenario in Figure 2-1.

In the designs of three models with interim analysis and with longitudinal modeling, the hierarchical model still performs best in all scenarios, and the performances of cluster hierarchical and pairwise independent model come to the second and third place. The power differences from hierarchical and cluster hierarchical models in one nugget scenario is larger than those from Figure 2-1 to 2-3.

When subgroup number increases to 8, the subgroup power of the hierarchical model is still the highest within each subgroup of each scenario under the batch designs with identical involvement of interim analysis and longitudinal modeling. The power of the cluster hierarchical model for all subgroups within each subgroup under each design batch is lower than that from the hierarchical model but higher than that from a pairwise independent model.

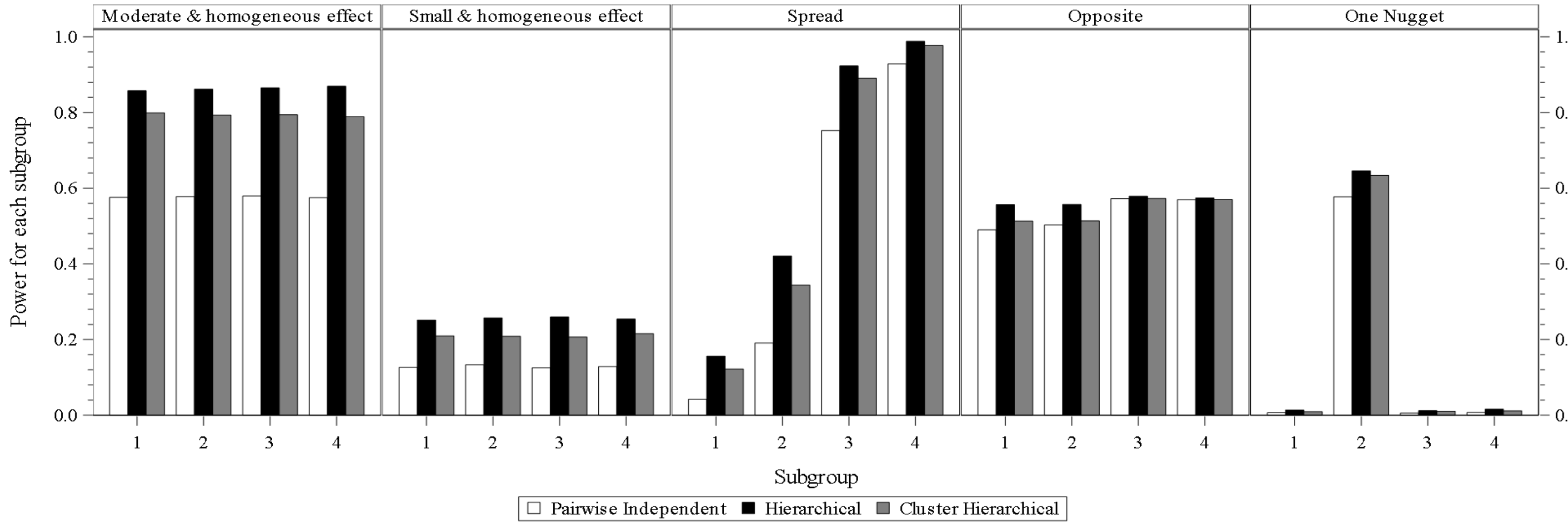


Figure 2-1 four subgroups power for design choice – model without interim analysis and without longitudinal modelling

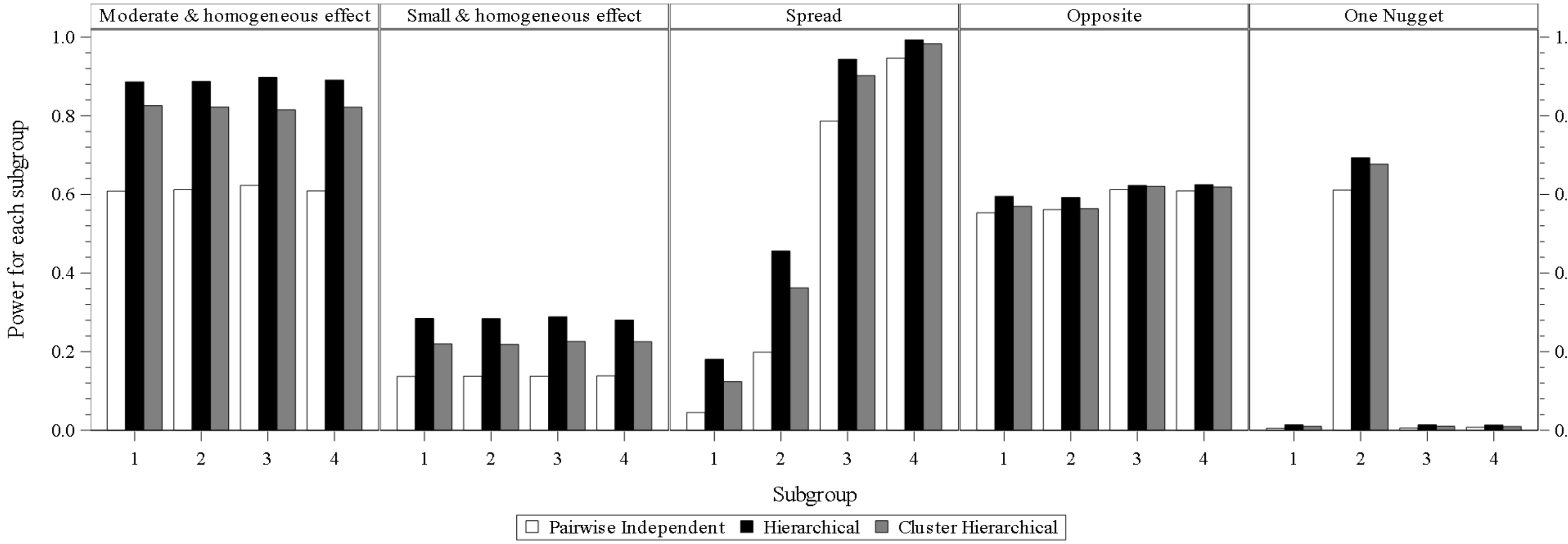


Figure 2-2 four subgroups power for design choice – model without interim analysis and with longitudinal modelling

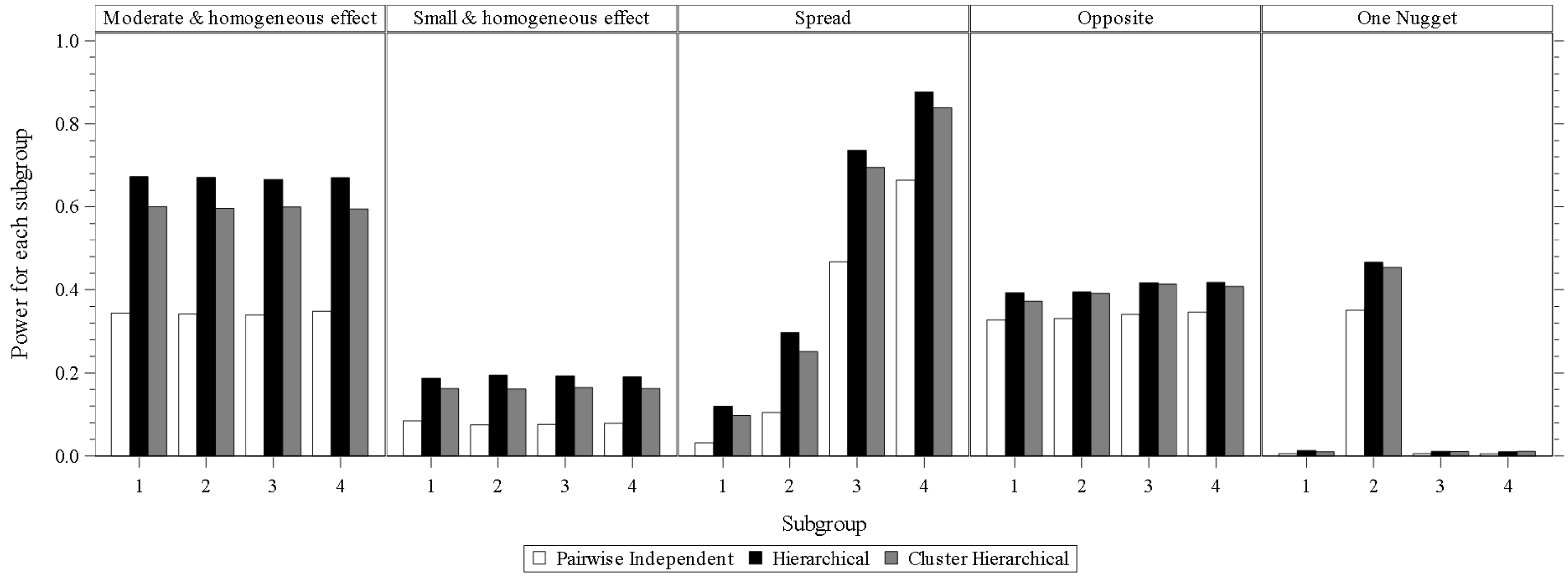


Figure 2-1 four subgroups power for design choice – model with interim analysis and without longitudinal modelling imputation

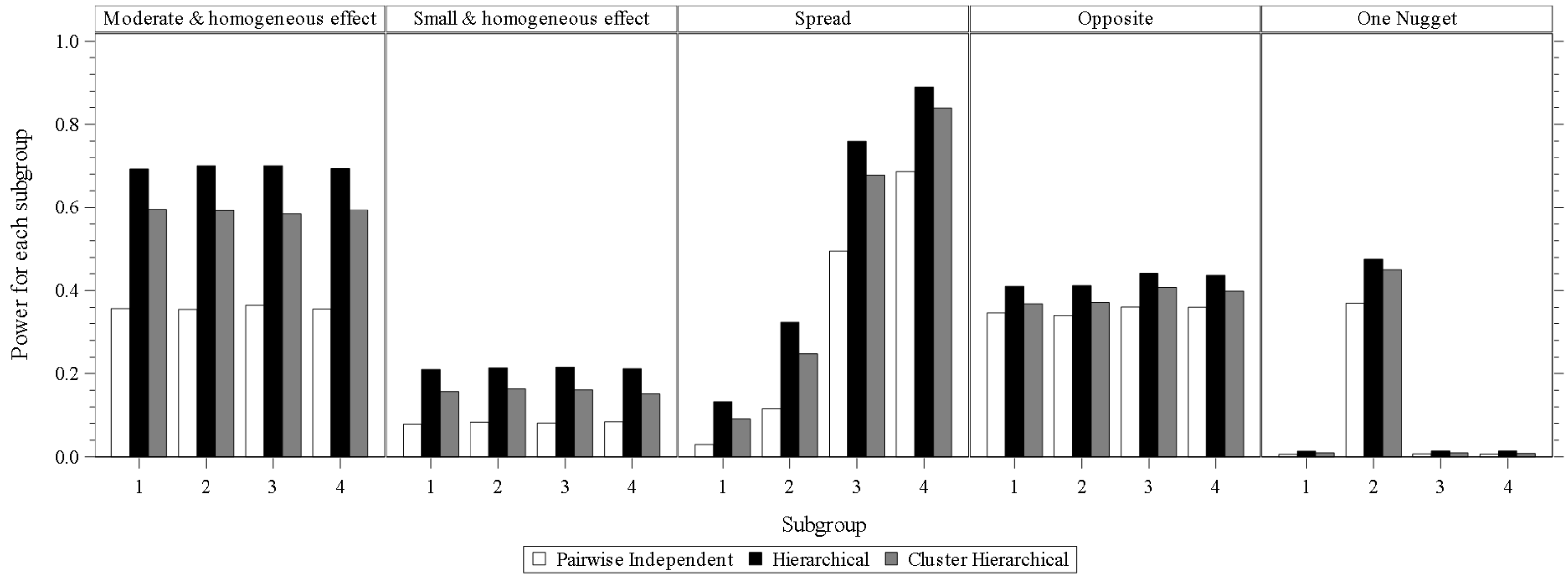


Figure 2-2 four subgroups power for design choice – model with interim analysis and with longitudinal modelling imputation

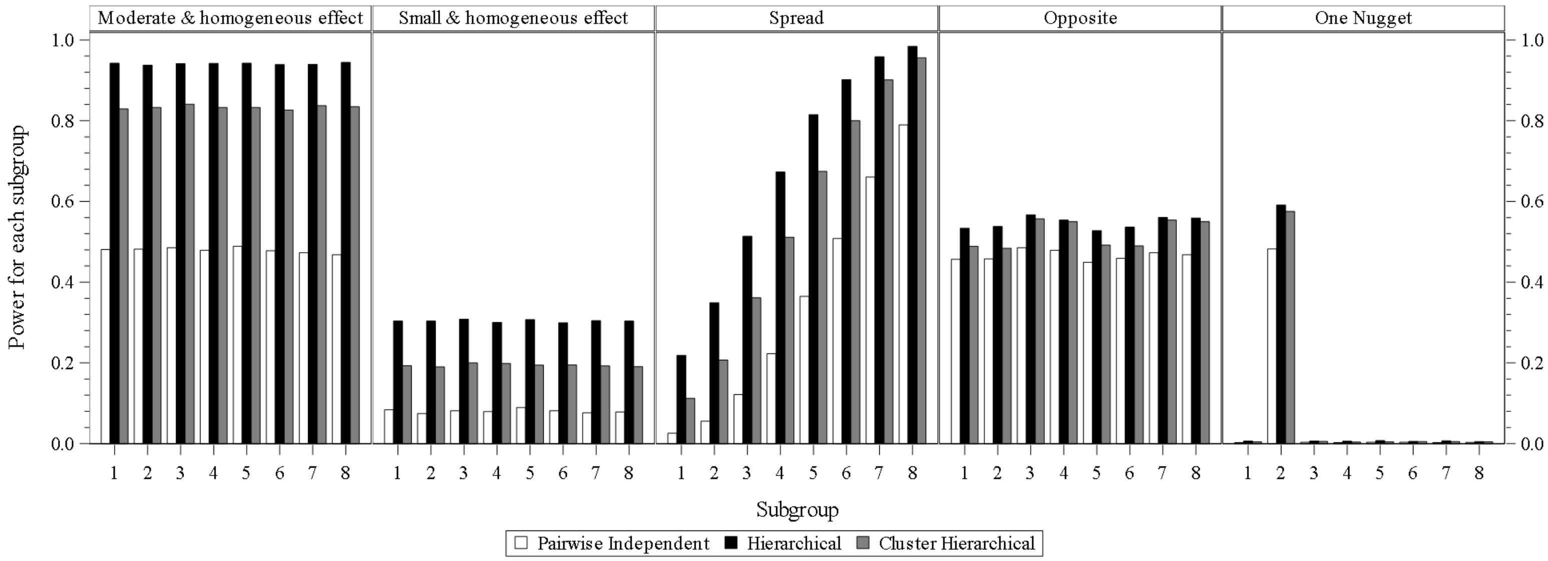


Figure 2-5 eight subgroups power for design choice – model without interim analysis and without longitudinal modelling

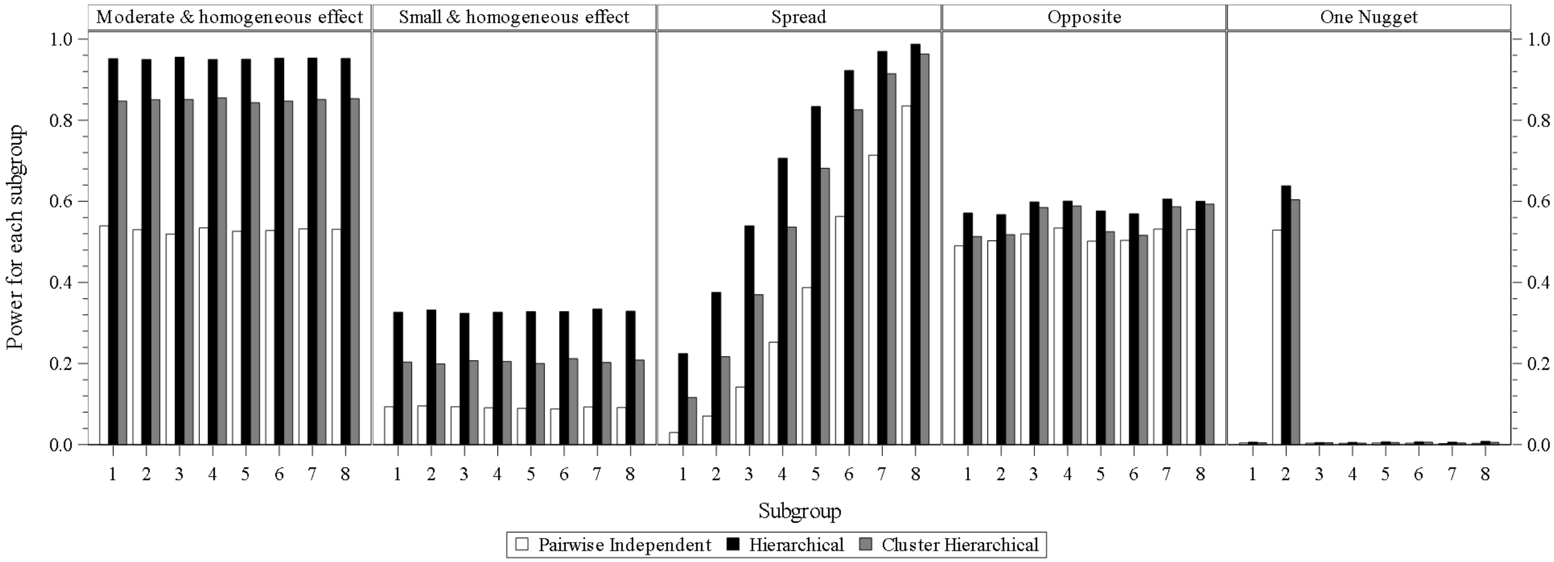


Figure 2-6 eight subgroups power for design choice – model without interim analysis and with longitudinal modelling

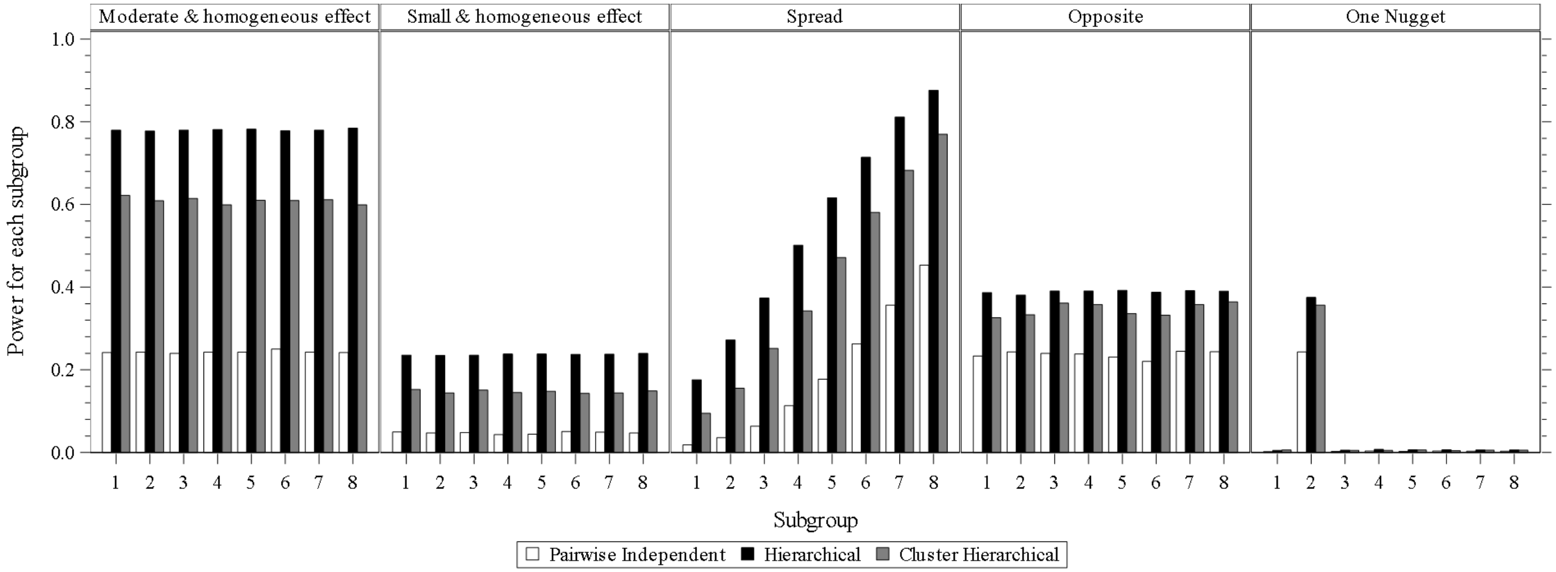


Figure 2-3 eight subgroups power for design choice – model with interim analysis and without longitudinal modelling

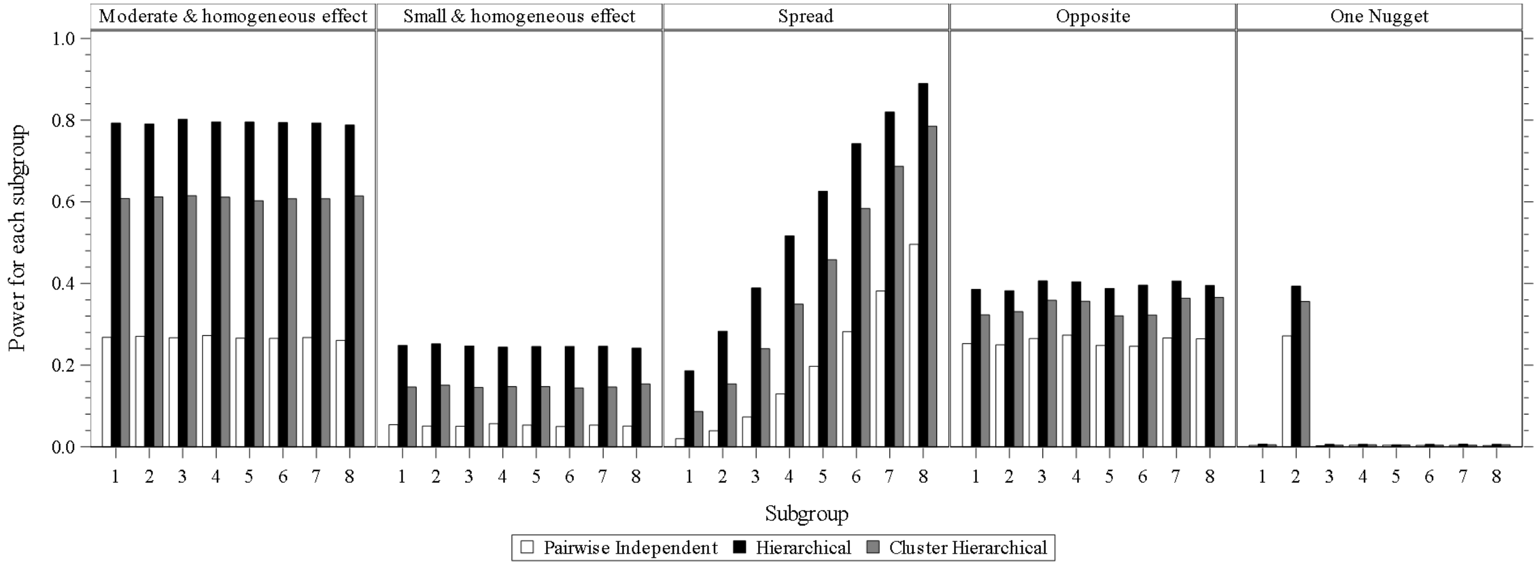


Figure 2-4 eight subgroups power for design choice – model with interim analysis and with longitudinal modelling imputation

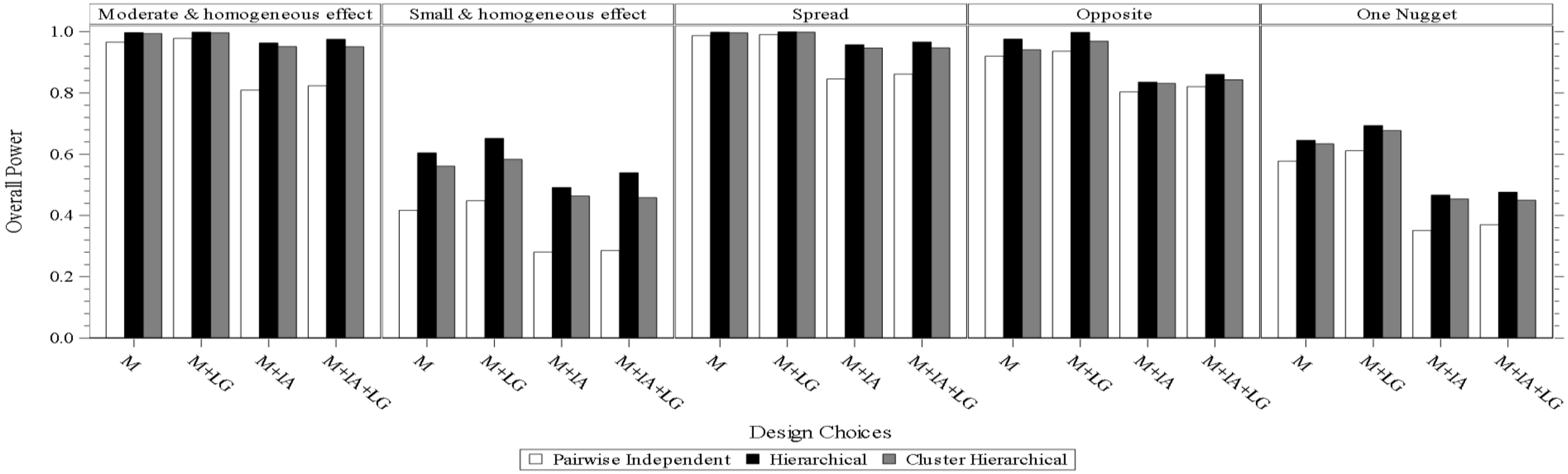


Figure 2-5 overall power for design choice under four subgroups. M = model without interim analysis and without longitudinal modelling imputation, M + LG = model without interim analysis and with longitudinal modelling imputation, M + IA = model with interim analysis and without longitudinal modelling imputation, M + IA + LG = model with interim analysis and with longitudinal modelling imputation

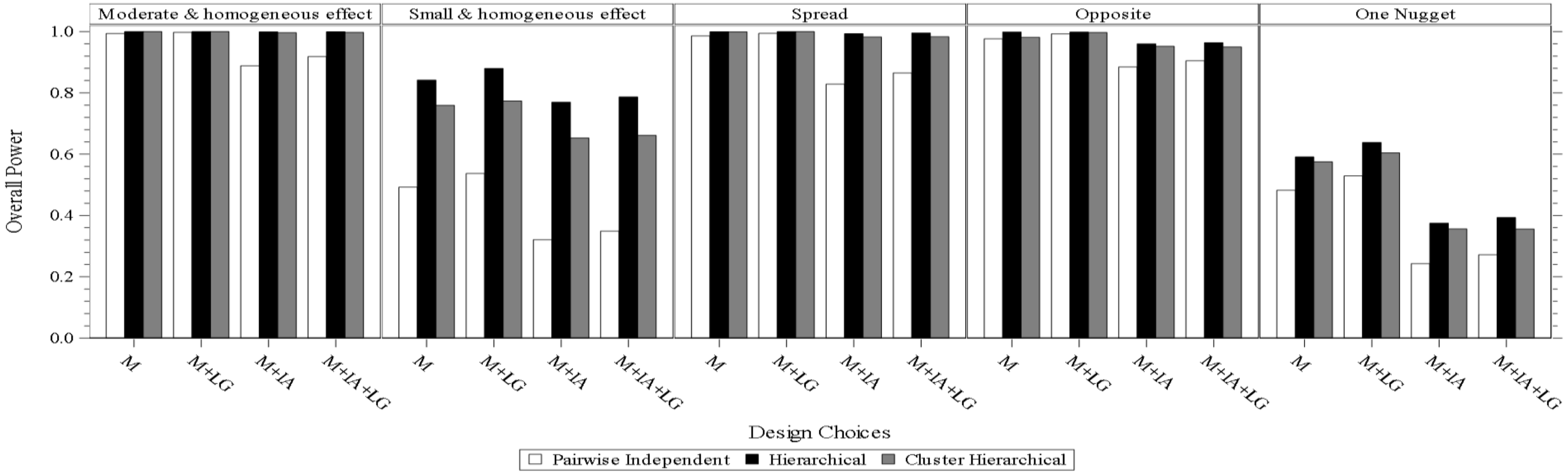


Figure 2-6 overall power for design choice under eight subgroups. M = model without interim analysis and without longitudinal modelling imputation, M + LG = model without interim analysis and with longitudinal modelling imputation, M + IA = model with interim analysis and without longitudinal modelling imputation, M + IA + LG = model with interim analysis and with longitudinal modelling imputation

Overall power (Study success). It is easy to directly obtain the study power for each of the scenarios since it is equal to the final success proportion output from the simulation except for the opposite case. In the opposite scenario for four and eight subgroups, the overall power is calculated by the summation of proportion of simulated studies with any of subgroups in which the posterior probability of response from one arm higher than the response from other one satisfying the success criteria. The logic for this calculation is that there exists two treatment comparators and the study is successful if either arm within any subgroup meets the criteria. The overall power for the one nugget is consistent to the power from subgroup 2 in Figure 2-1 to 2-8 presenting the subgroup power of related designs under different scenarios for both four and eight subgroups.

In the designs of three models without interim analysis and longitudinal modeling, overall power is high and quite similar to the three models under the scenarios of the *moderate and homogeneous effect* and *spread*. Under the *opposite* scenario, the power of the hierarchical model is still high, and the power goes down slightly but is still high for the cluster hierarchical and pairwise independent models. The power of the hierarchical model under the all scenarios of the *small and homogeneous effect*, and *one nugget* is the highest. The power of the cluster hierarchical model under the same two scenarios decreases slightly, and the power of the pairwise independent model under the two scenarios is lower and with relatively larger differences compared to that from the hierarchical model. Similar findings are identified for the designs of three models without interim analysis and with longitudinal modeling. Each of the three models is with mildly higher power compared to that in each scenario under the design without interim analysis and without longitudinal modeling.

In the designs of three models with interim analysis and without longitudinal modeling, hierarchical and cluster hierarchical models perform similarly and have higher power than that for a pairwise independent model under each scenario.

In the designs of the three models with interim analysis and with longitudinal modeling, the hierarchical model has the highest power compared to the other two models in each scenario, and cluster hierarchical model performs closely to the hierarchical model with mildly decreased power. Performance of the pairwise independent model, same as that from the other design batch, is with the lowest power in each scenario. The same or quite similar comparison results are observed from eight subgroups.

Sample size. Figure 2-11 & 2-12 present the expected sample size of designs under different scenarios for both four and eight subgroups. For the design batches of three models without interim analysis and with/without longitudinal modeling, the sample size is fixed as 100 subjects per subgroup. For the designs of the three models with interim analysis and without longitudinal modeling under the *moderate and homogeneous effect* and *spread* scenarios, the expected sample size dropped by 156 and 126 for hierarchical model, and by 141 and 115 for cluster hierarchical model, and by 119 and 104 for pairwise independent model. For the designs of the three models with interim analysis and with longitudinal modeling under the *moderate and homogeneous effect* and *spread* scenarios, the expected sample size approximately dropped by 167 and 134 for hierarchical model, and by 154 and 124 for cluster hierarchical model, and by 134 and 113 for pairwise independent model. The same trend is also observed under the *small and homogeneous effect* scenario, but all three models have higher expected sample size compared to the relevant one from the *moderate and homogeneous effect* and *spread* scenarios. However, under the scenarios of *opposite* and *one nugget*, pairwise independent is the best, and the other two models have higher expected sample size and perform similarly. The average expected sample sizes are approximately 330 and 360 for the two scenarios under the designs of the two models with interim analysis and without longitudinal modeling. The average expected sample sizes are approximately 310 and 360 for the two scenarios under the designs of the two models with interim analysis and with longitudinal modeling. Similar trends and comparison results are observed for eight subgroups.

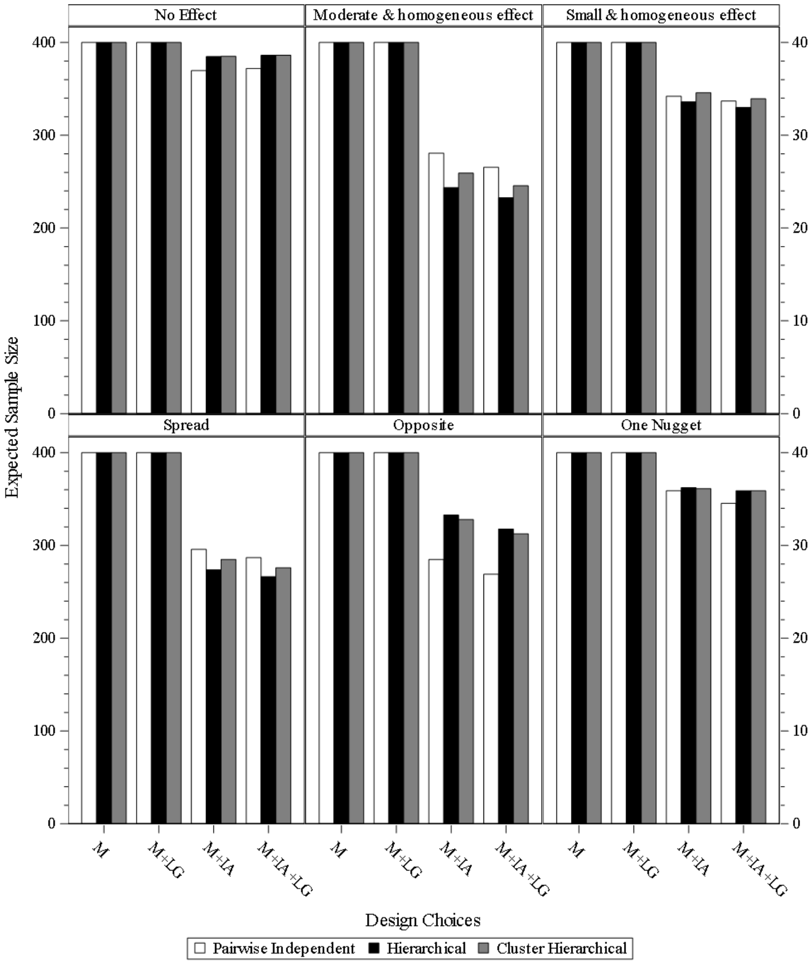


Figure 2-7 expected sample size for study design under four subgroups. M = model without interim analysis and without longitudinal modelling imputation, M + LG = model without interim analysis and with longitudinal modelling imputation, M + IA = model with interim analysis and without longitudinal modelling imputation, M + IA + LG = model with interim analysis and with longitudinal modelling imputation.

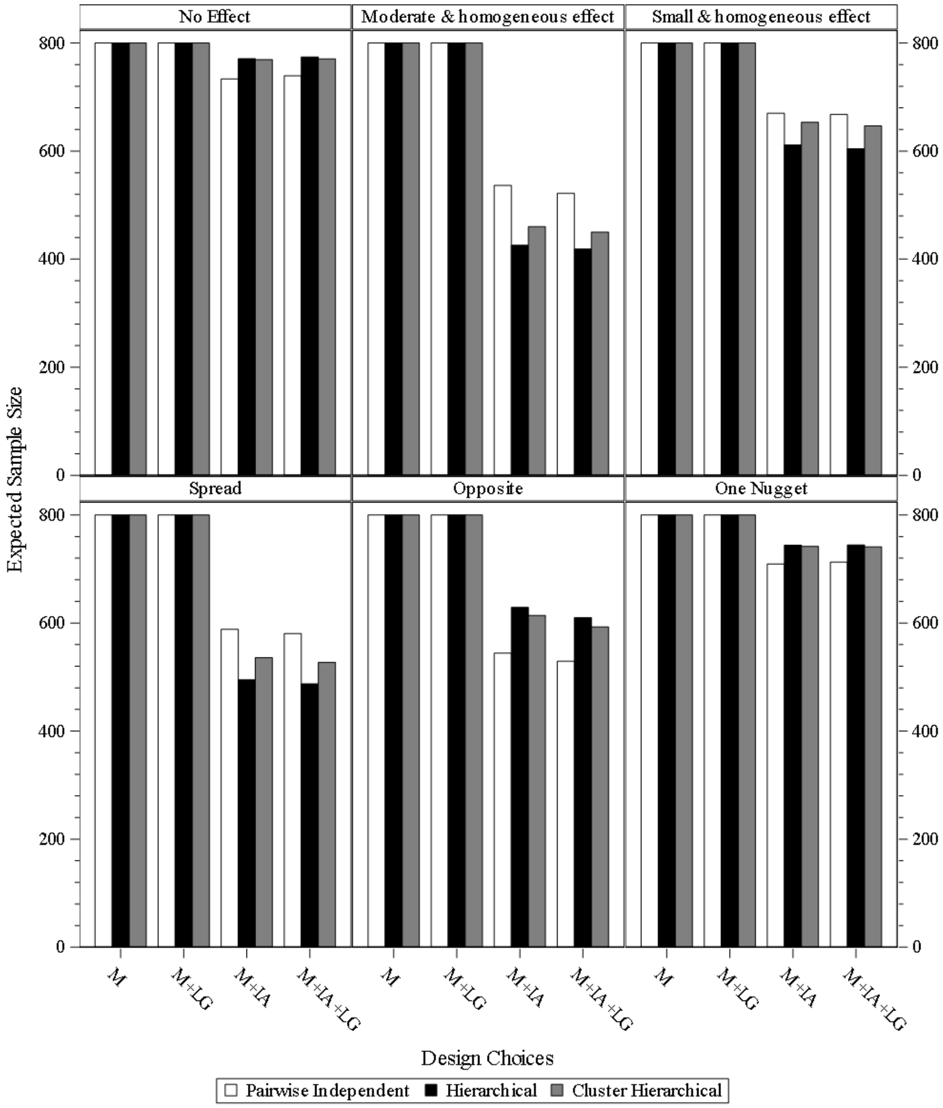


Figure 2-8 expected sample size for study design under eight subgroups. M = model without interim analysis and without longitudinal modelling imputation, M + LG = model without interim analysis and with longitudinal modelling imputation, M + IA = model with interim analysis and without longitudinal modelling imputation, M + IA + LG = model with interim analysis and with longitudinal modelling imputation.

Trial duration. Figure 2-13 & 2-14 presents the mean trial duration of the study designs under different scenarios for both four and eight subgroups. The same or similar findings of three models under different scenarios for both four and eight subgroups, as those from sample size observed since the trial duration is highly correlated to the sample size.

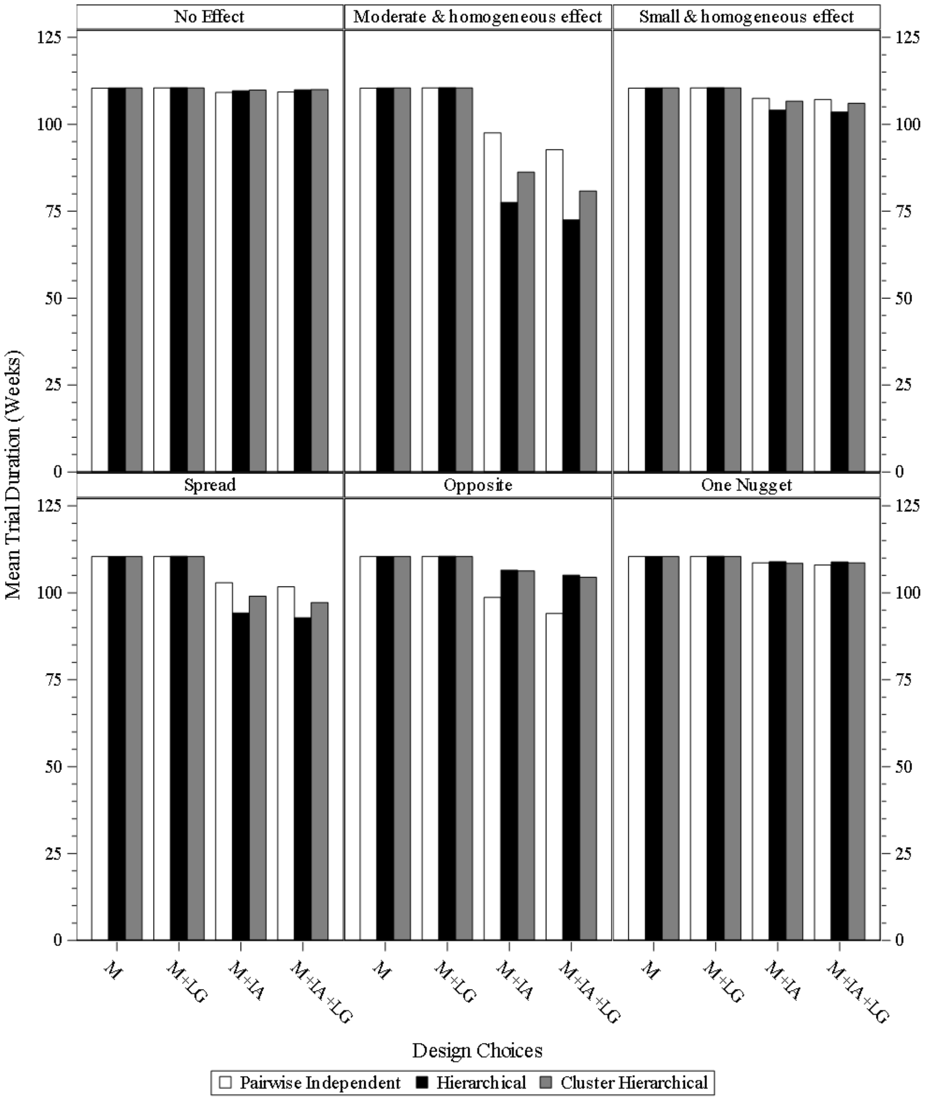


Figure 2-9 mean study duration for study desgin under four subgroups. M = model without interim analysis and without longitudinal modelling imputation, M + LG = model without interim analysis and with longitudinal modelling imputation, M + IA = model with interim analysis and without longitudinal modelling imputation, M + IA + LG = model with interim analysis and with longitudinal modelling imputation

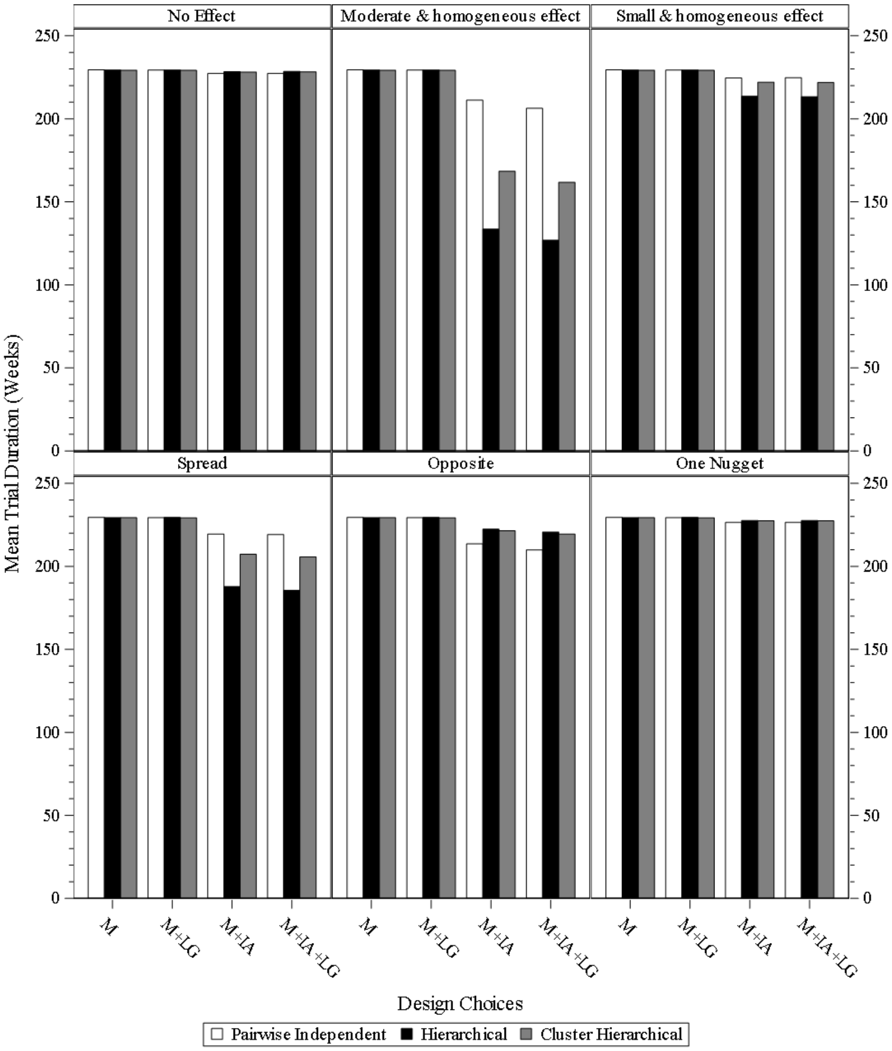


Figure 2-10 mean study duration for study desgin under eight subgroups. M = model without interim analysis and without longitudinal modelling imputation, M + LG = model without interim analysis and with longitudinal modelling imputation, M + IA = model with interim analysis and without longitudinal modelling imputation, M + IA + LG = model with interim analysis and with longitudinal modelling imputation.

Overall power comparison between hierarchical model and two independent sample t-test. We also explored the overall power (study success) comparison between the hierarchical model and an approach that ignores the different subgroup effects and uses a classical-frequentist method—t-test without the involvement of interim analysis and longitudinal data. Table 2 - 7 below presents the concrete values from the two approaches. The powers of the Bayesian hierarchical model are much higher for the *opposite* and *one nugget* scenarios. This is because the subgroup treatment effects for these two scenarios are a challenge to identify at the study level for frequentist approach.

Table 2-7 the overall power comparison between frequentist and our study

Scenario	Treatment	Four subgroups				Eight subgroups			
		Overall Effect	N Per Arm	Frequentist Power@	Our study Power *	Overall Effect	N Per Arm	Frequentist Power@	Our study Power *
No effect	A	0	200	0.05	0.05	0	400	0.05	0.05
	B	0	200			0	400		
Moderate and homogeneous effect	A	0	200	0.99	0.997	0	400	0.99	1
	B	0.17	200			0.17	400		
Small and homogeneous effect	A	0	200	0.81	0.6043	0	400	0.98	0.8417
	B	0.085	200			0.085	400		
Spread	A	0	200	0.99	0.9984	0	400	0.99	0.9997
	B	0.15	200			0.1375	400		
Opposite	A	0.085	200	0.05	0.9757	0.085	400	0.05	0.9987
	B	0.085	200			0.085	400		
One nugget	A	0	200	0.29	0.6456	0	400	0.17	0.591
	B	0.0425	200			0.0213	400		

@: Power is calculated via the t-test for two independent sample

*: All the powers for each scenario are from the design of hierarchical model without interim analysis and without longitudinal data under four or eight subgroups.

2.4. Discussion and Conclusion

This paper explores the performance of three Bayesian models—pairwise independent, hierarchical, and cluster hierarchical—under different virtual responses for subgroups, including versions with interim analysis and longitudinal modeling. For all scenarios under each design, the hierarchical model generally performs better than the other two. This is because the hierarchical model is able to analyze the data using a mixture model, flexibly borrowing information from all subgroups and shrinking the subgroup means towards the central one, according to how similar they appear. The final output is sensitive to prior distribution specification and related prior value setting, and thus the hyperprior setting is an essential factor in achieving the hierarchical model property, and different settings may affect the performance of the hierarchical model. The prior setting reflects the belief about the parameter before data is available. Informative prior, usually represented by location and scale parameters, is derived from researchers' clear understanding or the availability of highly relevant data. Otherwise, non-informative prior or weakly informative prior should be specified. The conjugate property of prior is another consideration when setting the prior from computing perspective. In our research, we incorporated the information from the example study and set the hyperprior following a normal distribution with mean and standard deviation equal to 0 and 0.1, which is weakly informative prior and conservative and leads to trials designs that mostly rely on data collected from the trial and not the prior. It is pronounced in the simulation results of the *spread* scenario of three models with interim analysis and longitudinal modeling involvement, the hierarchical model performs excellently in terms of reducing sample size by 40 percent and maintaining same power, compared to the simulation results of three models without interim analysis and longitudinal modeling involvement. For the scenarios of the *moderate and* *homogeneous effect* and *small and homogeneous effect*, the hierarchical model still provides an acceptable power and a decreased sample size, compared to the models with no interim analysis. Additionally, as the subgroup number expands from four to eight, the improvement of the hierarchical model is the most among the three models.

We also explored the study designs under the six scenarios for two subgroups. The performance of each model has a similar trend as that from four or eight subgroups in terms of subgroup power, overall power, sample size, and study duration. However, the three model performance differences for two subgroups are not as large as those from four or eight subgroups. It is mainly because a smaller number of subgroups limits the borrowing property of the hierarchical model. We consequently did not present them in this paper.

Cluster hierarchical model is a good candidate for hierarchical model backup. Under some cases of the *opposite* or *one nugget* scenarios, cluster hierarchical model even performs better than a hierarchical one. Generally, clustered hierarchical model considers there are some “clusters” that exist among the subgroups, and subgroups in the same cluster have considerable influence on each other than they do on subgroups from other clusters (Facts 2018). DP scale parameter plays a more critical role in the cluster hierarchical model since as DP scale parameter goes from zero to infinite, the random distribution drawn from the base distribution behaves from very discrete to asymptotical to base distribution, i.e., the cluster number correspondingly changes from one to infinity. Consequently, when the DP scale parameter is set as greater than zero, cluster hierarchical model dilutes the impact of the hyperpriors, and it makes the cluster hierarchical model robust to the different value setting for hyperpriors. In our study, we set the DP scale parameter equal to two since the subgroup number is either four or eight. Thus, cluster hierarchical model is a good choice when no substantial evidence exists to indicate the subgroup treatment difference, but the investigator believes it should exist.

Interim analysis based on ongoing study data provides valuable information for the researcher to take related actions, such as adjusting the dosage, randomization ratio, sample size, or even stop the study as either success or futility in case there is strong proof to demonstrate it. In our DACTPerM, we keep interim analysis as one important input component of the design, which will decrease the sample size and mean study duration but maintain similar power under scenarios of *moderate and homogeneous effect and spread* for hierarchical and cluster hierarchical model. Type I error needs to be adjusted accordingly for interim and final analysis to meet the criteria that the overall Type I error rate is 0.05. We spend less than 0.005 proportion of Type I error for interim analysis and 0.045 to 0.05 for final analysis. Additionally, we define the early success under the condition that all subgroups meet the related thresholds, and the final success under the condition that some certain subgroup meet the related threshold. The initial twisting value (0.9) of the threshold at interim analysis meets our strategy. It is smaller, compared to those from the final analysis. For the final one, we need to calibrate it to meet the overall type I error, the sum of the proportions spending on both interim and final analysis, equal to 0.05. The trade-off between power and expected sample size is made in the scenarios of *opposite* and *one nugget*. The scale of trade-off is adjusted via the early stopping criteria rather than interim analysis itself. More conservative criteria will result in slight power loss, more subject enrolled and a longer study.

Longitudinal modeling applied to clinical data is reasonable, and therefore, we applied it as one design factor to provide more study information and aid in the conclusion of subgroup treatment effect. ITP and SLR are used for longitudinal data simulation and imputation. There are other methods for longitudinal data imputation. For example, a hierarchical model is a common approach, and its rationale is to generate correlated data within the visit via random effect. Based on the data from the example study, which implied the medicines work slowly and stably since earlier visits, longitudinal data simulated via ITP provides a medical process much closer to the natural process. Specifically, the responses before final visit slowly achieve the final one and maintain stably with a small variance. There are also other methods to carry out the longitudinal data imputation, like Last Observation Carried Forward (LOCF), kernel density model which is a good candidate in a case where no model assumption for the responses between interim and final ones, and so on. From the example study, the data indicates that SLR fits the data well, and provides informative priors for imputation. That SLR is straightforward and easy to understand is also a contribution for choosing it as the final imputation method. We are also the first to use ITP and SLR for longitudinal data simulation and imputation.

Another important consideration of the longitudinal modeling application is rate of accrual and dropout (i.e., missing data). Lower accrual rate makes the application difficult to improve the performance since less data information is available when execution of the interim analysis. It is also necessary to specify a realistic dropout rate since an appropriate longitudinal modelling to impute the missing data will improve the design operation characteristics. Moreover, different imputation approaches will be applied based on the different missing data mechanism. In our research, we assume the data is missing at random (MAR). Meanwhile, it is a interesting topic for future research to explore the different imputation methods for other mechanism, like missing not at random (MNAR).

Generally, when referring Bayesian adaptive clinical design, it usually means the adjustment of treatment dosage, randomization ratio, sample size, and so on. However, we do not apply those in our DACTPerM project since it is based on Bayesian RAR design in which we adjust the randomization ratio based on interim analysis results. The main objective of DACTPerM is to identify the appropriate model to analyze the non-consistent treatment effect among different subgroups. All of the models we proposed are Bayesian related since our assumption is that there should have been some proof to indicate that the treatment effect is different among the subgroups before designing related subgroup analysis. The information from the proof should be served as the priors to facilitate the final findings. In consideration of the factors above, we propose and finalize our research, although there are many other interesting topics, even though we narrowed down the subgroup analysis for different treatment within the Bayesian adaptive design.

The expected sample size and power are determined by simulation in our research. Specifically, we propose 100 per subgroup, and we tune the criteria of the posterior probability of treatment difference between two arms under the *no effect* scenario to achieve Type I error rate equal to 0.05. It is calculated via the summation of the proportion with simulated studies identified as successful under *no effect* scenario. The identical criteria then applied to other alternative response scenarios under the same study design to have the expected sample size and power via the simulation.

Lastly, we explored the three models with interim analysis and longitudinal data model in a case where the endpoint is continuous. However, one can explore and apply the approach to categorical or time to event data. To sum up, the hierarchical model with interim analysis is a relatively better approach for different subgroup treatment effect identification, and cluster hierarchical model with interim analysis is a good backup for hierarchical model in case there is no sufficient information for hyperpriors.

Chapter 3: Historical Control Bayesian Designs Incorporating Historical Control Borrowing in Clinical Trials

Other Contributors for this Chapter: Zhaowei Hua, Geng Chen, Byron Gajewski

Abstract

Incorporating historical control to concurrent study can increase the power, decrease the sample size, minimize the patient burden. It is beneficial to patients and investigators. However, the appropriate borrowing method for the study design should be researched in terms of desired operating characteristics.

We investigate several Bayesian designs incorporating historical control borrowing: power prior via overlapping area, commensurate prior, and some other reference methods. The impact of historical data type and different types of the threshold used in Bayesian decision rule are also explored. The designs' performance in terms of power as a function of treatment effect, sample size, and posterior summary are investigated via simulation.

We found that it is a good consideration to apply the power prior adaptive design with power parameter determination via overlapping area of posterior distribution under certain values of true response rates of concurrent control, historical control, and treatment effect. Study design with commensurate prior is an admissible choice as well, however, appropriate priors need to be specified.

Key words: historical control borrowing, power prior, overlapping area, commensurate prior, adaptive design, threshold

3.1 Introduction

There are several researches that incorporate external information into the current study. The external information may come from real world data (RWD, such as medical chart (Clarke and Loudon 2011, Salman, Beller, et al. 2014), patient registry (Gliklich, Dreyer, et al. 2014, Richesson 2011), natural history (NH) trial (Groft 2010)) and completed clinical trials (Bhuyan, Chen, et al. 2015). It is beneficial to patients, especially for those studies aim of rare diseases treatment or unethical to provide placebo to the patients. The Food and Drug Administration (FDA) has released guidance to regulate how to design a trial that borrows historical information (Fda. 2019). It is appealing for pharmaceutical companies since usually there are large amount of related clinical data available before a new one is conducted, especially for the control arm (Liu 2018). More resources can be used for the treatment arm.

The use of a historical control has some desired properties, such as increase in power, decrease the in size (Liu 2018), minimize the patient burden (Lim, Walley, et al. 2018), etc. The important thought of historical control borrowing is how to connect the historical data to concurrent data. There are several structures of the connection (Spiegelhalter, Abrams, et al. 2004): full equal, discounted equal, biased, similar (i.e., exchangeable), and functional dependent. Then the related methods were derived and applied accordingly.

The test-then-pool is a straightforward and frequentist method to borrow the historical control (Ghadessi, Tang, et al. 2020, Viele, Berry, et al. 2014). The idea of this method is to combine the historical control with concurrent control if the null hypothesis of equality is not rejected at significance level. In such case, the historical control is treated identically as the concurrent ones. Otherwise, historical control data will be totally ignored.

$$H_0: \theta_{hc} = \theta_{cc} \text{ vs. } H_1 : \theta_{hc} \neq \theta_{cc}$$

It is the basic form of dynamic borrowing method. The important consideration to apply this approach is how to define the significance level of the equality hypothesis, and to measure the similarity of historical control and concurrent control accurately.

The propensity score is a method that can remove the effects of confounder to borrow the external historical control. It is essentially a conditional probability of each patient being assigned to the treatment arm based on the covariates (Austin 2011, Rosenbaum and Rubin 1984). There are generally four different propensity score methods - propensity score matching, stratification (or subclassification) on the propensity score, inverse probability of treatment weighting (IPTW) using the propensity score, and covariate adjustment using the propensity score (Austin 2011). In practice, an open-label and single arm study used the propensity score to evaluate the efficacy and safety of blinatumomab (i.e., Blincyto) in patients of minimal residual disease positive (MRD+) B-cell precursor acute lymphoblastic leukemia (ALL). It was approved by FDA (Ghadessi, Tang, et al. 2020). However, FDA commented that propensity score method can yield biased estimates due to the ignorance of important unmeasured or unknown covariates. Moreover, the comparability between groups after propensity score weighted analyses is not clear because of the small sample size. Consequently, it is necessary to have sufficient data when applying propensity score.

The hierarchical model is explored and researched in historical control borrowing (Spiegelhalter, Abrams, et al. 2004, Viele, Berry, et al. 2014). The general idea is that the parameters of control data from different studies follow a prior distribution. The borrowing and shrinkage properties of hierarchical model are used to estimate the parameter of concurrent data.

Attention should be placed to the prior variance specification. The prior variance represents the degree of heterogeneity of the control parameters among the studies. The different type of priors (e.g., informative prior (Gelman 2006) or non-informative prior (Lambert, Sutton, et al. 2005)) reflects the similarity among the historical controls and concurrent control, and it will impact the concurrent parameter estimation.

Some studies researched the meta analytic predictive (MAP) prior to borrow the historical control (Gsteiger, Neuenschwander, et al. 2013, Neuenschwander, Capkun-Niggli, et al. 2010). The MAP is essentially a hierarchical model. Generally, there are two steps in MAP prior methods (Neuenschwander, Capkun-Niggli, et al. 2010). The first one is to derive the predictive distribution of control based on the posterior distribution obtained from the multiple observed historical studies. Then the predictive distribution will be served as the prior and incorporated with current study to have the posterior of concurrent control. Thus, the application assumption of hierarchical model (i.e. the exchangeability of the study parameters or priors) should also be considered for MAP. Schmidli et al. (Schmidli, Gsteiger, et al. 2014) proposed the robust MAP prior to adjust the violation of MAP assumption. Specifically, the robust MAP prior is a mixture of a MAP prior and a comparatively vague prior. The weight of MAP prior depends on the similarity of historical control and concurrent control, which will affect differently on the final posterior estimation of concurrent control.

Commensurate prior can be used to borrow historical control (Hobbs, Carlin, et al. 2011, Papageorgiou, Koretsi, et al. 2017). The connection between the historical and concurrent control is the conditional distribution of parameter of concurrent data given the historical data (Gamalo-Siebers, Savic, et al. 2017). The conditional distribution is served as the prior and incorporated with the concurrent data to have the posterior estimation of control parameter. Commensurate prior is essentially a hierarchical model as well. However, it assumes that the historical response rate is non-systematically biased from the current response rate (Lim, Walley, et al. 2018).

There are some explorations of power prior borrowing the historical data (Gravestock and Held 2018, Hobbs, Carlin, et al. 2011, Liu 2018). The degree of power prior borrowing is controlled by the power parameter of power prior. The borrowing changes from “full borrowing” to “no borrowing” as the power parameter goes from 1 to 0. The limitation of power prior is to specify an appropriate power parameter. Some researchers proposed an estimated power parameter to adjust the limitation. Specifically, the power parameter follows a distribution rather than fixed (Neelon and O' Malley 2010). However, this adjustment tends to heavily discount historical data and does not efficiently borrow the historical data unless a very informative prior used for the power parameter (Lim, Walley, et al. 2018).

This study is to research the performance of several study designs incorporating historical control via different Bayesian borrowing methods – power prior, commensurate prior and some reference borrowing method. The performance is compared by the simulating trials. In Section 3.2, we introduce the motivating pilot study, effect of bazedoxifene and conjugated estrogen (duavee®) on breast cancer risk biomarkers in high risk women (Fabian, Nye, et al. 2019), and several Bayesian borrowing methods that a study design can incorporate. In Section 3.3, the parameters for simulations and related outputs (i.e., operating characteristics) for the different possible designs are presented and compared. We demonstrate the models’ simulation-based performance. Discussion and conclusion are presented in Section 3.4.

3.2 Method

3.2.1 Motivating Study

The objective of this research is to identify the most effective Phase II study design to borrow the historical control data from a pilot study (i.e. the motivating study) conducted in high risk women with breast cancer (Fabian, Nye, et al. 2019). The motivating study investigated the effect of treatment (bazedoxifene and conjugated estrogen, i.e., duavee®) via change from baseline in mammographic total fibroglandular volume at 6 months. Specifically, it is equal to $S_6 - S_0$, where S_6 and S_0 are the fibroglandular volume at month 6 and baseline. It was observed that the proportion of the subjects with a non-increase volume at month 6 from treatment group was larger than that from the non-randomized control group. Moreover, the researchers do not want to waste the data that have already collected in the pilot study when conducting a lager phase II study. We wish to design a future prospective trial that can borrow the historical control via Bayesian method.

3.2.2 Power Prior

Power prior is a method that has been existing for a long time. In our research, the data from simulation and application studies is binary. θ_{cc} and θ_{hc} represent the response rate for the concurrent and historical control. $\boldsymbol{D}$ and $\boldsymbol{D_0}$ denote the concurrent and historical control data. We specify the Jeffrey prior for the historical data. α is the power parameter.

$$\text{Power prior}\begin{cases}\text{Historical control}\begin{cases}\text{Prior: } \pi_0(\theta_{hc}) = Beta\ (0.5, 0.5)\\ \text{Posterior: } \pi(\theta_{hc}|\boldsymbol{D_0}) \propto L(\theta_{hc}|\boldsymbol{D_0})\pi_0(\theta_{hc})\end{cases}\\ \text{Concurrent control}\begin{cases}\text{Prior: } L(\theta_{hc}|\boldsymbol{D_0})^{\alpha}\pi_0(\theta_{hc})\\ \text{Posterior: } \pi(\theta_{cc}|\boldsymbol{D}, \boldsymbol{D_0}, \alpha) \propto L(\theta_{cc}|\boldsymbol{D})L(\theta_{hc}|\boldsymbol{D_0})^{\alpha}\pi_0(\theta_{hc}) \quad , \Rightarrow\\ \theta_{cc} \sim \text{Beta}(Y_{cc} + \alpha Y_{hc} + 0.5, (n_{cc} - Y_{cc}) + \alpha(n_{hc} - Y_{hc}) + 0.5)\end{cases}\end{cases} \tag{3.1}$$

where n_{cc} and n_{hc} represent the sample size of concurrent and historical control; Y_{cc} and Y_{hc} represent the responder of concurrent and historical control.

Conventionally, the power parameter (α) is specified before the new clinical trial is conducted. It ranges from zero to one, which indicates the imparity to identity of historical and concurrent control. In our research, we let the data determine the power parameter using a heuristic algorithm, which is advantageous because of its ease in interpretation. As the adaptive design graph (Figure 3-1) indicates, the study temporally stops at interim analysis (IA) to compare the similarity of the historical and concurrent control data when half of the pre-specified same size are enrolled. The similarity is measured via the overlapping area of the posterior probability distributions of the historical and concurrent control response rate. The overlapping area (OA) is calculated via formula (3.2) denoted below:

$$\mathrm{\alpha} = \mathrm{OA} = \frac{\min\left(P(\theta_{HC} \geq \theta_{CC}), P(\theta_{HC} < \theta_{CC})\right)}{0.5}, 0 \leq \ \mathrm{OA} \leq 1 \ (3.2)$$

It is equal to the multiplication of two and minimal value of posteriors of the historical control response rate (θ_{hc}) greater than or equal to the concurrent control response rate (θ_{cc}), and the historical control response rate (θ_{hc}) less than the concurrent control response rate (θ_{cc}). We specify that the power parameter is equal to the overlapping area because they both naturally range between zero and one. Moreover, as the value changes from zero to one, they both indicate the imparity and identity of historical and concurrent control. After the interim analysis, the concurrent control enrollment will decrease accordingly based on the similarity compared to historical control. As the Figure 3-1 indicates, the actual concurrent control enrolled after the comparison is equal to the half of the proposed concurrent control sample size minus the multiplication of the overlapping area and historical control. If the calculated patients after the interim analysis is decimal, then we will have smallest integer that is greater than decimal. We assume the historical control sample size is no more than that from proposed concurrent control. We will not enroll the concurrent control patients if half of the proposed concurrent control is less than the multiplication of the overlapping area and historical control. There is no stopping rule applied at interim analysis since the treatment arm is always needed to be enrolled after interim analysis. The posterior probability of the control response rate via the power prior borrowing method follows a Beta $(Y_{cc} + \alpha Y_{hc} + 0.5, (n_{cc} - Y_{cc}) + \alpha(n_{hc} - Y_{hc}) + 0.5)$, and we provide the related derivation in Appendix 3.1. It should be noted that power prior with interim analysis is the only one incorporated into the adaptive design, all other methods introduced in the following sections are under fixed design. Moreover, the treatment arm is not involved in interim analysis.

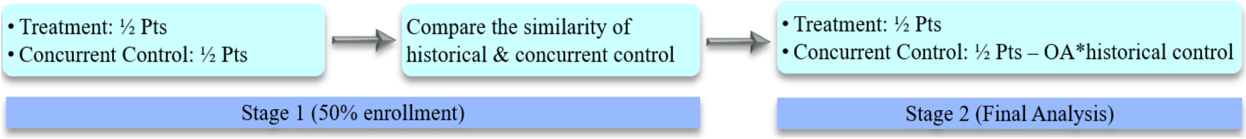

• Treatment: ½ Pts
• Concurrent Control: ½ Pts
Compare the similarity of historical & concurrent control
• Treatment: ½ Pts
• Concurrent Control: ½ Pts – OA*historical control
Stage 1 (50% enrollment)
Stage 2 (Final Analysis)

Figure 3-1 adaptive design based on power prior borrowing. Pts stands for “patients” and OA stands for “overlapping area.”

3.2.3 Commensurate Prior

Commensurate prior is essentially a hierarchical model, and we adopt the framework from Gamalo-Siebers’ research (Gamalo-Siebers, Savic, et al. 2017). The conditional distribution of θ_{cc} given θ_{hc} follows a Beta distribution with parameters $\kappa\theta_{hc}$, and $\kappa(1 - \theta_{hc})$. κ follows a Gamma distribution with the location parameter equal to K and scale parameter equal to 1. In this notation, both mean and variance are equal to K, which is convenient to specify the different types of priors. In our research, we specify K = 1, 50 and 100 to see the difference performances of the commensurate prior. The initial prior of θ_{hc} follows a non-informative Jeffrey prior. The commensurate prior is applied under the fixed study design, which is different from the power prior applied under the adaptive circumstance. Since the posterior probability of control response rate via the commensurate prior borrowing method does not have a close form, and we provide the related Stan code in the Appendix 3.2.

$$\text{Commensurate prior}\begin{cases}\text{Initial Prior for historical data } \pi(\theta_{hc}) = Beta\ (0.5, 0.5)\\ \text{Prior}\begin{cases}\pi(\theta_{cc}|\theta_{hc}) = Beta\ (\kappa\theta_{hc}, \kappa(1-\ \theta_{hc}))\\ \kappa \sim Gamma\ (K, 1)\\ K = 1, 50, 100\end{cases}\\ \text{Posterior: } \pi(\theta_{cc}|\boldsymbol{D}, \boldsymbol{D_0}, \theta_{hc}) \propto L(\theta_{cc}|\boldsymbol{D})L(\theta_{hc}|\boldsymbol{D_0})\pi(\theta_{cc}|\theta_{hc})\pi(\kappa)\pi(\theta_{hc})\end{cases} \tag{3.3}$$

3.2.4 Other borrowing methods

Full borrowing. It means that the control posterior is obtained under the combination of the historical and concurrent control. Jeffrey prior is specified for both historical and concurrent data. The posterior of the response rate follows a Beta distribution based on the conjugate property of Beta-Binomial distribution. The two parameters of Beta distribution are $(Y_{cc} + Y_{hc} + 0.5)$ and $\big((n_{cc} - Y_{cc}) + (n_{hc} - Y_{hc}) + 0.5\big)$. Attention should be placed if it is applied in a real study since the combination without differentiation may cause the incorrect posterior estimation. It is served as the reference in our research.

$$\text{Full borrowing framework}\begin{cases}\text{Prior: } \pi_0(\theta_{hc}) = Beta\ (0.5, 0.5)\\ \text{Posterior:}\begin{cases}\pi(\theta_{cc}|\boldsymbol{D}, \boldsymbol{D_0}, \theta_{hc}) \propto L(\theta_{cc}|\boldsymbol{D})L(\theta_{hc}|\boldsymbol{D_0})\pi_0(\theta_{hc})\\ \theta_{cc} \sim Beta(Y_{cc} + Y_{hc} + 0.5, (n_{cc} - Y_{cc}) + (n_{hc} - Y_{hc}) + 0.5)\end{cases}\end{cases} \tag{3.4}$$

No Borrowing. It is supposed that no historical data is involved in the posterior estimation. Still, Jeffrey prior is specified for the data. The posterior of the response rate also follows a Beta distribution with the parameters of $(Y_{cc} + 0.5)$ and $\big((n_{cc} - Y_{cc}) + 0.5\big)$. Similarly, it is served as the reference to be compared with the power prior and commensurate prior.

$$\text{No borrowing framework}\begin{cases} \text{Prior: } \pi_0(\theta_{cc}) = Beta\ (0.5, 0.5) \\ \text{Posterior:} \begin{cases} \pi(\theta_{cc}|\boldsymbol{D}) \propto L(\theta_{cc}|\boldsymbol{D})\pi_0(\theta_{cc}) \\ \theta_{cc} \sim Beta(Y_{hc} + 0.5, (n_{cc} - Y_{cc}) + 0.5) \end{cases} \end{cases} \tag{3.5}$$

Frequentist method. The frequentist estimation should be quite similar with the ones from Bayesian estimation under the no borrowing framework. The specific method applied is Chi-square test. We adopt the frequentist estimation in terms of point estimation, bias and MSE to validate this assumption. It should be noted that it means Chi-square test when referring the Frequentist method in this paper. All other methods, including full borrowing, no borrowing and Frequentist, are under fixed designs.

3.3 Simulation

3.3.1 Simulation Input

Control data. As specified in the method part, the data in our research is binary. The historical (θ_{hc}) and concurrent control response rates (θ_{cc}) range from 0.1 to 0.5 by 0.1. Table 3 - 1 summarizes the parameter value for simulation. The concurrent control data is obtained via simulation. The historical control data is generated via simulation and "observation" which means the responder is calculated via the multiplication of response rate and the historical control sample size, supposing the historical data is observed.

Effect size & treatment. In our research, we use the difference of response rate from the treatment and concurrent control arms as the effect size (denoted as $\theta_t - \theta_{cc}$). The proposed effect sizes range from 0.1 to 0.4 by 0.1. Together with the span of control data, they will evaluate the different Bayesian methods thoroughly and comprehensively. The treatment response rate is equal to the summation of concurrent response rate and effect size.

Table 3-1 Summary of the parameter values for simulation

Parameter	Value
θ_{hc}	0.1, 0.2, 0.3, 0.4, 0.5
θ_{cc}	0.1, 0.2, 0.3, 0.4, 0.5
$\theta_t - \theta_{cc}$	0, 0.1, 0.2, 0.3, 0.4

Proposed Sample size. From practical perspective, it is seldom that pivotal studies from routine disease area (e.g., hypertension, diabetes, oncology, etc.) incorporate historical control and get approved by FDA. Most of the pivotal studies that incorporated historical control are from rare disease (Ghadessi, Tang, et al. 2020). For our research, the proposed sample size cannot be large. The proposed historical control, concurrent control and treatment sample sizes in our research are 20, 20 and 40, respectively. The sample sizes of historical control and treatment are fixed. The expected sample size of concurrent control may be not identical to the proposed one depending on the similarity of historical and concurrent control.

Threshold. We propose three types of the threshold – global, local and regional- for Bayesian decision rule. For the global threshold, it means that it controls type I error less than or equal to 0.025 for the study designs under all concurrent control response rates (i.e., 0.1 to 0.5 by 0.1) given a specific response rate of historical control under the null hypothesis (i.e., effect size is equal to zero). For the local threshold, it means it controls type I error equal to 0.025 for the study design under the concurrent control response rate equal to the specific historical control response rate under the null hypothesis. For the regional threshold is chosen to partially guarantee that the type I error less than or equal to 0.025 for the study designs borrowing the historical control with a specific response rate and with a limited and related concurrent control response rates, i.e., $\theta_{cc} \in [\theta_{hc} - \mathrm{s.e.}, \theta_{hc} + \mathrm{s.e.}]$. The different threshold types reflect researchers' belief of the similarity of concurrent and historical control. The number of simulated studies for each method is 20,000, and the iteration number is 40,000 times for those designs with Bayesian borrowing.

3.3.2 Simulation Output

Type I error. Our research hypothesis is one-sided. Specifically, the null hypothesis is that the response rates from both arms are equal. The alternative hypothesis is that the treatment response rate is greater than the concurrent control response rate (expression (3.6) below). Type I error is controlled to 0.025. Given a specific study design, the threshold is twisted and determined to make the proportion of the simulated studies with the quantity of interest $\left(\text{i. e.}, P(\theta_t > \theta_{cc}|Data)\right)$ greater than the threshold under the null hypothesis is equal to 0.025.

$$H_0: \theta_t = \theta_{cc} \text{ vs. } H_1 : \theta_t > \theta_{cc}, \text{ where } \theta_t \text{ denotes the treatment response rate. (3.6)}$$

The thresholds of different type are determined by the definition accordingly. The Bayesian decision rule is that under the null hypothesis, the proportion of simulated studies with the posterior probability of quantity of interest great than threshold is less than or equal to 0.025 (expression (3.7) below). The posterior of concurrent control has already incorporated historical control based on the specific borrowing method. The global threshold is chosen to guarantee that the type I error less than or equal to 0.025 for the study designs under all possible concurrent control response rates and borrowing the historical control with a specific response rate.

$$[\text{Pr}[P(\theta_t > \theta_{cc}|Data) > \text{threshold}, \text{where } \theta_{cc} \in [0.1 \text{ to } 0.5 \text{ by } 0.1]\ |H_0] \leq 0.025 \text{ (3.7)}$$

As the expression (3.8) below involved in the local threshold, the Bayesian decision rule is that under the null hypothesis, the proportion of simulated studies with the posterior probability of quantity of interest (i.e., $\theta_t > \theta_{cc}$) great than threshold is equal to 0.025. Still, the posterior of concurrent control has already incorporated historical control based on the specific borrowing method. The local threshold is determined to only guarantee that the type I error is equal to 0.025 under the condition of the concurrent control response rate equal to the historical control for borrowing.

$$\Pr[P(\theta_t > \theta_{cc}|Data) > \text{threshold, where } \theta_{cc} = \theta_{hc}\ |H_0] = 0.025 \text{ (3.8)}$$

The definition of the regional threshold has the identical rationale as that from the global threshold. It is defined to guarantee that the type I error is less than or equal to 0.025 for the study designs borrowing the historical control with a specific response rate and with a limited range of and related concurrent control response rates that related to historical control response rate (expression (3.9) below). The different threshold types are only applicable for the methods that historical control is borrowed (i.e., power prior, commensurate prior and full borrowing), otherwise, threshold is only identified via the current study simulated data. The specific thresholds are provided in Appendix 3.3.

$$\Pr[P(\theta_t > \theta_{cc}|Data) > \text{threshold, where } \theta_{cc} \in [\theta_{hc} - \text{s.e.}, \theta_{hc} + \text{s.e.}]\ |H_0]\ \leq 0.025 \text{ (3.9)}$$

Power (Study success). Simulations track the proportion of studies that show success based on the evaluation criterion (i.e., threshold) identified under the hypothesis. Based on the Bayesian decision rule, the power is generally defined as the proportion of the simulated studies that meet the evaluation criteria under the alternative hypothesis (i.e., $\theta_t > \theta_{cc}$). The evaluation criterion is the quantity of interest satisfying the related threshold. The specific powers of study design with the historical borrowing based on different threshold types are then calculated based on each threshold value. (expression (3.10), (3.11) and (3.12) below).

$$\Pr[P(\theta_t > \theta_{cc}|Data) > \text{threshold, where } \theta_{cc} \in [0.1 \text{ to } 0.5 \text{ by } 0.1]\ |H_1]\ (3.10)$$

$$\Pr[P(\theta_t > \theta_{cc}|Data) > \text{threshold, where } \theta_{cc} = \theta_{hc}\ |H_1]\ (3.11)$$

$$\Pr[P(\theta_t > \theta_{cc}|Data) > \text{threshold, where } \theta_{cc} \in [\theta_{hc} - \text{s.e.}, \theta_{hc} + \text{s.e.}]\ |H_1]\ (3.12)$$

Expected sample size. Sample size is important operation characteristic since it directly relates to the difficulty and cost of running a trial, especially for the trials with a low accrual rate. A study design with a lower sample size but similar power to a competing design is desirable. In our research, only the design with power prior may have different expected sample size. All other designs are fixed, and the expected sample size is equal to the proposed sample size.

Posterior summary. The posterior summary in terms of point estimation, credible interval, bias and mean square error (MSE) are presented to compare the performance of different study designs.

3.3.3 Simulation Result

Figure 3-2 presents the power of different study designs under different observed historical control rate and effect sizes via global thresholds. When historical control response rate (θ_{hc}) and effect size ($\theta_t - \theta_{cc}$) are both equal to 0.1, the powers of all the study designs with different borrowing methods are generally below 0.2 for all values of concurrent control response rate ($\theta_{cc}'s$). When ($\theta_t - \theta_{cc}$) becomes 0.2 and θ_{hc} is still equal to 0.1, the powers of study designs with no borrowing and frequentist are generally between 0.3 and 0.4. They are higher than those of the study designs with other borrowing methods when θ_{cc} is equal to 0.1, 0.2 and 0.3. The powers of the study designs with commensurate priors ($K = 50, 100$) and full borrowing are quite similar (around 0.35) with that from the study design with no borrowing or frequentist when θ_{cc} is equal to 0.4, and they are higher (around 0.5) when θ_{cc} is equal to 0.5. The powers of study design with power prior borrowing and commensurate priors ($K = 1$) are generally lower than those from the study design with other borrowing methods when θ_{cc} is equal to 0.4 or 0.5, but higher than those from the study design with commensurate priors ($K = 50, 100$) and full borrowing when θ_{cc} is equal to 0.1. Similar findings are identified for the figure panel where θ_{hc} is equal to 0.1 and $(\theta_t - \theta_{cc})$ is equal to 0.3 and 0.4. The main difference is that when θ_{cc} are close to 0.3 or equal to 0.3, the powers of the designs with no borrowing or frequentist are similar with those from the study designs with commensurate priors ($K = 50, 100$) and full borrowing.

When θ_{hc} increases to 0.2, $(\theta_t - \theta_{cc})$ ranges from 0.1 to 0.4 and, θ_{cc} ranges from 0.1 to 0.5, the power profiles are quite similar with those θ_{hc} equal to 0.1. All study designs are with a general higher power. When θ_{hc} increases to 0.3, the major change is from the power profiles where study designs with power prior. There is a clear trend that the power increases as θ_{cc} ranges from 0.1 to 0.5. When θ_{hc} increases to 0.4 and 0.5, the overall power profiles are still similar compared to them where θ_{hc} is equal to 0.2. Moreover, when θ_{hc} is equal to 0.4, the power of the study design with power prior is almost close to the highest ones where θ_{cc} is equal to 0.5 and $(\theta_t - \theta_{cc})$ is equal to 0.2 or 0.3. When θ_{hc} is equal to 0.5, the power profile of the study design with power prior is quite like a “bowl” where θ_{cc} ranges from 0.1 to 0.5 and $(\theta_t - \theta_{cc})$ is equal to 0.3. The power profiles of different study designs under different simulated historical control rate and effect sizes via global thresholds are generally similar with the related ones from Figure 3-2. The main distinction is that the power differences among the study designs are not so large as those from Figure 3-2.

Figure 3-4 & 3-5 present the power of different study designs under different observed and simulated historical control rate and effect sizes via local thresholds. It only presents the power profiles where $\theta_{hc} = \theta_{cc}$, because it is more important to observe the power points in the graphs where $\theta_{hc}'s$ are equal to θ_{cc}'s since the thresholds are locally controlled type I error equal to 0.025 at θ_{hc} = θ_{cc}. It is clearly observed that the powers of the study designs with different methods are quite similar with each other when $(\theta_t - \theta_{cc})$ are equal to 0.1 and θ_{hc} ranges from 0.1 to 0.5. Generally, all the powers increase accordingly when $(\theta_t - \theta_{cc})$ increases to 0.4. The powers of study designs with commensurate priors ($K = 50, 100$) and full borrowing are quite similar and higher than those from other study designs when $(\theta_t - \theta_{cc})$ is larger than 0.1 and θ_{hc} is greater than 0.1 as well. The powers of study designs with commensurate priors ($K = 1$), full borrowing and no borrowing are quite similar and lower than those from other study designs, except for the scenarios where $(\theta_t - \theta_{cc})$ is larger than 0.1 and θ_{hc} is equal to 0.1. In some scenarios, the powers of the study designs with commensurate priors ($K = 1$) are the lowest one. The powers of the study designs with power priors are generally between these two “clusters”. Moreover, under related observed historical control rate scenarios, the power from the power prior borrowing method is quite close to the higher ones in the scenarios where θ_{hc} and $(\theta_t - \theta_{cc})$ are both equal to 0.3 and 0.4, and θ_{hc} is equal to 0.5 and $(\theta_t - \theta_{cc})$ are equal to 0.3.

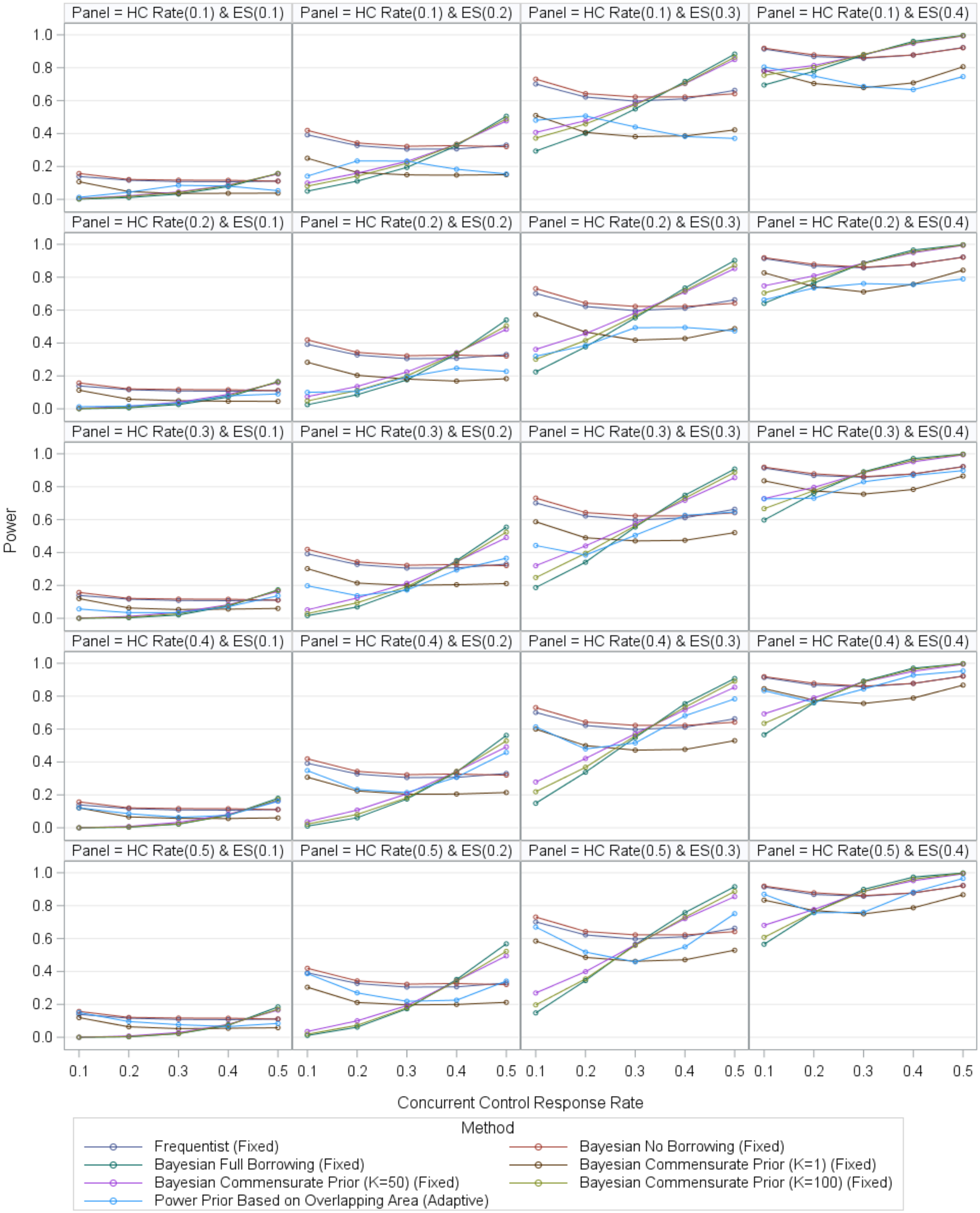


Figure 3-2 Power of different study designs with different borrowing method under different observed historical control rate (HC Rate) (θ_{hc} ϵ [0.1 to 0.5 by 0.1]) and effect sizes (ES) ($\theta_t - \theta_{cc}$ϵ[0.1 to 0.4 by 0.1]) via global thresholds. Concurrent Control Response Rate (θ_{cc} ϵ [0.1 to 0.5 by 0.1]). “Fixed” and “Adaptive” in the parenthesis of legend mean the related methods incorporated in the fixed or adaptive design.

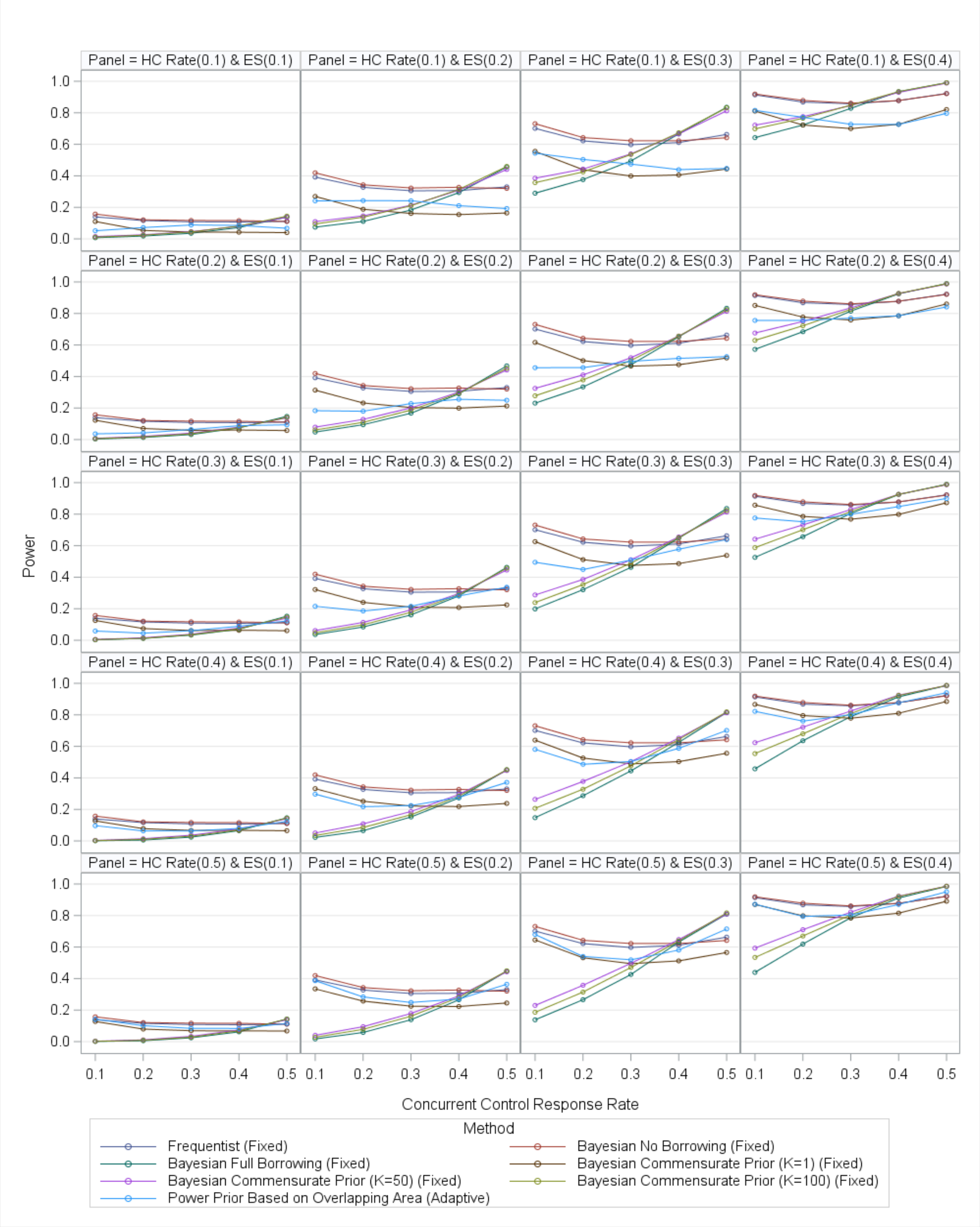


Figure 3-3 Power of different study designs with different borrowing method under different observed historical control rate (HC Rate) (θ_{hc} ϵ [0.1 to 0.5 by 0.1]) and effect sizes (ES) ($\theta_t - \theta_{cc}$ϵ[0.1 to 0.4 by 0.1]) via global thresholds. Concurrent Control Response Rate (θ_{cc} ϵ [0.1 to 0.5 by 0.1]). “Fixed” and “Adaptive” in the parenthesis of legend mean the related methods incorporated in the fixed or adaptive design.

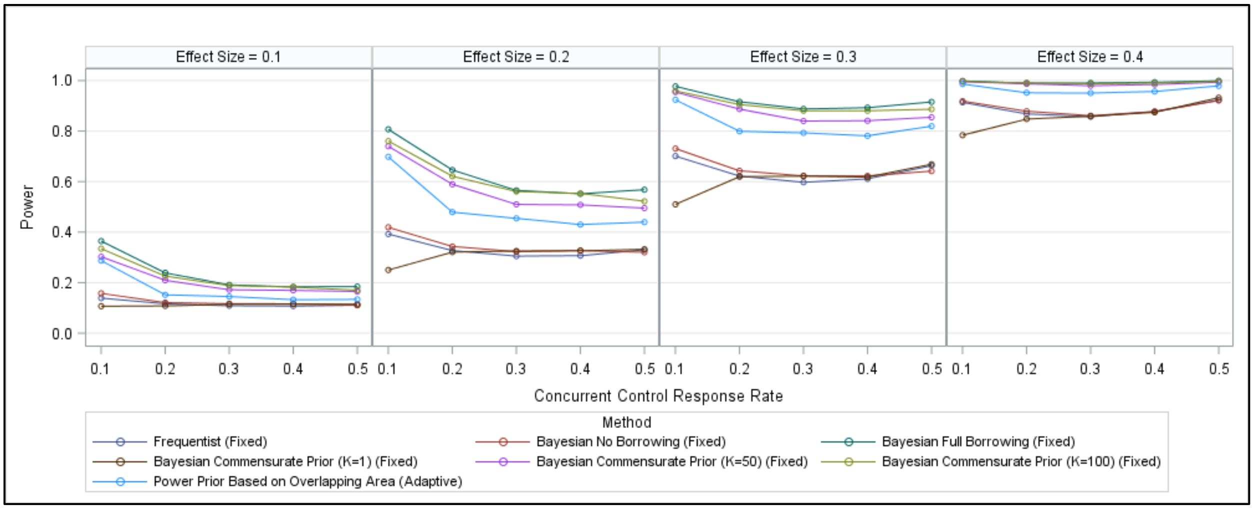


Figure 3-4 Power of different study designs with different borrowing method under different observed historical control rate ($\theta_{hc} = \theta_{cc}$) and effect sizes (0.1 to 0.4 by 0.1) via local thresholds. “Fixed” and “Adaptive” in the parenthesis of legend mean the related methods incorporated in the fixed or adaptive design.

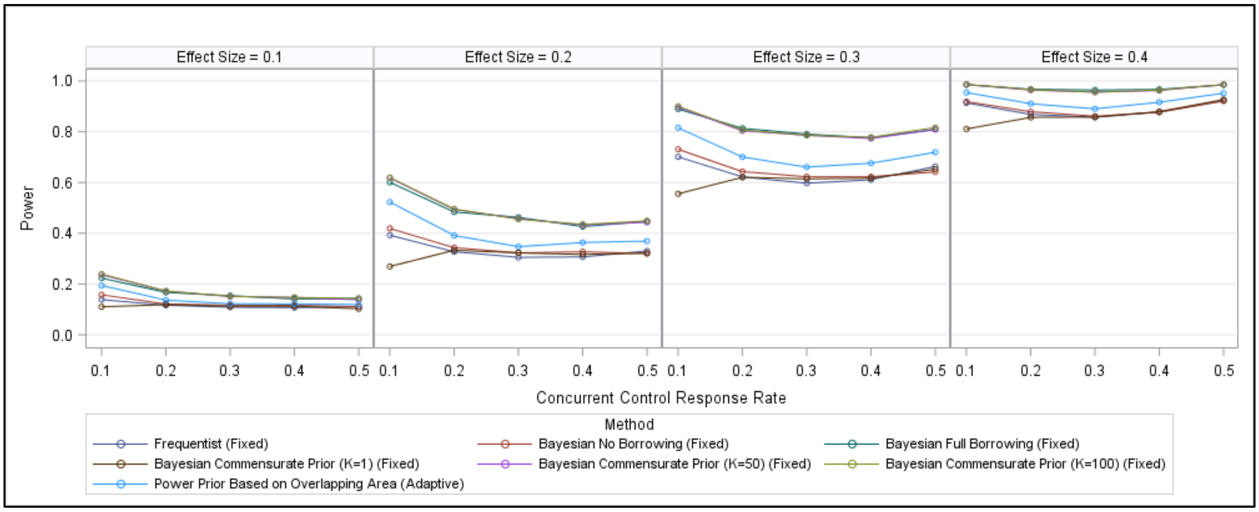


Figure 3-5 Power of different study designs with different borrowing method under different simulated historical control rate ($\theta_{hc} = \theta_{cc}$) and effect sizes (0.1 to 0.4 by 0.1) via local thresholds. “Fixed” and “Adaptive” in the parenthesis of legend mean the related methods incorporated in the fixed or adaptive design.

Figure 3-6 & 3-7 present the power of different study designs under different observed and simulated historical control rate and effect sizes via regional thresholds. We mainly focus on the response rate between 0.1 and 0.5. Thus, when θ_{hc} is equal to 0.1, the stick values of θ_{cc} on X-axis represent 0.1, 0.1 + 0.25se, 0.1 + 0.5se, 0.1 + 0.75se and 0.1 + se. When θ_{hc} is equal to 0.5, the stick values of θ_{cc} on X-axis represent 0.5 – se, 0.5 - 0.75se, 0.5 - 0.5se, 0.5 - 0.25se, and

0.5. When θ_{hc} is equal to 0.2, 0.3 or 0.4, the stick values of θ_{cc} on X-axis represent θ_{hc} ± se, θ_{hc} ± 0.5se, and θ_{hc}. When θ_{hc} is equal to 0.2, 0.3 or 0.4, the power profiles are generally similar with the pattern from the related study designs with global thresholds. The powers of the study designs with different methods are quite similar with each other when $\theta_t - \theta_{cc}$is equal to 0.1 and θ_{hc} ranges from 0.1 to 0.5. The powers of the study designs with commensurate priors ($K = 50, 100$) and full borrowing are quite similar and higher than those from other study designs when $\theta_t - \theta_{cc}$ is greater than 0.1, θ_{hc} is greater than 0.2, and θ_{cc} is equal to θ_{hc} + 0.75se or θ_{hc} + se. Correspondingly, the powers of study designs with commensurate priors ($K = 1$), full borrowing and no borrowing are quite similar and lower than those from other study designs, except for the scenarios where θ_{hc} is equal to 0.1 or 0.2. However, when θ_{cc} is equal to θ_{hc} - 0.75se or θ_{hc} - se, the powers of study designs with commensurate priors ($K = 1$), full borrowing and no borrowing are quite similar and higher than those from other study designs, except for the scenarios where θ_{hc} is equal to 0.1 or 0.2. The powers of the study designs with power priors are generally between these two “clusters”.

Table 3 - 2 below presents overlapping area (OA) and related concurrent control enrollment after interim analysis of the study designs with power prior borrowing under the global threshold. It is clearly observed that the OA is generally the largest and the concurrent control enrollment after the interim analysis is correspondingly the least when θ_{hc} is equal to θ_{cc}. The OA generally decreases and the concurrent control enrollment after the interim analysis is correspondingly increase as the differences between θ_{hc} and θ_{cc} increase. The OA for simulated historical control is generally smaller and the concurrent control enrollment after the interim analysis is correspondingly larger, comparing related OA and enrollment from those designs with observed historical control.

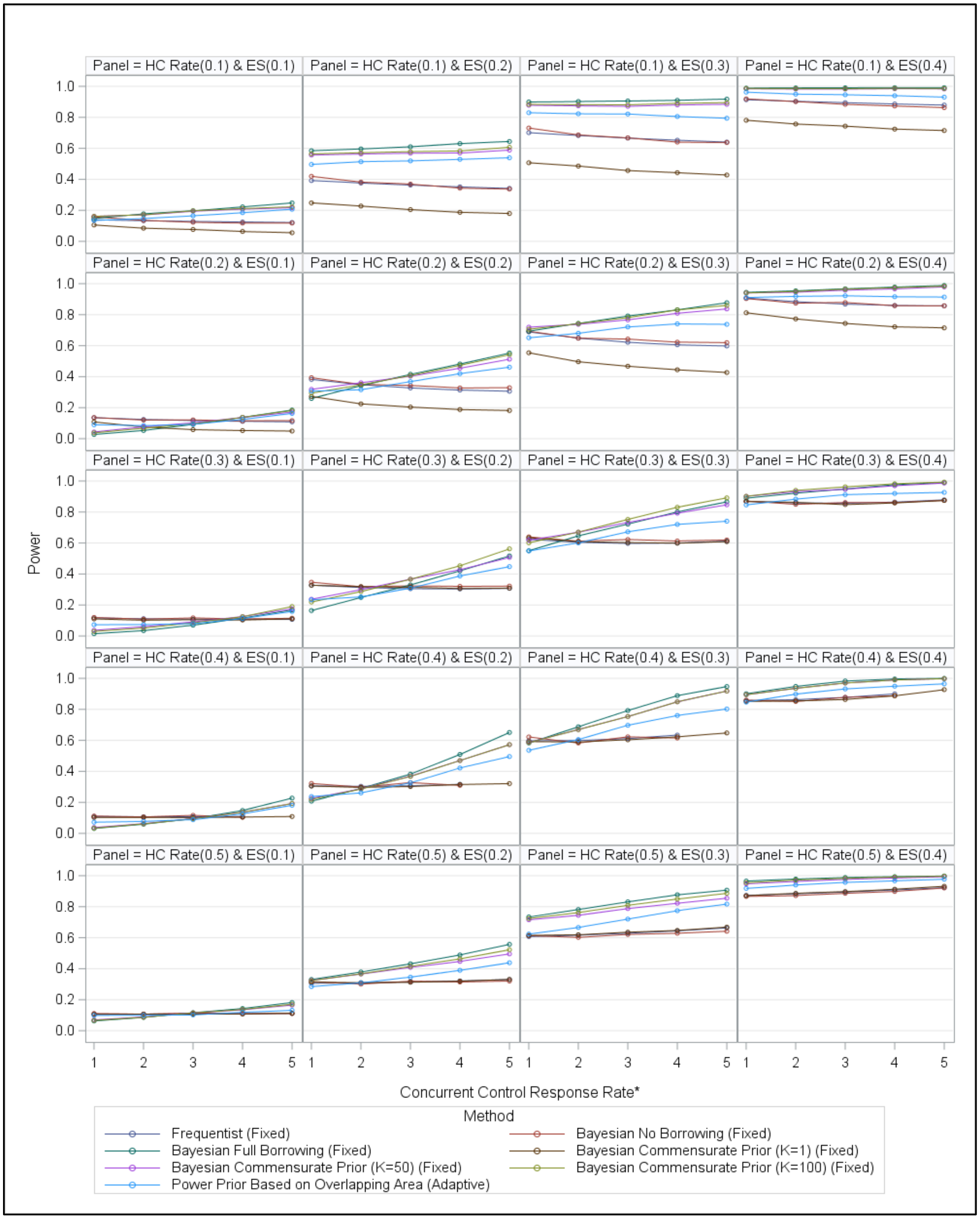


Figure3-6 Power of different study designs with different borrowing method under different observed historical control rate (HC Rate) (θ_{hc} ϵ [0.1 to 0.5 by 0.1]) and effect sizes (0.1 to 0.4 by 0.1) via regional thresholds.

*: For θ_{hc} = 0.1, the θ_{cc} value on X-axis: 1= θ_{hc}, 2= θ_{hc}+ 0.25se, 3= θ_{hc}+ 0.5se, 4= θ_{hc}+ 0.75se, 5=θ_{hc}+ se.
For θ_{hc} = 0.2, 0.3 and 0.4, the X-axis stick represents θ_{cc}: 1= θ_{hc}- se, 2= θ_{hc}- 0.5se, 3= θ_{hc}, 4= θ_{hc}+ 0.5se, 5=θ_{hc}+ se.
For θ_{hc} = 0.5, the θ_{cc} value on X-axis: 1= θ_{hc}- se, 2= θ_{hc}- 0.75se, 3= θ_{hc}- 0.5se, 4= θ_{hc}- 0.25se, 5=θ_{hc}.
In each panel, “HC rate” means historical response rate, and ES means effect size. The “HC(0.1) & ES(0.1)” means historical response rate equal to 0.1 and effect size equal to 0.1. Same rationale for all other panels.
“Fixed” and “Adaptive” in the parenthesis of legend mean the related methods incorporated in the fixed or adaptive design.

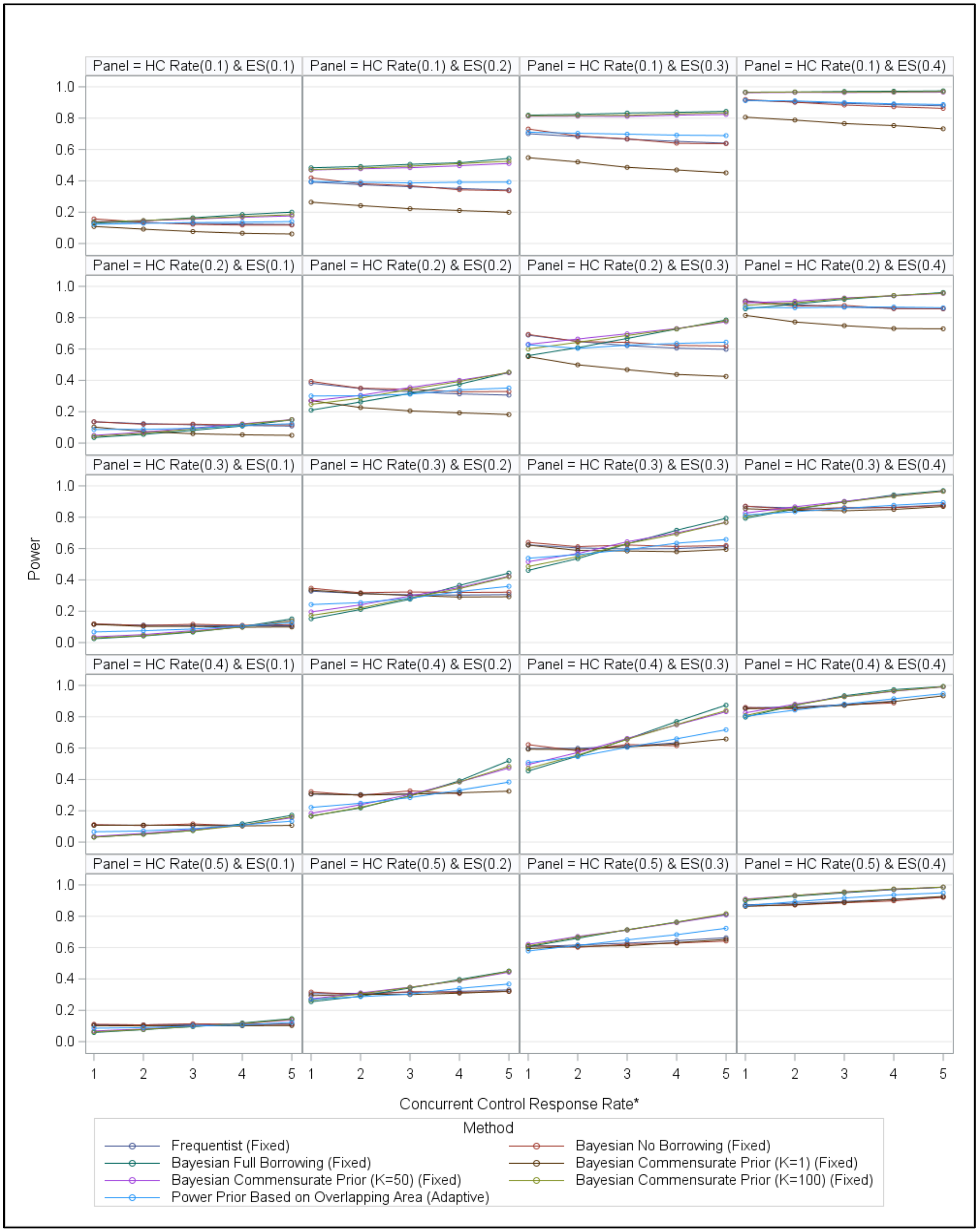
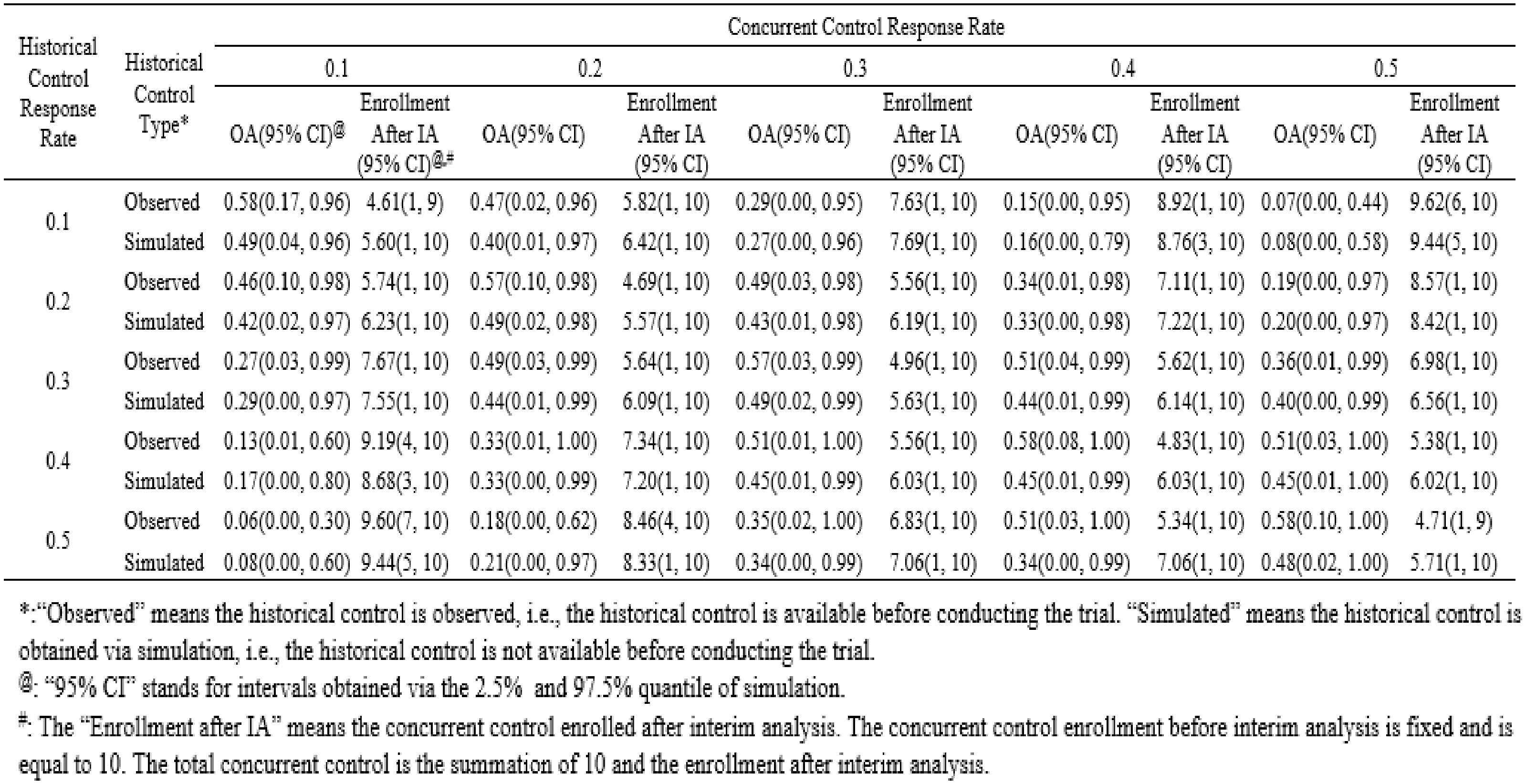


Figure 3-7 Power of different study designs with different borrowing method under different simulated historical control rate (θ_{hc} ϵ [0.1 to 0.5 by 0.1]) and effect sizes (0.1 to 0.4 by 0.1) via regional thresholds.

*: For θ_{hc} = 0.1, the θ_{cc} value on X-axis: 1= θ_{hc}, 2= θ_{hc}+ 0.25se, 3= θ_{hc}+ 0.5se, 4= θ_{hc}+ 0.75se, 5=θ_{hc}+ se.
For θ_{hc} = 0.2, 0.3 and 0.4, the X-axis stick represents θ_{cc}: 1= θ_{hc}- se, 2= θ_{hc}- 0.5se, 3= θ_{hc}, 4= θ_{hc}+ 0.5se, 5=θ_{hc}+ se.
For θ_{hc} = 0.5, the θ_{cc} value on X-axis: 1= θ_{hc}- se, 2= θ_{hc}- 0.75se, 3= θ_{hc}- 0.5se, 4= θ_{hc}- 0.25se, 5=θ_{hc}.
In each panel, “HC rate” means historical response rate, and ES means effect size. The “HC(0.1) & ES(0.1)” means historical response rate equal to 0.1 and effect size equal to 0.1. Same rationale for all other panels.
“Fixed” and “Adaptive” in the parenthesis of legend mean the related methods incorporated in the fixed or adaptive design.

Table 3-2 the overlapping area (OA) and related concurrent control enrollment after interim analysis (IA) for power prior borrowing under global threshold



Historical Control Response Rate	Historical Control Type*	Concurrent Control Response Rate									
		0.1		0.2		0.3		0.4		0.5	
		OA(95% CI)@	Enrollment After IA (95% CI)@,#	OA(95% CI)	Enrollment After IA (95% CI)	OA(95% CI)	Enrollment After IA (95% CI)	OA(95% CI)	Enrollment After IA (95% CI)	OA(95% CI)	Enrollment After IA (95% CI)
0.1	Observed	0.58(0.17, 0.96)	4.61(1, 9)	0.47(0.02, 0.96)	5.82(1, 10)	0.29(0.00, 0.95)	7.63(1, 10)	0.15(0.00, 0.95)	8.92(1, 10)	0.07(0.00, 0.44)	9.62(6, 10)
	Simulated	0.49(0.04, 0.96)	5.60(1, 10)	0.40(0.01, 0.97)	6.42(1, 10)	0.27(0.00, 0.96)	7.69(1, 10)	0.16(0.00, 0.79)	8.76(3, 10)	0.08(0.00, 0.58)	9.44(5, 10)
0.2	Observed	0.46(0.10, 0.98)	5.74(1, 10)	0.57(0.10, 0.98)	4.69(1, 10)	0.49(0.03, 0.98)	5.56(1, 10)	0.34(0.01, 0.98)	7.11(1, 10)	0.19(0.00, 0.97)	8.57(1, 10)
	Simulated	0.42(0.02, 0.97)	6.23(1, 10)	0.49(0.02, 0.98)	5.57(1, 10)	0.43(0.01, 0.98)	6.19(1, 10)	0.33(0.00, 0.98)	7.22(1, 10)	0.20(0.00, 0.97)	8.42(1, 10)
0.3	Observed	0.27(0.03, 0.99)	7.67(1, 10)	0.49(0.03, 0.99)	5.64(1, 10)	0.57(0.03, 0.99)	4.96(1, 10)	0.51(0.04, 0.99)	5.62(1, 10)	0.36(0.01, 0.99)	6.98(1, 10)
	Simulated	0.29(0.00, 0.97)	7.55(1, 10)	0.44(0.01, 0.99)	6.09(1, 10)	0.49(0.02, 0.99)	5.63(1, 10)	0.44(0.01, 0.99)	6.14(1, 10)	0.40(0.00, 0.99)	6.56(1, 10)
0.4	Observed	0.13(0.01, 0.60)	9.19(4, 10)	0.33(0.01, 1.00)	7.34(1, 10)	0.51(0.01, 1.00)	5.56(1, 10)	0.58(0.08, 1.00)	4.83(1, 10)	0.51(0.03, 1.00)	5.38(1, 10)
	Simulated	0.17(0.00, 0.80)	8.68(3, 10)	0.33(0.00, 0.99)	7.20(1, 10)	0.45(0.01, 0.99)	6.03(1, 10)	0.45(0.01, 0.99)	6.03(1, 10)	0.45(0.01, 1.00)	6.02(1, 10)
0.5	Observed	0.06(0.00, 0.30)	9.60(7, 10)	0.18(0.00, 0.62)	8.46(4, 10)	0.35(0.02, 1.00)	6.83(1, 10)	0.51(0.03, 1.00)	5.34(1, 10)	0.58(0.10, 1.00)	4.71(1, 9)
	Simulated	0.08(0.00, 0.60)	9.44(5, 10)	0.21(0.00, 0.97)	8.33(1, 10)	0.34(0.00, 0.99)	7.06(1, 10)	0.34(0.00, 0.99)	7.06(1, 10)	0.48(0.02, 1.00)	5.71(1, 10)

*:"Observed" means the historical control is observed, i.e., the historical control is available before conducting the trial. "Simulated" means the historical control is obtained via simulation, i.e., the historical control is not available before conducting the trial.

@: "95% CI" stands for intervals obtained via the 2.5% and 97.5% quantile of simulation.

#: The "Enrollment after IA" means the concurrent control enrolled after interim analysis. The concurrent control enrollment before interim analysis is fixed and is equal to 10. The total concurrent control is the summation of 10 and the enrollment after interim analysis.

Table 3-3 the overlapping area (OA) and related concurrent control enrollment after interim analysis (IA) for power prior borrowing under local threshold

Historical Control Type*	Concurrent Control Response Rate									
	0.1		0.2		0.3		0.4		0.5	
	OA(95% CI) @	Enrollment After IA(95% CI) @#	OA(95% CI)	Enrollment After IA(95% CI)	OA(95% CI)	Enrollment After IA(95% CI)	OA(95% CI)	Enrollment After IA(95% CI)	OA(95% CI)	Enrollment After IA(95% CI)
Observed	0.58(0.17, 0.96)	4.61(1.00, 9.00)	0.57(0.10, 0.98)	4.69(1.00, 10.00)	0.57(0.03, 0.99)	4.96(1.00, 10.00)	0.58(0.078, 0.998)	4.83(1.00, 10.00)	0.58(0.10, 1.00)	4.71(1.00, 9.00)
Simulated	0.49(0.04, 0.96)	5.60(1.00, 10.00)	0.49(0.02, 0.98)	5.57(1.00, 10.00)	0.49(0.02, 0.99)	5.63(1.00, 10.00)	0.49(0.018, 0.996)	5.65(1.00, 10.00)	0.50(0.02, 1.00)	5.59(1.00, 10.00)

*:"Observed" means the historical control is observed, i.e., the historical control is available before conducting the trial. "Simulated" means the historical control is obtained via simulation, i.e., the historical control is not available before conducting the trial.

@: "95% CI" stands for interval obtained via the 2.5% and 97.5% quantile of simulation.

#: The "Enrollment after IA" means the concurrent control enrolled after interim analysis. The concurrent control enrollment before interim analysis is fixed and is equal to 10. The total concurrent control is the summation of 10 and the enrollment after interim analysis.

Table 3-4 the overlapping area (OA) and related concurrent control enrollment after interim analysis (IA) for power prior borrowing under regional threshold

Historical Control Response Rate	Historical Control Type$	Concurrent Control Response Rate									
		= θ_{hc}-se%		= θ_{hc}-0.5se		= θ_{hc}		θ_{hc}+0.5se		θ_{hc}+se	
		OA (95% CI) @	Enrollment After IA (95% CI) @#	OA (95% CI)	Enrollment After IA (95% CI)	OA (95% CI)	Enrollment After IA (95% CI)	OA (95% CI)	Enrollment After IA (95% CI)	OA (95% CI)	Enrollment After IA (95% CI)
0.1*	Observed	0.58(0.17, 0.96)	4.60(1, 9)	0.57(0.17, 0.96)	4.71(1, 9)	0.56(0.06, 0.96)	4.88(1, 10)	0.54(0.06, 0.96)	5.05(1, 10)	0.52(0.06, 0.96)	5.28(1, 10)
	Simulated	0.49(0.04, 0.96)	5.63(1, 10)	0.48(0.03, 0.96)	5.66(1, 10)	0.47(0.02, 0.96)	5.80(1, 10)	0.45(0.01, 0.96)	5.93(1, 10)	0.44(0.01, 0.96)	6.07(1, 10)
0.2	Observed	0.48(0.10, 0.98)	5.53(1, 10)	0.55(0.10, 0.98)	4.88(1, 10)	0.57(0.10, 0.98)	4.67(1, 10)	0.55(0.09, 0.98)	4.94(1, 10)	0.51(0.03, 0.98)	5.37(1, 10)
	Simulated	0.44(0.02, 0.97)	6.07(1, 10)	0.48(0.03, 0.98)	5.65(1, 10)	0.49(0.02, 0.98)	5.61(1, 10)	0.48(0.02, 0.98)	5.72(1, 10)	0.45(0.01, 0.98)	6.03(1, 10)
0.3	Observed	0.48(0.03, 0.99)	5.73(1, 10)	0.55(0.03, 0.99)	5.12(1, 10)	0.58(0.03, 0.99)	4.93(1, 10)	0.55(0.04, 0.99)	5.17(1, 10)	0.51(0.04, 0.99)	5.62(1, 10)
	Simulated	0.44(0.01, 0.99)	6.13(1, 10)	0.48(0.02, 0.99)	5.71(1, 10)	0.49(0.02, 0.99)	5.65(1, 10)	0.48(0.02, 0.99)	5.76(1, 10)	0.44(0.01, 0.99)	6.13(1, 10)
0.4	Observed	0.49(0.01, 1.00)	5.75(1, 10)	0.56(0.08, 1.00)	5.06(1, 10)	0.58(0.08, 1.00)	4.85(1, 10)	0.55(0.04, 1.00)	5.04(1, 10)	0.50(0.03, 1.00)	5.50(1, 10)
	Simulated	0.44(0.01, 0.99)	6.12(1, 10)	0.48(0.02, 1.00)	5.76(1, 10)	0.50(0.02, 1.00)	5.60(1, 10)	0.48(0.02, 1.00)	5.77(1, 10)	0.44(0.01, 1.00)	6.14(1, 10)
0.5*	Observed	0.49(0.02, 1.00)	5.50(1, 10)	0.53(0.02, 1.00)	5.18(1, 10)	0.55(0.03, 1.00)	4.93(1, 10)	0.57(0.10, 1.00)	4.74(1, 9)	0.58(0.10, 1.00)	4.71(1, 9)
	Simulated	0.44(0.01, 1.00)	6.14(1, 10)	0.47(0.01, 1.00)	5.85(1, 10)	0.48(0.02, 1.00)	5.73(1, 10)	0.49(0.02, 1.00)	5.64(1, 10)	0.50(0.02, 1.00)	5.57(1, 10)

*: For θ_{hc} = 0.1, the θ_{cc} values are θ_{hc}, θ_{hc}+ 0.25se, θ_{hc}+ 0.5se, θ_{hc}+ 0.75se, and θ_{hc}+ se. For θ_{hc} = 0.5, the θ_{cc} values are θ_{hc}- se, 2= θ_{hc}- 0.75se, 3= θ_{hc}- 0.5se, 4= θ_{hc}- 0.25se, 5=θ_{hc}.

$:"Observed" means the historical control is observed, i.e., the historical control is available before conducting the trial. "Simulated" means the historical control is obtained via simulation, i.e., the historical control is not available before conducting the trial.

@: "95% CI" stands for interval obtained via the 2.5% and 97.5% quantile of simulation.

#: The "Enrollment after IA" means the concurrent control enrolled after interim analysis. The concurrent control enrollment before interim analysis is fixed and is equal to 10. The total concurrent control is the summation of 10 and the enrollment after interim analysis.

%: "se" stands for standard error.

Table 3 - 3 below presents overlapping area (OA) and related concurrent control enrollment after interim analysis (IA) of the study designs with power prior borrowing under the local threshold. It is observed that only about half of the patients need to be enrolled after the interim analysis. The enrollment from the related simulated historical data needs slightly more patients enrolled.

Table 3 - 4 below presents overlapping area (OA) and related concurrent control enrollment after interim analysis (IA) of the study designs with power prior borrowing under the regional threshold. Generally, it is clearly observed that only about half of the patients need to be enrolled after the interim analysis. the OA is generally the largest and the concurrent control enrollment after the interim analysis is correspondingly the least when θ_{hc} is equal to θ_{cc}. The enrollment from the related simulated historical data needs slightly more patients enrolled.

It may cover multiple pages to present the estimations, related bias and mean square error (MSE) for different study designs under different historical data type, threshold type and effect size. Table 3 - 5 presents the estimations, bias and MSE of different study designs with different borrowing methods under all the related parameters θ_{hc}, θ_{cc} and $\theta_t - \theta_{cc}$ equal to 0.3. It clearly shows that all the borrowing methods are with quite close estimations to the parameter values for different historical data type and threshold type.

Table 3-5 Estimation summary of different methods at θ_{hc} = 0.3, θ_{cc} = 0.3 and effect size = 0.3

Scenario	method#	Trt.(95% CI)*	Cctrl.(95% CI)*	Eff.(95% CI)*	MSE	Bias
Under Observed historical data and global threshold	Full Borrowing	0.598(0.451, 0.744)	0.305(0.207, 0.403)	0.293(0.121, 0.464)	0.008	-0.007
	Power Prior	0.599(0.451, 0.744)	0.307(0.162, 0.458)	0.292(0.077, 0.496)	0.011	-0.008
	Frequentist	0.600(0.450, 0.750)	0.301(0.100, 0.500)	0.300(0.050, 0.550)	0.016	0.000
	No Borrowing	0.598(0.451, 0.744)	0.309(0.119, 0.500)	0.290(0.047, 0.528)	-0.010	0.016
	Commensurate Prior(K=1)	0.598(0.451, 0.744)	0.300(0.113, 0.488)	0.298(0.058, 0.533)	-0.002	0.015
	Commensurate Prior(K=50)	0.596(0.451, 0.744)	0.303(0.187, 0.420)	0.294(0.105, 0.472)	-0.006	0.009
	Commensurate Prior(K=100)	0.596(0.451, 0.744)	0.303(0.196, 0.412)	0.293(0.113, 0.466)	-0.007	0.009

Under Simulated historical data and global threshold	Full Borrowing	0.597(0.451, 0.744)	0.306(0.182, 0.451)	0.292(0.074, 0.488)	0.011	-0.008
	Power Prior	0.598(0.451, 0.744)	0.308(0.135, 0.491)	0.290(0.059, 0.513)	0.014	-0.010
	Frequentist	0.600(0.450, 0.750)	0.301(0.100, 0.500)	0.300(0.050, 0.550)	0.016	0.000
	No Borrowing	0.598(0.451, 0.744)	0.309(0.119, 0.500)	0.290(0.047, 0.528)	-0.010	0.016
	Commensurate Prior(K=1)	0.599(0.451, 0.744)	0.302(0.115, 0.491)	0.298(0.057, 0.537)	-0.002	0.015
	Commensurate Prior(K=50)	0.598(0.451, 0.744)	0.302(0.166, 0.443)	0.296(0.094, 0.493)	-0.004	0.011
	Commensurate Prior(K=100)	0.599(0.451, 0.744)	0.303(0.175, 0.452)	0.296(0.095, 0.499)	-0.004	0.011
Under Observed historical data and local threshold	Full Borrowing	0.598(0.451, 0.744)	0.305(0.207, 0.403)	0.293(0.121, 0.464)	0.008	-0.007
	Power Prior	0.599(0.451, 0.744)	0.307(0.162, 0.458)	0.292(0.077, 0.496)	0.011	-0.008
	Frequentist	0.600(0.450, 0.750)	0.301(0.100, 0.500)	0.300(0.050, 0.550)	0.016	0.000
	No Borrowing	0.598(0.451, 0.744)	0.309(0.119, 0.500)	0.290(0.047, 0.528)	-0.010	0.016
	Commensurate Prior(K=1)	0.598(0.451, 0.744)	0.300(0.113, 0.488)	0.298(0.058, 0.533)	-0.002	0.015
	Commensurate Prior(K=50)	0.596(0.451, 0.744)	0.304(0.187, 0.42)	0.292(0.105, 0.477)	-0.008	0.009
	Commensurate Prior(K=100)	0.597(0.451, 0.744)	0.304(0.196, 0.412)	0.294(0.115, 0.468)	-0.006	0.009
Under Simulated historical data and local threshold	Full Borrowing	0.597(0.451, 0.744)	0.306(0.182, 0.451)	0.292(0.074, 0.488)	0.011	-0.008
	Power Prior	0.598(0.451, 0.744)	0.308(0.135, 0.491)	0.290(0.059, 0.513)	0.014	-0.010
	Frequentist	0.600(0.450, 0.750)	0.301(0.100, 0.500)	0.300(0.050, 0.550)	0.016	0.000
	No Borrowing	0.598(0.451, 0.744)	0.309(0.119, 0.500)	0.290(0.047, 0.528)	-0.010	0.016
	Commensurate Prior(K=1)	0.594(0.451, 0.744)	0.299(0.113, 0.491)	0.295(0.059, 0.537)	-0.005	0.015
	Commensurate Prior(K=50)	0.597(0.451, 0.744)	0.302(0.166, 0.451)	0.295(0.077, 0.489)	-0.005	0.011
	Commensurate Prior(K=100)	0.599(0.451, 0.744)	0.303(0.175, 0.452)	0.296(0.095, 0.499)	-0.004	0.011
Under Observed historical data and regional threshold	Full Borrowing	0.597(0.451, 0.744)	0.305(0.207, 0.403)	0.292(0.121, 0.464)	-0.008	0.008
	Power Prior	0.598(0.451, 0.744)	0.307(0.162, 0.458)	0.29(0.072, 0.498)	-0.010	0.011
	Frequentist	0.600(0.450, 0.750)	0.301(0.100, 0.500)	0.300(0.050, 0.550)	0.016	0.000
	No Borrowing	0.598(0.451, 0.744)	0.309(0.119, 0.500)	0.290(0.047, 0.528)	-0.010	0.016
	Commensurate Prior(K=1)	0.597(0.451, 0.744)	0.302(0.113, 0.488)	0.294(0.056, 0.515)	-0.006	0.014
	Commensurate Prior(K=50)	0.598(0.451, 0.744)	0.304(0.187, 0.42)	0.295(0.104, 0.477)	-0.005	0.009
	Commensurate Prior(K=100)	0.598(0.451, 0.744)	0.303(0.196, 0.412)	0.295(0.116, 0.472)	-0.005	0.009
Under Simulated historical data and regional threshold	Full Borrowing	0.597(0.451, 0.744)	0.306(0.182, 0.451)	0.291(0.097, 0.488)	-0.009	0.011
	Power Prior	0.598(0.451, 0.744)	0.308(0.135, 0.490)	0.290(0.060, 0.514)	-0.010	0.013
	Frequentist	0.600(0.450, 0.750)	0.301(0.100, 0.500)	0.300(0.050, 0.550)	0.016	0.000
	No Borrowing	0.598(0.451, 0.744)	0.309(0.119, 0.500)	0.290(0.047, 0.528)	-0.010	0.016
	Commensurate Prior(K=1)	0.594(0.451, 0.744)	0.299(0.113, 0.491)	0.295(0.059, 0.537)	-0.005	0.015
	Commensurate Prior(K=50)	0.597(0.451, 0.744)	0.302(0.166, 0.451)	0.295(0.077, 0.489)	-0.005	0.011
	Commensurate Prior(K=100)	0.597(0.451, 0.744)	0.302(0.163, 0.452)	0.295(0.087, 0.491)	-0.005	0.011

#: Methods of No borrowing and Frequentist are not involved in historical data. Thus, the related estimations, bias and MSE are identical for each scenarios.
*: “Trt.”, “Cctrl.”, and “Eff.” represent the “treatment”, “Concurrent Control” and “Effect Size”. “CI” means credible interval obtained based on 2.5% and 97.5% quantile of the posterior distribution for Bayesian method, and confidence interval for Frequentist (Chi-square test) methods.

4. Discussion & Conclusion

In our research, we explored several different methods of incorporating historical control to concurrent control via Bayesian design. Power prior with interim analysis has been proposed and researched for a long time (Chen, Ibrahim, et al. 2000, Ibrahim and Chen 2000). Usually, the power parameter is fixed before the study based on the related expertise and knowledge. We propose that the data itself determines the power prior parameter at interim analysis via the OA of the posterior distributions of historical and concurrent control. It has the flexibility to adjust the power parameter between zero and one, which is correspondent to the methods of no and full borrowing. The proposed calculation method is straightforward. It is easy to interpret the adaptive design with power prior and the OA calculation to the study team. Moreover, there is no concerns of the bias of the posterior estimation due to the flexibility of adjustment. Under some scenarios (e.g., θ_{hc} and θ_{cc} are equal to 0.4, and $(\theta_t - \theta_{cc})$ is close to 0.3), the power of the study designs with power prior is quite similar with those from commensurate prior or full borrowing, and it has fewer expected sample size. They are the desired properties that power maintains high and sample size is smaller. The response rates [(θ_{cc} = 0.44, θ_t= 0.72, and θ_{t-cc}= 0.28] from the motivating study are just located in the “sweet spot”, and we recommend the adaptive design with power prior to the study team.

There are plenty of researches regarding commensurate prior (Hong, Fu, et al. 2018, Murray, Hobbs, et al. 2014). Although there are bias between historical controls and concurrent ones, commensurate prior essentially is hierarchical model, and the conditional distribution of concurrent control response rate given the historical control response rate is the measure of similarity between the "prior" from hierarchical model. The Gamma distribution that κ follows is equivalent to hyperprior of hierarchical model. In our research, we specify K equal to 1, 50 and 100 to evaluate the different performances of commensurate prior borrowing. The commensurate prior is close to the full borrowing method when K is equal to 50 or 100, and close to the no borrowing method when K is equal to 1. Similar with the power parameter from power prior, the input and adjustment from expertise and knowledge is necessary when specifying the K.

The methods of full borrowing, no borrowing and frequentist are served as the reference in our research. The full borrowing method is hard to be applied in the practice since it highly believes that the historical control is identical to concurrent control, which is difficult to persuade the researchers to accept it. The no borrowing method is not efficient, and it is served as reference as well. On the other hand, it is clearly observed that the performance similarity between the no borrowing and frequentist method.

Another factor we considered in the research is historical control date type (i.e. the historical data is simulated or observed). Both sources are possible and depend on the research process status, and we mimic the cases that could happen in the real world to assess the study comprehensively. Generally, it can be observed that the related power from the simulated historical control is slightly lower than that from observed ones, which is caused by the variation of the simulated data.

We also proposed three different types of thresholds (i.e. global, local and regional threshold). They reflect the different degrees of the researchers' belief in the similarity of the historical control and concurrent data. The power will decrease when global threshold is applied for the cases where there are obvious differences of historical and concurrent control data. However, type I error will increase largely when local threshold is applied and there are obvious differences of historical and concurrent control data. Regional threshold is optimal option between conservative and false positive result. The historical and concurrent control response rates are both located from 0.1 to 0.5 by 0.1. We did not explore the response rates less than 0.1 due to the unlikeness occurrence in practice, and the response rates greater than 0.5 since the results will be symmetric to those corresponding response rates less than 0.5 (i.e., one minus the response rate).

Our research focus is binary data, and the variance is associated with the response rate. It is worth researching other data types, especially the continuous ones that the variance is independent of location parameter. It should also be noted that there is no difference to specify the subject level or study level data for binary data if response rate and sample size are known. However, the methods may require different level data to conduct the borrowing. We mainly focus on small sample size. However, researchers can probably have larger data when designing a new related study (Liu 2018), the performance of those borrowing methods is worth being explored under a moderate or large sample size. Another limitation is that we did not consider the variety of the covariates. There are some proposed methods(Han, Zhan, et al. 2017), and it is a good future exploring.

To sum up, it is a good consideration to apply the power prior adaptive design with power parameter determination via overlapping area of posterior distribution under θ_{hc} and θ_{cc} close to 0.4, and effect size close to 0.3. Study design with commensurate prior is a general choice as well, however, appropriate priors need to be specified before study conducts.

Chapter 4: Subgroup identification of early preterm birth (ePTB): informing a future prospective enrichment clinical trial design

Zhang, C., Garrard, L., Keighley, J., Carlson, S. E., & Gajewski, B. J. (2017). Subgroup identification of early preterm birth (ePTB): informing a future prospective enrichment clinical trial design. *BMC Pregnancy and Childbirth, 17*, 18.

Abstract

Background: Despite the widely recognized association between the severity of early preterm birth (ePTB) and its related severe diseases, little is known about the potential risk factors of ePTB and the sub-population with high risk of ePTB. Moreover, motivated by a future confirmatory clinical trial to identify whether supplementing pregnant women with docosahexaenoic acid (DHA) has a different effect on the risk subgroup population or not in terms of ePTB prevalence, this study aims to identify potential risk subgroups and risk factors for ePTB, defined as babies born less than 34 weeks of gestation.

Methods: The analysis data (N = 3,994,872) were obtained from CDC and NCHS' 2014 Natality public data file. The sample was split into independent training and validation cohorts for model generation and model assessment, respectively. Logistic regression and CART models were used to examine potential ePTB risk predictors and their interactions, including mothers' age, nativity, race, Hispanic origin, marital status, education, pre-pregnancy smoking status, pre-pregnancy BMI, pre-pregnancy diabetes status, pre-pregnancy hypertension status, previous preterm birth status, infertility treatment usage status, fertility enhancing drug usage status, and delivery payment source.

Results: Both logistic regression models with either 14 or 10 ePTB risk factors produced the same C-index (0.646) based on the training cohort. The C-index of the logistic regression model based on 10 predictors was 0.645 for the validation cohort. Both C-indexes indicated a good discrimination and acceptable model fit. The CART model identified preterm birth history and race as the most important risk factors, and revealed that the subgroup with a preterm birth history and a race designation as Black had the highest risk for ePTB. The c-index and misclassification rate were 0.579 and 0.034 for the training cohort, and 0.578 and 0.034 for the validation cohort, respectively.

Conclusions: This study revealed 14 maternal characteristic variables that reliably identified risk for ePTB through either logistic regression model and/or a CART model. Moreover, both models efficiently identify risk subgroups for further enrichment clinical trial design.

Key Words: early preterm birth; risk factor; interaction; classification and regression tree; logistic regression; enrichment trial design

4.1 Background

Preterm birth, also known as premature birth, is the birth of a baby at less than 37 weeks of gestational age (Cdc.). Preterm birth occurs in 9.57% of all U.S. births each year (Hamilton, Martin, et al. 2015) . Worldwide, approximately 15 million babies are born prematurely each year (Who. 2018). Preterm birth increases the risk of many severe health outcomes. Infants born preterm are more likely to experience early death than are infants born at term (Blencowe, Cousens, et al. 2012, Catov, Bertolet, et al. 2014); and preterm birth is the leading cause of both neonatal death and long-term neurological disabilities for children in the United States (Cdc. , Witt, Cheng, et al. 2014). Moreover, adults who were born preterm are at increased risk of having hypertension (Keijzer-Veen, Dulger, et al. 2010, Norman 2010), mental health disorders, chronic respiratory disease, and neurologic and learning disabilities (Gravett and Rubens 2012). Preterm birth causes great social and medical burdens both in the U.S. (Mccormick 1985, Russell, Green, et al. 2007) and worldwide (Christopherson and Penrose 2010, Lawn, Gravett, et al. 2010, Treyvaud, Doyle, et al. 2011). Early preterm birth (ePTB)—birth at less than 34 weeks—has the highest risk of mortality and other diseases in adulthood (Creasy 1993, Martius, Steck, et al. 1998). The importance of prevention is evident for preterm birth, including ePTB. Consequently, to identify the risk factors of preterm birth, especially for ePTB, is a highly important step that will provide valuable information for subsequent enrichment clinical trial designs of targeted preventions and/or treatment.

Several recent studies have explored the risk factors for ePTB (Connealy, Carreno, et al. 2014, Gandhimadhi and Mythili 2010, Little, Janiak, et al. 2015, Saccone, Perriera, et al. 2015). Researchers have identified a few potential maternal risk factors associated with preterm birth including maternal hypertension (Norman 2010), Factor V Leiden (Hiltunen, Laivuori, et al. 2011), lower genital tract inflammatory milieu (Simhan, Bodnar, et al. 2011), prior preeclampsia (Connealy, Carreno, et al. 2014), and Crohn's disease (Stephansson, Larsson, et al. 2010). Not only were these trials limited in statistical power, few studies explored potential risk factors for ePTB, which has a higher risk for poor health outcomes (Martius, Steck, et al. 1998, Saigal and Doyle 2008). In addition, interaction among the risk factors was typically not considered, despite the important role played by the interaction among risk factors in the prevention and treatment of preterm birth, including ePTB. From a practical perspective, this analysis is motivated by a desire to inform a future confirmatory clinical trial designed to identify whether supplementing pregnant women with docosahexaenoic acid (DHA) can differently reduce the rate of ePTB for the subgroups. DHA supplementation provides a high yield, low risk provocative strategy to reduce ePTB delivery in the U.S. by up to 75% (Carlson, Colombo, et al. 2013). However, little is known regarding the effect profile of DHA on various populations; and it is possible for DHA to have different effects on different risk subgroups.

Based on findings from previous studies on preterm birth and our future research interest, the specific aim for this study is to identify potential risk subgroups and risk factors for the main outcome, ePTB, defined previously as babies born prior to 34 weeks of gestation (Creasy 1993, Neerhof, Cravello, et al. 1999). We applied and compared both logistic regression and classification and regression tree (CART) models to identify potential risk subgroups and risk factors from maternal demographic characteristics (Tan, Wen, et al. 2007, Witt, Cheng, et al. 2014) and maternal pre-pregnancy characteristics for ePTB. To the author's best knowledge, this is the first study to explore the association of ePTB with risk factors, the interactions among the risk factors, and to identify potential subgroups to inform future enrichment trial designs.

4.2 Method

4.2.1 2014 Natality Public Data File

The ePTB population data used for these analyses were obtained from the National Vital Statistics System's 2014 Natality public data file, compiled by the Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS). Since federal law mandates national collection and publication of births and other vital statistical data, all births occurring and registered within the U.S. in 2014 were collected directly from the 50 U.S. states, New York City, and the District of Columbia (DC) (Cdc 2014). The overall database contains 3,998,175 records comprised of demographic characteristics of the mother, father, and the child (e.g., gestation), maternal prenatal care, pregnancy history, and health data, etc. The public data and the corresponding user's guide are available from the website: http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm

4.2.2 Study Population

After excluding 3303 cases for which the gestation period from the original 2014 Natality public data file was unknown, the final analysis file for the current study included 3,994,872 records. Since the main outcome variable is ePTB, a binary flag variable representing the ePTB status (i.e., 1 = $<$ 34 Wks: ePTB and 0 = $\geq$ 34 Wks) was created in the analysis file. The analysis file included selected maternal demographic characteristics considered relevant to ePTB, such as mothers' age, mothers' nativity, mothers' race, mothers' Hispanic origin, marital status, mothers' education, delivery payment source. Delivery payment source was included as an additional covariate that may provide additional information on the implications of socioeconomic status for ePTB. Maternal pre-pregnancy characteristics and medical history were also included in the ePTB risk factor analysis. These factors included smoking status, body mass index (BMI), diabetes status, hypertension status, previous preterm birth status, infertility treatment usage status and fertility enhancing drug usage status. In total, 14 maternal variables from the database were used as risk predictors in statistical models. The father's demographic characteristics were not considered for this study.

A total of 142,851 (3.58%) observations from the analysis file contained at least one missing value for some of the predictors and those predictors were categorized as "missing." Predictors with responses of "Unknown," "Not Stated," "Not Applicable," and "Other," were categorized together as shown in the descriptive statistics listed in Table 4 - 1 & 4 - 2.

4.2.3 Statistical Analysis

Training and validation datasets. The large sample size allowed for independent training and validation cohorts. The overall sample was divided randomly into a training cohort (70%) and a validation cohort (30%), stratifying by ePTB status to ensure a balanced partition.

Descriptive statistics were summarized to compare the demographic and pre-pregnancy information between the two cohorts of data. The training sample was used to build models via both logistic regression and CART and the validation sample was used to evaluate the models obtained from the training cohort.

Logistic Regression. In order to investigate the association of ePTB with the potential risk factors, a multivariate logistic regression model was applied to estimate odds ratios (OR) and the corresponding 95% confidence intervals (CI). All predictors entered the model and they were selected via backward elimination. We set the significance level to stay in the model for a predictor to 0.05. A further simplified logistic regression model was fitted using 10 covariates to explore risk subgroups of ePTB. The predicted probabilities were calculated for the validation cohort based on the simplified model obtained from the training cohort. Based on the validation cohort, the calibration plot was generated to compare the average predicted probabilities and the average observed probabilities. The c-index was calculated to identify the model discriminatory capacity in terms of the training and validation cohorts.

CART model. CART model can be a very useful complement to a logistic regression model because the CART model can identify unknown interactions among the risk factors of ePTB. CART is a nonparametric method that derives hidden patterns in data by constructing a series of binary splits on the outcome of interest (Lei, Nollen, et al. 2015, Loh 2011, Nollen, Ahluwalia, et al. 2015). The most discriminating predictor is selected to form the first partition based on the ability of the variables to minimize the within-group variance of the dependent variable, so the observations within each subgroup share the same characteristics that influence the probability of belonging to the interested response group (Lemon, Roy, et al. 2003). This step is executed repeatedly to each partition until the sample size of each subgroup (i.e., a terminal node) is at or below a pre-specified level. In this study, the terminal node was specified as 0.5% of the total sample (either the training sample or the validation sample). A maximum tree first was constructed and standard pruning strategies were then applied to arrive at a parsimonious tree with a low misclassification rate and a high discriminatory capacity (Breiman, Friedman, et al. 1984). The final CART model can be visualized as an upside-down tree with the parent node of the tree containing the entire sample. Additional child nodes can be created using the Gini splitting rule for binary outcomes(Gordon 2013), and the terminal nodes are where predictions and inferences are made. The training cohort was used to generate an appropriate CART tree, and the validation cohort was utilized to evaluate the CART tree via the C-index and the misclassification rate.

All statistical tests were two-tailed with $p \leq 0.05$ as the statistically significant level. The CART analysis was executed in SAS Enterprise Miner Workstation 13.1 (Gordon 2013), and all other statistical analyses and the data management were conducted with SAS 9.4.

4.3 Results

4.3.1 Characteristics of the Study Population and Training and Validation Datasets

As previously mentioned, the analysis file included 3,994,872 records which contained 134,009 cases of ePTB (< 34 weeks) and 3,860,863 cases of baby birth ≥ 34 weeks of gestation. The characteristics of the subjects stratified by ePTB status are shown in Table 4 - 1. For the training and validation cohorts, 70% (N = 2,796,411) and 30% (N = 1,198,461) of the total sample were generated for each cohort, respectively. The frequencies and related percentages of each predictor were similar after the random split stratified by the ePTB status, indicating that the partition is well-balanced (Table 4 - 2).

Table 4-1 Subject demography information

Variable	Newborn Gestational Age	
	< 34 Wks: ePTB	≥ 34 Wks
	N = 134009	N = 3860863
Mothers' Age (%)		
≤ 24 Years	40711 (30.38)	1094793 (28.36)
25-29 Years	34831 (25.99)	1112643 (28.82)
30-34 Years	33578 (25.06)	1049775 (27.19)
≥ 35 Years	24889 (18.57)	603652 (15.64)
Mothers' Nativity (%)		
Born in U.S.	107578 (80.28)	2996531 (77.61)
Born Outside U.S. /Unknown/Not Stated	26431 (19.72)	864332 (22.39)
Mothers' Race (%)		
White	88185 (65.81)	2938466 (76.11)
Black	36554 (27.28)	603921 (15.64)
American Indian/Alaskan Native/Asian or Pacific Islander	9270 (6.92)	318476 (8.25)
Mothers' Hispanic Origin (%)		
Non-Hispanic/Hispanic Origin Not Stated	105011 (78.36)	2968422 (76.88)
Hispanic	28998 (21.64)	892441 (23.12)
Marital Status (%)		
Married	65594 (48.95)	2323620 (60.18)
Unmarried	68415 (51.05)	1537243 (39.82)
Mothers' Education (%)		
≤ High School or GED/Unknown	62819 (46.88)	1512489 (39.17)
Associate/Some College Credit	37338 (27.86)	1086153 (28.13)
≥ Bachelor's	29145 (21.75)	1124077 (29.11)
Missing	4707 (3.51)	138144 (3.58)
Pre-pregnancy Smoking Status (%)		
Nonsmoker	108663 (81.09)	3258557 (84.40)
Smoker/Unknown/Not Stated	20639 (15.40)	464162 (12.02)
Missing	4707 (3.51)	138144 (3.58)
Pre-pregnancy BMI (%)		
Under Weight-Normal ≤ 24.9	55824 (41.66)	1785913 (46.26)
Overweight 25.0-29.9	30288 (22.60)	918380 (23.79)

Obesity ≥ 30.0/Unknown/Not Stated	43190 (32.23)	1018426 (26.38)
Missing	4707 (3.51)	138144 (3.58)
Pre-pregnancy Diabetes Status (%)		
No/Unknown/Not Stated	126901 (94.70)	3694967 (95.70)
Yes	2401 (1.79)	27752 (0.72)
Missing	4707 (3.51)	138144 (3.58)
Pre-pregnancy Hypertension Status (%)		
No/Unknown/Not Stated	123932 (92.48)	3667289 (94.99)
Yes	5370 (4.01)	55430 (1.44)
Missing	4707 (3.51)	138144 (3.58)
Previous Preterm Birth Status (%)		
No/Unknown/Not Stated	118468 (88.40)	3626879 (93.94)
Yes	10834 (8.08)	95840 (2.48)
Missing	4707 (3.51)	138144 (3.58)
Infertility Treatment Usage Status (%)		
No/Unknown/Not Stated	122859 (91.68)	3669850 (95.05)
Yes	6443 (4.81)	52869 (1.37)
Missing	4707 (3.51)	138144 (3.58)
Fertility Enhancing Drug Usage Status (%)		
No/Not Applicable/Unknown/Not Stated	126582 (94.46)	3697856 (95.78)
Yes	2720 (2.03)	24863 (0.64)
Missing	4707 (3.51)	138144 (3.58)
Delivery Payment Source (%)		
Medicaid	65048 (48.54)	1598851 (41.41)
Private Insurance	51753 (38.62)	1771814 (45.89)
Self-pay/Other/Unknown	12501 (9.33)	352054 (9.12)
Missing	4707 (3.51)	138144 (3.58)

Table 4-2 Univariate difference between training sample and validation sample

Variables	Cohort	
	Training	Validation
	N = 2796411	N = 1198461
Mothers' Age (%)		
≤ 24 Years	794486 (28.41)	341018 (28.45)
25-29 Years	803113 (28.72)	344361 (28.73)

30-34 Years	758087 (27.11)	325266 (27.14)
≥ 35 Years	440725 (15.76)	187816 (15.67)
Mothers’ Nativity (%)		
Born in U.S.	2172903 (77.70)	931206 (77.70)
Born Outside U.S. /Unknown/Not Stated	623508 (22.30)	267255 (22.30)
Mothers’ Race (%)		
White	2119115 (75.78)	907536 (75.73)
Black	447972 (16.02)	192503 (16.06)
American Indian/Alaskan Native/Asian or Pacific Islander	229324 (8.20)	98422 (8.21)
Mothers’ Hispanic Origin (%)		
Non-Hispanic/Hispanic Origin Not Stated	2151766 (76.95)	921667 (76.90)
Hispanic	644645 (23.05)	276794 (23.10)
Marital Status (%)		
Married	1672583 (59.81)	716631 (59.80)
Unmarried	1123828 (40.19)	481830 (40.20)
Mothers’ Education (%)		
≤ High School or GED/Unknown	1102757 (39.43)	472551 (39.43)
Associate/Some College Credit	786618 (28.13)	336873 (28.11)
≥ Bachelor's	806822 (28.85)	346400 (28.90)
Missing	100214 (3.58)	42637 (3.56)
Pre-pregnancy Smoking Status (%)		
Nonsmoker	2357285 (84.30)	1009935 (84.27)
Smoker/Unknown/Not Stated	338912 (12.12)	145889 (12.17)
Missing	100214 (3.58)	42637 (3.56)
Pre-pregnancy BMI (%)		
Under Weight-Normal ≤ 24.9	1288811 (46.09)	552926 (46.14)
Overweight 25.0-29.9	664673 (23.77)	283995 (23.70)
Obesity ≥ 30.0/Unknown/Not Stated	742713 (26.56)	318903 (26.61)
Missing	100214 (3.58)	42637 (3.56)
Pre-pregnancy Diabetes Status (%)		
No/Unknown/Not Stated	2675048 (95.66)	1146820 (95.69)
Yes	21149 (0.76)	9004 (0.75)
Missing	100214 (3.58)	42637 (3.56)

Pre-pregnancy Hypertension Status (%)		
No/Unknown/Not Stated	2653410 (94.89)	1137811 (94.94)
Yes	42787 (1.53)	18013 (1.50)
Missing	100214 (3.58)	42637 (3.56)
Previous Preterm Birth Status (%)		
No/Unknown/Not Stated	2621496 (93.75)	1123851 (93.77)
Yes	74701 (2.67)	31973 (2.67)
Missing	100214 (3.58)	42637 (3.56)
Infertility Treatment Usage Status (%)		
No/Unknown/Not Stated	2654757 (94.93)	1137952 (94.95)
Yes	41440 (1.48)	17872 (1.49)
Missing	100214 (3.58)	42637 (3.56)
Fertility Enhancing Drug Usage Status (%)		
No/Not Applicable/Unknown/Not Stated	2676910 (95.73)	1147528 (95.75)
Yes	19287 (0.69)	8296 (0.69)
Missing	100214 (3.58)	42637 (3.56)
Delivery Payment Source (%)		
Medicaid	1164617 (41.65)	499282 (41.66)
Private Insurance	1276362 (45.64)	547205 (45.66)
Self-pay/Other/Unknown	255218 (9.13)	109337 (9.12)
Missing	100214 (3.58)	42637 (3.56)
Newborn Gestational Age (%)		
< 34 Wks: ePTB	93751 (3.35)	40258 (3.36)
≥ 34 Wks	2702660 (96.65)	1158203 (96.64)

4.3.2 Logistic Regression

14-Predictor model. Table 4 - 3 showed results from the logistic regression analysis for prevalence of ePTB with all 14 predictor variables. A relatively higher ePTB prevalence was observed in the older mother populations compared to younger mothers in the ≤ 24 years old reference group. The adjusted OR (95% CI) were 1.013 (0.995, 1.032), 1.130 (1.108, 1.152), and 1.354 (1.325, 1.385) for mothers in the age groups of 25-29 years (non-significant, p=0.169), 30-

34 years, and ≥ 35 years, respectively. Mothers born outside of the U.S. were less likely to experience ePTB compared to mothers born in the U.S. with an adjusted OR (95% CI) of 0.880 (0.863, 0.898). Black mothers and American Indian/Alaskan Native/Asian or Pacific Islander mothers were more likely to have an ePTB compared to White mothers with adjusted OR (95% CI) of 1.773 (1.743, 1.803) and 1.096 (1.066, 1.127), respectively. Mothers of Hispanic origin had a slightly higher ePTB prevalence compared to mothers of non-Hispanic origin with an adjusted OR (95% CI) of 1.033 (1.013, 1.053). ePTB was more likely to occur in the unmarried mother population compared to married mothers with an adjusted OR (95% CI) of 1.326 (1.304, 1.347).

Mothers with an associate degree or some college credit and mothers with a bachelor's degree or higher education were less likely to experience ePTB compared to mothers with a high school/general educational development (GED) or less education. The corresponding adjusted OR (95% CI) for each subgroup was 0.842 (0.828, 0.856) and 0.713 (0.698, 0.729), respectively. Results from the subgroup with missing mother's education were non-significant (p=0.873). In addition, since all the observations with missing predictors were all from the same subset, for the following parameters after mothers' education, missing observations were automatically excluded from the analysis, and the corresponding parameters were automatically set to 0 due to they are from the same subset.

Some maternal pre-pregnancy characteristics and medical history factors were also found to be related to ePTB. For Pre-pregnancy BMI, mothers in the overweight subgroup had a slightly lower prevalence of ePTB (p=0.047), with an adjusted OR (95% CI) of 0.983 (0.966, 1.000) compared to mothers with underweight and/or normal BMI. However, the opposite result was obtained for the obese subgroup with an adjusted OR (95% CI) of 1.127 (1.109, 1.145), compared with the underweight and/or normal BMI mothers. For other pre-pregnancy risk factors (i.e., smoking status, diabetes status, hypertension status, and previous preterm birth status), mothers in each risk sub-category were more likely to have a higher prevalence of ePTB compared to mothers who did not have the abovementioned risk factors. The corresponding adjusted OR (95% CI) were 1.183 (1.160, 1.206), 1.776 (1.685, 1.871), 1.984 (1.913, 2.056), 3.004 (2.929, 3.081), respectively.

In addition, mothers who used infertility treatment were much more likely to experience ePTB than those who had not used the infertility treatment, with an adjusted OR (95% CI) of 5.103 (4.888, 5.328). On the other hand, a different outcome was observed with the usage of fertility enhancing drug. Mothers who used fertility enhancing drugs were less likely to have an ePTB compared to women who did not, with an adjusted OR (95% CI) of 0.820 (0.769, 0.873). Compared to women whose payer was Medicaid, the adjusted OR (95% CI) were 0.965 (0.948, 0.983) and 1.079 (1.054, 1.105) for women who had private insurance and self-pay, respectively. Mothers with private insurance had a slightly lower prevalence of ePTB; whereas mothers with self-paid delivery had a slightly higher prevalence of ePTB. Although the p-values for both comparisons were statistically significant (< 0.0001), the numerical differences were small.

Table 4-3. The estimate and adjusted OR of logistic regression analysis on the training cohort

Parameter	Estimat e	Adjusted OR (95% CI)	*P* value
Intercept	-3.7154	-	<.0001
Mothers' Age (%)			
≤ 24 Years	-	1.0 (1.0–1.0)	-
25-29 Years	0.0129	1.013 (0.995, 1.032)	0.169
30-34 Years	0.1221	1.130 (1.108, 1.152)	<.0001
≥ 35 Years	0.3034	1.354 (1.325, 1.385)	<.0001

Mothers’ Nativity (%)			
Born in U.S.	-	1.0 (1.0–1.0)	-
Born Outside U.S. /Unknown/Not Stated	-0.1274	0.880 (0.863, 0.898)	<.0001
Mothers’ Race (%)			
White	-	1.0 (1.0–1.0)	-
Black	0.5727	1.773 (1.743, 1.803)	<.0001
American Indian/Alaskan Native/Asian or Pacific Islander	0.0917	1.096 (1.066, 1.127)	<.0001
Mothers’ Hispanic Origin (%)			
Non-Hispanic/Hispanic Origin Not Stated	-	1.0 (1.0–1.0)	-
Hispanic	0.0323	1.033 (1.013, 1.053)	0.009
Marital Status (%)			
Married	-	1.0 (1.0–1.0)	-
Unmarried	0.2819	1.326 (1.304, 1.347)	<.0001
Mothers’ Education (%)			
≤ High School or GED/Unknown	-	1.0 (1.0–1.0)	-
Associate/Some College Credit	-0.1725	0.842 (0.828, 0.856)	<.0001
≥ Bachelor's	-0.3382	0.713 (0.698, 0.729)	<.0001
Missing	0.0031	1.003 (0.966, 1.042)	0.8727
Pre-pregnancy Smoking Status (%) [a]			
Nonsmoker	-	1.0 (1.0–1.0)	-
Smoker/Unknown/Not Stated	0.1677	1.183 (1.160, 1.206)	<.0001
Pre-pregnancy BMI (%) [a]			
Under Weight-Normal ≤24.9	-	1.0 (1.0–1.0)	-
Overweight 25.0-29.9	-0.0174	0.983 (0.966, 1.000)	0.0472
Obesity ≥30.0/Unknown/Not Stated	0.1195	1.127 (1.109, 1.145)	<.0001
Pre-pregnancy Diabetes Status (%) [a]			
No/Unknown/Not Stated	-	1.0 (1.0–1.0)	-
Yes	0.5741	1.776 (1.685, 1.871)	<.0001
Pre-pregnancy Hypertension Status (%) [a]			
No/Unknown/Not Stated	-	1.0 (1.0–1.0)	
Yes	0.6849	1.984 (1.913, 2.056)	<.0001
Previous Preterm Birth Status (%) [a]			
No/Unknown/Not Stated	-	1.0 (1.0–1.0)	-
Yes	1.0999	3.004 (2.929, 3.081)	<.0001

Infertility Treatment Usage Status (%) [a]			
No/Unknown/Not Stated	-	1.0 (1.0–1.0)	-
Yes	1.6299	5.103 (4.888, 5.328)	<.0001
Fertility Enhancing Drug Usage Status (%) [a]			
No/Not Applicable/Unknown/Not Stated	-	1.0 (1.0–1.0)	-
Yes	-0.1988	0.820 (0.769, 0.873)	<.0001
Delivery Payment Source (%) [a]			
Medicaid	-	1.0 (1.0–1.0)	-
Private Insurance	-0.0352	0.965 (0.948, 0.983)	<.0001
Self-pay/Other/Unknown	0.0762	1.079 (1.054, 1.105)	<.0001

[a]: For the following parameters after mothers' education, missing observations were automatically excluded from the analysis, and the corresponding parameters were automatically set to 0 due to they are from the same subset.

10-Predictor model. After examining results from the 14-predictor model, four covariates - mothers' nativity, mothers' Hispanic origin, fertility enhancing drug usage status, and delivery payment source - were excluded for having minimal effects on ePTB and to explore further a smaller set of potential risk subgroups for ePTB. Moreover, the same C-index (0.646) was obtained from both logistic regression models with either 14 or 10 predictors based on the training cohort (Figure 4-1). The C-index was 0.645 after fitting the 10-predictor model on the validation data, indicating an acceptable model fit. Figure 4-2 showed the calibration plot based on the validation cohort to compare the average predicted probabilities and the average observed probabilities across quartiles. The average and range of both predicted and observed probability for each of the four potential subgroups were shown in Table 4 - 4, along with summarized maternal characteristics for each subgroup from the validation cohort.

For the first subgroup (i.e., first quartile), the average predicted and observed probabilities were 1.92% and 1.83% respectively, with a range of 0.55% for the predicted probability. A typical mother from this potential subgroup was between 30-34 years old, with a designation as white, married, with a bachelor's degree or higher education level, non-smoking, underweight to normal weight (BMI ≤24.9) before pregnancy, without notable pre-pregnancy risk factors (i.e., diabetes, hypertension, previous preterm birth), and without infertility treatment. The second subgroup (i.e., second quartile) had an average predicted and an average observed probability of 2.46% and 2.33% respectively, with a range of 0.52% for the predicted probability. Mothers from the second potential subgroup shared very similar characteristics with a typical mother from the first subgroup, with the exception of age (slightly younger, 25-29 years old) and slightly lower education level (associate degree or some college credit). The average and range of predicted probability for the third subgroup (i.e., third quartile) were 3.22% and 0.95%; and the observed probability was 3.24%. Similar to trends observed from the second subgroup (in comparison with the first subgroup), a typical mother from the third subgroup was younger (≤ 24 years old) and with less education (≤ high school or GED/unknown). Lastly, the average predicted and observed probabilities for the highest risk subgroup (i.e., last 25% of data) were 6.02% and 6.07% respectively, with the predicted probability range of 60.6%. Mothers in this high-risk subgroup exhibit much different characteristics from the other three subgroups. They tended to be younger (≤ 24 years old), Black, unmarried, with a high school/GED or less education level, and generally obese (≥ 30.0 BMI). Moreover, compared to the other three subgroups, a relatively higher percentage of mothers in this high-risk subgroup had pre-pregnancy diabetes, hypertension, previous preterm birth, and infertility treatment usage.

Table 4-4 The ePTB subgroup predicted /observed probability and maternal characteristics in validation cohort via logistic regression

Variable	Subgroup			
	1st Quartile	2nd Quartile	3rd Quartile	4th Quartile
	N = 299529	N = 299078	N = 299993	N = 299861
Probability (%)				
Average Predicted	1.92	2.46	3.22	6.02
Range Predicted	0.55	0.52	0.95	60.6
Average Observed	1.83	2.33	3.24	6.07
Mothers' Age (%)				
≤ 24 Years	36603 (12.22)	70681 (23.63)	127739 (42.58)	105995 (35.35)
25-29 Years	120779 (40.32)	83600 (27.95)	68003 (22.67)	71979 (24.00)
30-34 Years	129538 (43.25)	78439 (26.23)	56362 (18.79)	60927 (20.32)
≥ 35 Years	12609 (4.21)	66358 (22.19)	47889 (15.96)	60960 (20.33)
Mothers' Race (%)				
White	259978 (86.80)	273311 (91.38)	260128 (86.71)	114119 (38.06)
Black	0 (0.00)	872 (0.29)	18661 (6.22)	172970 (57.68)
American Indian/Alaskan Native/Asian or Pacific Islander	39551 (13.20)	24895 (8.32)	21204 (7.07)	12772 (4.26)
Marital Status (%)				
Married	296804 (99.09)	246717 (82.49)	92320 (30.77)	80790 (26.94)
Unmarried	2725 (0.91)	52361 (17.51)	207673 (69.23)	219071 (73.06)
Mothers' Education (%)				
≤ High School or GED/Unknown	10988 (3.67)	93778 (31.36)	192086 (64.03)	175699 (58.59)
Associate/Some College Credit	69843 (23.32)	117843 (39.40)	69455 (23.15)	79732 (26.59)
≥ Bachelor's	217614 (72.65)	71541 (23.92)	21886 (7.30)	35359 (11.79)
Missing	1084 (0.36)	15916 (5.32)	16566 (5.52)	9071 (3.03)
Pre-pregnancy Smoking Status (%)				
Nonsmoker	295313 (98.59)	262159 (87.66)	234907 (78.30)	217556 (72.55)
Smoker/Unknown/Not Stated	3132 (1.05)	21003 (7.02)	48520 (16.17)	73234 (24.42)
Missing	1084 (0.36)	15916 (5.32)	16566 (5.52)	9071 (3.03)
Pre-pregnancy BMI (%)				
Under Weight-Normal ≤ 24.9	183032 (61.11)	142007 (47.48)	119757 (39.92)	108130 (36.06)
Overweight 25.0-29.9	82956 (27.70)	67818 (22.68)	70451 (23.48)	62770 (20.93)
Obesity ≥ 30.0/Unknown/Not Stated	32457 (10.84)	73337 (24.52)	93219 (31.07)	119890 (39.98)
Missing	1084 (0.36)	15916 (5.32)	16566 (5.52)	9071 (3.03)

Pre-pregnancy Diabetes Status (%)				
No/Unknown/Not Stated	298445 (99.64)	283149 (94.67)	282480 (94.16)	282746 (94.29)
Yes	0 (0.00)	13 (0.00)	947 (0.32)	8044 (2.68)
Missing	1084 (0.36)	15916 (5.32)	16566 (5.52)	9071 (3.03)
Pre-pregnancy Hypertension Status (%)				
No/Unknown/Not Stated	298445 (99.64)	283162 (94.68)	282293 (94.10)	273911 (91.35)
Yes	0 (0.00)	0 (0.00)	1134 (0.38)	16879 (5.63)
Missing	1084 (0.36)	15916 (5.32)	16566 (5.52)	9071 (3.03)
Previous Preterm Birth Status (%)				
No/Unknown/Not Stated	298445 (99.64)	283162 (94.68)	283427 (94.48)	258817 (86.31)
Yes	0 (0.00)	0 (0.00)	0 (0.00)	31973 (10.66)
Missing	1084 (0.36)	15916 (5.32)	16566 (5.52)	9071 (3.03)
Infertility Treatment Usage Status (%)				
No/Unknown/Not Stated	298445 (99.64)	283162 (94.68)	283427 (94.48)	272918 (91.01)
Yes	0 (0.00)	0 (0.00)	0 (0.00)	17872 (5.96)
Missing	1084 (0.36)	15916 (5.32)	16566 (5.52)	9071 (3.03)

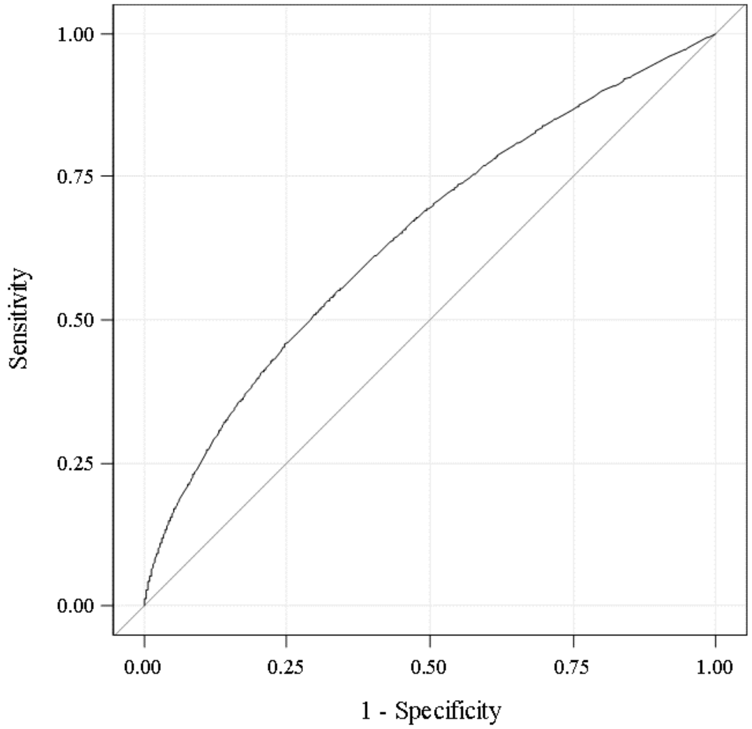


Figure 4-1 ROC curve from logistic regression on the training dataset

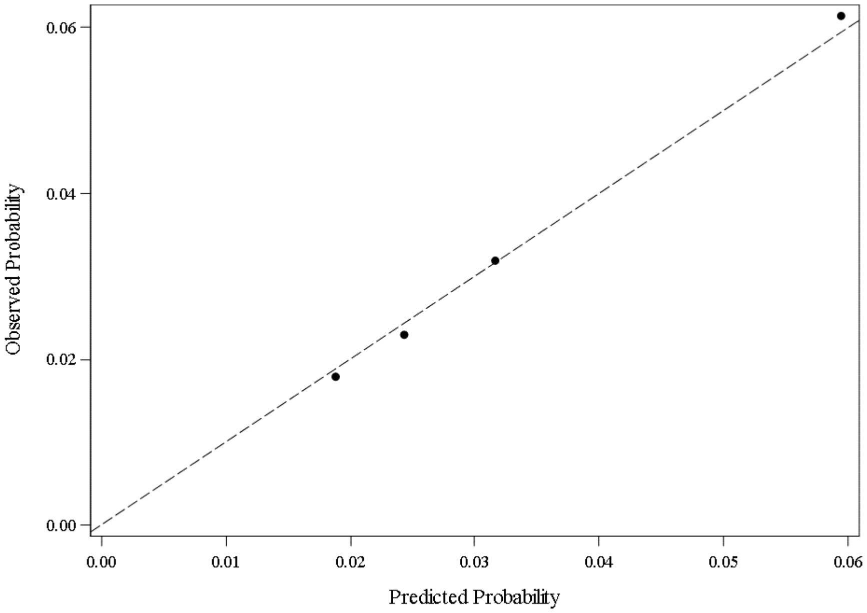


Figure 4-2 Calibration plot from the validation sample. Observed vs. Predicted Probability across the quartiles

4.3.3 CART model

For the CART model, sub-categories were collapsed for a couple of risk factors. The missing subgroup of previous preterm birth status was combined with the “no” group; and the race category of American Indian/Alaskan Native/Asian or Pacific Islander was combined with the White group. Based on a pre-specified stopping rule of having the terminal node size no less than 0.5% of the total sample and the binary Gini splitting rule, the CART tree was created to explore the unknown interactions among the risk factors and identify potential risk subgroups (Figure 4-3). Overall, the CART model from the training cohort produced a misclassification rate of 0.034 and a C-index of 0.579. Moreover, the misclassification rate was 0.034 and the c-index was 0.578 from the validation cohort. By the percentage representing the observed prevalence of ePTB, CART identified four subgroups. Previous preterm birth status was identified as the most discriminating predictor for ePTB, followed by mothers' race.

From training cohort, 14.41% of mothers with a preterm birth history and a race designation as Black had an ePTB experience (n =16,750), indicating a higher risk of ePTB for Black mothers with a preterm birth history. The correspondent percentage of this subgroup from the validation cohort is 15.02% (n=7,085). This subgroup totally accounted for 0.60% of the overall 2014 U.S. births. 8.96% and 8.70% of mothers with a preterm birth history and a race designation as White had an ePTB experience from training (n = 57,951) and validation (n = 24,888), and the subgroup birth prevalence (SBP) was 2.07%. Women without a preterm birth history who were Black had an ePTB experience of 5.37% (n = 431,222); while 2.75% of mothers without a preterm birth history who were White had an ePTB experience (n = 2,290,488). The correspondent rates for the identical subgroups from the validation cohort are 5.35% (n =185,418) and 2.76% (n = 981,070). These two subgroups accounted for 15.44% and 81.89% of the overall birth data, respectively.

It is also informative to interpret the CART tree in terms of risk factors that increase or decrease the probability of ePTB. One can compare the rates of ePTB among the four potential subgroups to the average rate of ePTB of the total sample (3.35%, 3.36% for training and validation cohort, respectively). Three subgroups (with preterm birth history and Black, with preterm birth history and White, without preterm birth history and Black) had an increased probability of ePTB compared to the subgroup without a preterm birth history who were White.

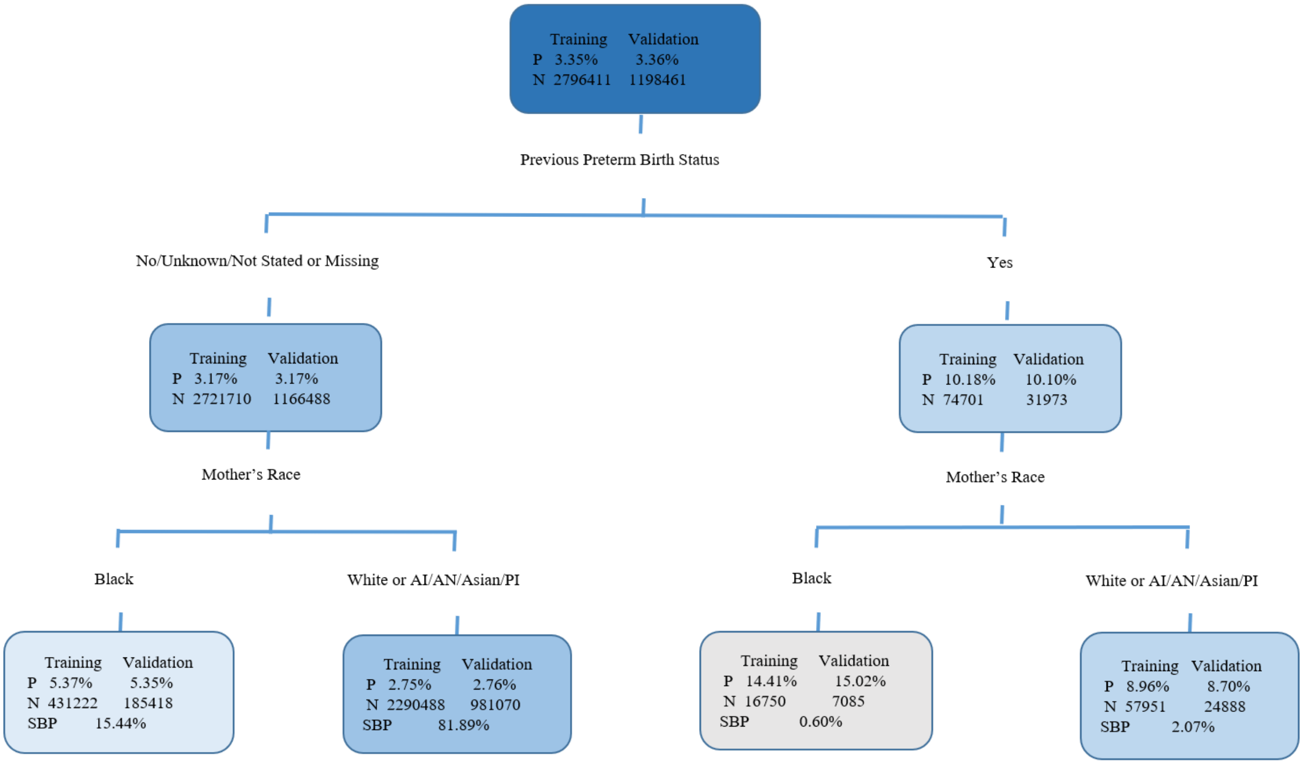


Figure 4-3 Classification and Regression Tree model for predicting ePTB

The probability of ePTB (P) and the number of subject (N) are all given inside of each node for both training and validation cohort. In each end node, the subgroup birth prevalence (SBP) is also calculated. AI = American Indian; AN = Alaskan Native; PI = Pacific Islander.

4.4 Discussion

This large sampled pioneer study aimed to explore potential risk factors and their interactions, and identify subgroup for the ePTB population via both logistic regression model and the CART model. Several important findings emerged from the current study. First, a subset of the most important and relevant covariates have been identified among the 14 risk factors examined, such as race, diabetes history, hypertension history, preterm birth history, and infertility treatment usage. Second, although logistic regression model identified a set of 10 predictors for the prevalence of ePTB, the CART model was able to examine multiple and complicated interactions among the selected predictors. The CART model clearly identified that the subgroup with a preterm birth history and a race designation as Black had the highest risk for ePTB. Third, although not presented in the current work, the risk ratios (RR) of a particular subgroup from the CART terminal nodes can be calculated to compare with the RR of other subgroups via the observed probabilities. RR also indirectly can inform the risk factors for ePTB.

Previous preterm birth status and race were the most discriminating predictors for ePTB by the CART model, while another eight predictors were identified by the logistic regression analyses. As a well-known traditional statistical approach, logistic regression provided predicted probabilities based on the important demographics and characteristics for ePTB; however, it cannot identify complicated interactions among risk factors. On the other hand, the CART model presents a more straightforward picture of the potential high risk subgroups for ePTB for whom targeted prevention efforts can be implemented. Moreover, each subgroup accounted for a different percent of the overall simple size. Thus the difference in ePTB prevalence among the four subgroups identified by the CART model was much larger than that identified by the logistic regression model. Coupling both statistical approaches provides more efficiency for analyzing the overall objective of this study. It also further exemplifies the statistical analysis for similar studies.

Additionally, from a long-term perspective, this pioneering study provides valuable information and direction for our further targeted subgroup enrichment clinical trials aiming at decreasing the prevalence of ePTB among the interactive risk subgroups via supplement pregnant women with DHA.

There are some limitations with this study. Some risk factors contained missing values and/or values of “Not Applicable”, “Unknown,” and “Not Stated,” which added complexity to the proposed analyses. However, data management is unavoidable for any concrete project, and we face the same issue for such a large database regarding birth data for the whole country. The solution taken was from an objective and general perspective, which could deduce the reasonable and acceptable results. Additionally, the risk predictors explored in this paper mainly from mothers’ demographics factors and Maternal pre-pregnancy characteristics, and it does include more highly specific biomarkers. This is due to no such predictors collected in the analysis database. Potentially, this limitation may lead to the relatively low c-index for both models. Further application and reference for these two models should be precautioned.

4.5 Conclusions

This study revealed 14 maternal characteristic variables that can be used reliably to identify risk factor subgroups for ePTB either through a logistic regression model and/or a CART model. Moreover, both models may be used efficiently to identify high risk subgroups for further enrichment clinical trial design.

4.6 List of abbreviations

BMI – body mass index

CART - classification and regression tree

CDC- Centers for Disease Control and Prevention’s

CI - confidence intervals

DC - District of Columbia

DHA - docosahexaenoic acid

ePTB - early preterm birth

GED - general educational development

NCHS - National Center for Health Statistics

OR - odds ratios

RR - risk ratios

SBP - subgroup birth prevalence

4.7 Availability of data and materials

The dataset supporting the conclusions of this article is available in the Centers for Disease Control and Prevention repository: http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm

4.8 Competing interests

The authors declare that they have no competing interests. Moreover, this manuscript reflects the views of the authors and should not be construed to represent the FDA's views or policies. Lili Garrard completed this work as a PhD student in the Department of Biostatistics at the University of Kansas Medical Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

4.9 Funding

This study was supported in part by a National Institutes of Health (NIH) Clinical and Translational Science Award Grant (UL1 TR000001, formerly UL1RR033179), awarded to the University of Kansas Medical Center.

Chapter 5: Summary and Future Directions

In Chapter Two, we investigate batch of adaptive designs which are composed by analysis models (pairwise independent, hierarchical, and cluster hierarchical achieved via Dirichlet Process (DP)), interim analysis (Yes vs. No) and longitudinal data modeling (Yes vs. No). We found that the hierarchical model with interim analysis and longitudinal modelling is an optimal approach to identifying subgroup treatment effects, and the cluster hierarchical model with interim analysis and longitudinal imputation is an excellent alternative approach in cases where sufficient information is not available for specifying the related priors. There are several points that is worth exploring in the future. Firstly, our research is based on continues data, and it is interesting to validate that these findings can be applied to the with discrete or time-to-event endpoints. Secondly, there is only one interim analysis and randomization ratio is fixed in our research, however, it is good to explore Bayesian response adaptive randomization (Bayesian RAR) to update the randomization ratio based on each interim analysis result when no indication of effective treatment arms. Other factors, such as treatment dosage, sample size, etc. may also be adjusted accordingly under Bayesian adaptive designs. Lastly, we assume the missing data pattern is missing at random (MAR). Meanwhile, it is an interesting topic for future research to explore the different imputation methods for other mechanism, like missing not at random (MNAR).

In Chapter Three, we investigate several Bayesian designs incorporating historical control borrowing: power prior via overlapping area, commensurate prior, and some other methods. The impact of historical data type and different types of the threshold used in Bayesian decision rule are also explored. We found that it is a good consideration to apply the power prior adaptive design with power parameter determination via overlapping area of posterior distribution under certain values of true response rates of concurrent control, historical control, and treatment effect. Study design with commensurate prior is an admissible choice as well, however, appropriate priors need to be specified. Still, there are several points that is worth exploring in the future. Firstly, the commensurate prior is incorporated in the adaptive scenarios, is it possible to connect the interim analysis result to the commensurate prior parameter setting? If yes, then compare it with the designs incorporated power prior via overlapping area. Secondly, it is data type. Data type in our research is binary, and the summary data level is equivalent to the subject data level. Moreover, the variance of binary data is associated with the response rate. It is worth researching other data type, especially the continuous ones that the variance is independent of location parameter. Thirdly, we mainly focus on small sample size. However, researchers can probably have larger data when designing a new related study, the performance of those borrowing methods should be explored under a moderate or large sample size. Lastly, we did not consider the variety of the covariates. It is a good future exploring how to adjust the difference between the concurrent and historical controls via the different covariates.

In Chapter Four, we logistic regression and CART models to identify the risk factors of ePTB from maternal perspective based on the birth data from CDC and NCHS' 2014 Natality public file. We identify that the subgroup with a preterm birth history and a race designation as Black had the highest risk for ePTB. Those findings can provide valuable information for a future enrichment trial design. Moreover, both models can be applied to identify risk factors for other studies.

References

Paving the way for personalized medicine: FDA's role in a new era of medical product development. *https://wwwfdanewscom/ext/resources/files/10/10-28-13-Personalized-Medicinepdf*.

Almirall D, Compton SN, Gunlicks - Stoessel M, Duan N, Murphy SA. (2012) Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. *Statistics in medicine*, 31(17):1887-1902.

Alosh M, Fritsch K, Huque M, Mahjoob K, Pennello G, Rothmann M, Russek-Cohen E, Smith F, Wilson S, Yue L. (2015) Statistical considerations on subgroup analysis in clinical trials. *Statistics in Biopharmaceutical Research*, 7(4):286-303.

Alosh M, Huque MF, Bretz F, D'Agostino Sr RB. (2017) Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. *Statistics in medicine*, 36(8):1334-1360.

Altstein LL, Li G, Elashoff RM. (2011) A method to estimate treatment efficacy among latent subgroups of a randomized clinical trial. *Statistics in medicine*, 30(7):709-717.

Austin PC. (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399-424.

Barohn R, Gajewski B, Pasnoor M, Brown L, Herbelin L, Kimminau K, Jawdat O, Liu T, Parks C, Shlemon P. (2018) *Patient Assisted Intervention for Neuropathy: Comparison of treatment in real life situations (PAIN-CONTRoLS)(P1. 435)*. In.: AAN Enterprises.

Bayman EÖ, Chaloner K, Cowles MK. (2010) Detecting qualitative interaction: a Bayesian approach. *Statistics in medicine*, 29(4):455-463.

Berry SM, Broglio KR, Groshen S, Berry DA. (2013) Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials. *Clinical Trials*, 10(5):720-734.

Bhuyan P, Chen C, Desai J, Hart C, Helvering L, Jia D, Lim G, Lu J. (2015) DEVELOPMENT AND IMPLEMENTATION OF A PHARMA-COLLABORATIVE LARGE HISTORICAL CONTROL DATABASE. *WHITE PAPER Available at http://wwwtranscelleratebiopharmainccom/wp-content/uploads/2015/04/TransCelerate-PSoC-Data-Sharing-White-Paperpdf.*

Blencowe H, Cousens S, Oestergaard MZ, Chou D, Moller AB, Narwal R, Adler A, Vera Garcia C, Rohde S, Say L *et al*. (2012) National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *Lancet*, 379(9832):2162-2172.

Breiman L, Friedman J, Olshen R, Stone C. (1984) *Classification and regression trees. Monterey, Calif., USA: Wadsworth*. In.: Inc.

Carlson SE, Colombo J, Gajewski BJ, Gustafson KM, Mundy D, Yeast J, Georgieff MK, Markley LA, Kerling EH, Shaddy DJ. (2013) DHA supplementation and pregnancy outcomes. *The American journal of clinical nutrition*, 97(4):808-815.

Catov JM, Bertolet M, Chen YF, Evans RW, Hubel CA. (2014) Nonesterified fatty acids and spontaneous preterm birth: a factor analysis for identification of risk patterns. *Am J Epidemiol*, 179(10):1208-1215.

CDC. (2014) User Guide to the 2014 NatalityPublic Use File.

CDC. Preterm Birth. https://www.cdc.gov/Reproductivehealth/Maternalinfanthealth/Pretermbirth.Htm.

Chen M-H, Ibrahim JG, Shao Q-M. (2000) Power prior distributions for generalized linear models. *Journal of Statistical Planning Inference*, 84(1-2):121-137.

Christopherson M, Penrose C. (2010) 1240 An Evaluation of the Burden of Premature Birth on a United Kingdom Regional Paediatric Intensive Care Service. *Pediatric Research*, 68:614-615.

Clarke M, Loudon K. (2011) Effects on patients of their healthcare practitioner's or institution's participation in clinical trials: a systematic review. *Trials*, 12(1):16.

Connealy BD, Carreno CA, Kase BA, Hart LA, Blackwell SC, Sibai BM. (2014) A history of prior preeclampsia as a risk factor for preterm birth. *Am J Perinatol*, 31(6):483-488.

Creasy RK. (1993) Preterm birth prevention: where are we? *Am J Obstet Gynecol*, 168(4):1223-1230.

Dmitrienko A, Muysers C, Fritsch A, Lipkovich I. (2016) General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. *Journal of biopharmaceutical statistics*, 26(1):71-98.

Fabian CJ, Nye L, Powers KR, Nydegger JL, Kreutzjans AL, Phillips TA, Metheny T, Winblad O, Zalles CM, Hagan CR. (2019) Effect of Bazedoxifene and Conjugated Estrogen (Duavee) on Breast Cancer Risk Biomarkers in High-Risk Women: A Pilot Study. *Cancer Prevention Research*, 12(10):711-720.

FACTS. (2018) FACTS Adaptive Indication and Population Finder (AIPF) Design Engine Specification.

FDA. (2018) Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry.

FDA. (2012) Guidance for Industry Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products.

FDA. (2019) Rare Diseases: Common Issues in Drug Development Guidance for Industry.

FDA. (2019) Rare Diseases: Natural History Studies for Drug Development Guidance for Industry.

Foster JC, Taylor JM, Ruberg SJ. (2011) Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867-2880.

Gajewski BJ, Berry SM, Barsan WG, Silbergleit R, Meurer WJ, Martin R, Rockswold GL. (2016) Hyperbaric oxygen brain injury treatment (HOBIT) trial: a multifactor design with response adaptive randomization and longitudinal modeling. *Pharmaceutical statistics*, 15(5):396-404.

Gajewski BJ, Berry SM, Quintana M, Pasnoor M, Dimachkie M, Herbelin L, Barohn R. (2015) Building efficient comparative effectiveness trials through adaptive designs, utility functions, and accrual rate optimization: finding the sweet spot. *Statistics in medicine*, 34(7):1134-1149.

Gamalo-Siebers M, Savic J, Basu C, Zhao X, Gopalakrishnan M, Gao A, Song G, Baygani S, Thompson L, Xia HA *et al*. (2017) Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation. *Pharm Stat*, 16(4):232-249.

Gamalo-Siebers M, Tiwari R, LaVange L. (2016) Flexible shrinkage estimation of subgroup effects through Dirichlet process priors. *Journal of biopharmaceutical statistics*, 26(6):1040-1055.

Gandhimadhi D, Mythili R. (2010) Periodontal infection as a risk factor for preterm low birth weight. *J Indian Soc Periodontol*, 14(2):114-120.

Gelman A. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3):515-534.

Ghadessi M, Tang R, Zhou J, Liu R, Wang C, Toyoizumi K, Mei C, Zhang L, Deng C, Beckman RA. (2020) A roadmap to using historical controls in clinical trials–by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). *Orphanet Journal of Rare Diseases*, 15(1):1-19.

Gliklich RE, Dreyer NA, Leavy MB. (2014) *Registries for evaluating patient outcomes: a user's guide*: Government Printing Office.

Gordon L. (2013) Using Classification and Regression Trees (CART) in SAS® Enterprise Miner TM For Applications in Public Health. *2013*.

Gravestock I, Held L. (2018) Power priors based on multiple historical studies for binary outcomes. *Biometrical Journal*.

Gravett MG, Rubens CE. (2012) A framework for strategic investments in research to reduce the global burden of preterm birth. *Am J Obstet Gynecol*, 207(5):368-373.

Grieve AP. (2016) Idle thoughts of a ‘well‐calibrated’Bayesian in clinical drug development. *Pharmaceutical statistics*, 15(2):96-108.

Groft SC. (2010) Rare diseases-avoiding misperceptions and establishing realities: the need for reliable epidemiological data. *Advances in experimental medicine biology*, 686:3-14.

Gsteiger S, Neuenschwander B, Mercier F, Schmidli H. (2013) Using historical control information for the design and analysis of clinical trials with overdispersed count data. *Statistics in Medicine*, 32(21):3609-3622.

Hamilton BE, Martin JA, Osterman M, Curtin S, Matthews T. (2015) Births: Final Data for 2014. *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 64(12):1-64.

Han B, Zhan J, John Zhong Z, Liu D, Lindborg S. (2017) Covariate‐adjusted borrowing of historical control data in randomized clinical trials. *Pharmaceutical statistics,* 16(4):296-308.

Hiltunen LM, Laivuori H, Rautanen A, Kaaja R, Kere J, Krusius T, Rasi V, Paunio M. (2011) Factor V Leiden as a risk factor for preterm birth--a population-based nested case-control study. *J Thromb Haemost*, 9(1):71-78.

Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. (2011) Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics,* 67(3):1047-1056.

Hobbs BP, Landin R. (2018) Bayesian basket trial design with exchangeability monitoring. *Statistics in medicine*.

Hong H, Fu H, Carlin BP. (2018) Power and commensurate priors for synthesizing aggregate and individual patient level data in network meta‐analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 67(4):1047-1069.

Ibrahim JG, Chen M-H. (2000) Power prior distributions for regression models. *Statistical Science,* 15(1):46-60.

Jenkins M, Stone A, Jennison C. (2011) An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical statistics,* 10(4):347-356.

Keijzer-Veen MG, Dulger A, Dekker FW, Nauta J, van der Heijden BJ. (2010) Very preterm birth is a risk factor for increased systolic blood pressure at a young adult age. *Pediatr Nephrol*, 25(3):509-516.

Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. (2005) How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in medicine*, 24(15):2401-2428.

Lawn JE, Gravett MG, Nunes TM, Rubens CE, Stanton C. (2010) Global report on preterm birth and stillbirth (1 of 7): definitions, description of the burden and opportunities to improve data. *Bmc Pregnancy Childb*, 10(1):1.

Lei Y, Nollen N, Ahluwahlia JS, Yu Q, Mayo MS. (2015) An application in identifying high-risk populations in alternative tobacco product use utilizing logistic regression and CART: a heuristic comparison. *Bmc Public Health*, 15.

Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. (2003) Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine*, 26(3):172-181.

Lim J, Walley R, Yuan J, Liu J, Dabral A, Best N, Grieve A, Hampson L, Wolfram J, Woodward P. (2018) Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: review of methods and opportunities. *Therapeutic innovation & regulatory science*, 52(5):546-559.

Lipkovich I, Dmitrienko A, Denne J, Enas G. (2011) Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in medicine*, 30(21):2601-2621.

Little SE, Janiak E, Bartz D, Smith NA. (2015) Second trimester dilation and evacuation: a risk factor for preterm birth? *J Perinatol*, 35(12):1006-1010.

Liu GF. (2018) A dynamic power prior for borrowing historical data in noninferiority trials with binary endpoint. *Pharmaceutical statistics*, 17(1):61-73.

Loh WY. (2011) Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14-23.

Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JP, Salman RA-S, Chan A-W, Glasziou P. (2014) Biomedical research: increasing value, reducing waste. *The Lancet*, 383(9912):101-104.

Magnusson BP, Turnbull BW. (2013) Group sequential enrichment design incorporating subgroup selection. *Statistics in medicine*, 32(16):2695-2714.

Martius JA, Steck T, Oehler MK, Wulf K-H. (1998) Risk factors associated with preterm (< 37+ 0 weeks) and early preterm birth (< 32+ 0 weeks): univariate and multivariate analysis of 106 345 singleton births from the 1994 statewide perinatal survey of Bavaria. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 80(2):183-189.

McCormick MC. (1985) The contribution of low birth weight to infant mortality and childhood morbidity. *New England journal of medicine*, 312(2):82-90.

Mehta CR, Gao P. (2011) Population enrichment designs: case study of a large multinational trial. *Journal of biopharmaceutical statistics*, 21(4):831-845.

Morita S, Yamamoto H, Sugitani Y. (2014) Biomarker‐based Bayesian randomized phase II clinical trial design to identify a sensitive patient subpopulation. *Statistics in medicine*, 33(23):4008-4016.

Murray TA, Hobbs BP, Lystig TC, Carlin BP. (2014) Semiparametric Bayesian commensurate survival model for post‐market medical device surveillance with non‐exchangeable historical data. *Biometrics*, 70(1):185-191.

Neelon B, O' Malley AJ. (2010) Bayesian analysis using power priors with application to pediatric quality of care. *Journal of Biometrics and Biostatistics*, 1(1):1-9.

Neerhof MG, Cravello C, Haney EI, Silver RK. (1999) Timing of labor induction after premature rupture of membranes between 32 and 36 weeks' gestation. *Am J Obstet Gynecol*, 180(2):349-352.

Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter DJ. (2010) Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5-18.

Nollen NL, Ahluwalia JS, Lei Y, Yu Q, Scheuermann TS, Mayo MS. (2015) Adult Cigarette Smokers at Highest Risk for Concurrent Alternative Tobacco Product Use Among a Racially/Ethnically and Socioeconomically Diverse Sample. *Nicotine & Tobacco Research*.

Norman M. (2010) Preterm birth--an emerging risk factor for adult hypertension? *Semin Perinatol*, 34(3):183-187.

Papageorgiou SN, Koretsi V, Jäger A. (2017) Bias from historical control groups used in orthodontic research: a meta-epidemiological study. *European Journal of Orthodontics*, 39(1):98-105.

Richesson RL. (2011) *Data standards in diabetes patient registries*. In.: SAGE Publications.

Rosenbaum PR, Rubin DB. (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516-524.

Rosenblum M, Luber B, Thompson RE, Hanley D. (2016) Group sequential designs with prospectively planned rules for subpopulation enrichment. *Statistics in medicine*, 35(21):3776-3791.

Ruberg SJ, Shen L. (2015) Personalized medicine: four perspectives of tailored medicine. *Statistics in Biopharmaceutical Research*, 7(3):214-229.

Rufibach K, Chen M, Nguyen H. (2016) Comparison of different clinical development plans for confirmatory subpopulation selection. *Contemporary clinical trials*, 47:78-84.

Russell RB, Green NS, Steiner CA, Meikle S, Howse JL, Poschman K, Dias T, Potetz L, Davidoff MJ, Damus K. (2007) Cost of hospitalization for preterm and low birth weight infants in the United States. *Pediatrics*, 120(1):e1-e9.

Saccone G, Perriera L, Berghella V. (2015) Prior uterine evacuation of pregnancy as independent risk factor for preterm birth: a systematic review and metaanalysis. *Am J Obstet Gynecol*.

Saigal S, Doyle LW. (2008) An overview of mortality and sequelae of preterm birth from infancy to adulthood. *The Lancet*, 371(9608):261-269.

Salman RA-S, Beller E, Kagan J, Hemminki E, Phillips RS, Savulescu J, Macleod M, Wisely J, Chalmers I. (2014) Increasing value and reducing waste in biomedical research regulation and management. *The Lancet*, 383(9912):176-185.

Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. (2014) Robust meta‐analytic‐predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023-1032.

Simhan HN, Bodnar LM, Kim KH. (2011) Lower genital tract inflammatory milieu and the risk of subsequent preterm birth: an exploratory factor analysis. *Paediatr Perinat Epidemiol*, 25(3):277-282.

Simon N, Simon R. (2013) Adaptive enrichment designs for clinical trials. *Biostatistics*, 14(4):613-625.

Spiegelhalter DJ, Abrams KR, Myles JP. (2004) *Bayesian approaches to clinical trials and health-care evaluation*, vol. 13: John Wiley & Sons.

Stephansson O, Larsson H, Pedersen L, Kieler H, Granath F, Ludvigsson JF, Falconer H, Ekbom A, Sorensen HT, Norgaard M. (2010) Crohn's disease is a risk factor for preterm birth. *Clin Gastroenterol Hepatol*, 8(6):509-515.

Su X, Peña AT, Liu L, Levine RA. (2018) Random forests of interaction trees for estimating individualized treatment effects in randomized trials. *Statistics in medicine*, 37(17):2547-2560.

Tan HZ, Wen SW, Chen XK, Demissie K, Walker M. (2007) Early prediction of preterm birth for singleton, twin, and triplet pregnancies. *Eur J Obstet Gyn R B*, 131(2):132-137.

Treyvaud K, Doyle LW, Lee KJ, Roberts G, Cheong JL, Inder TE, Anderson PJ. (2011) Family functioning, burden and parenting stress 2years after very preterm birth. *Early human development*, 87(6):427-431.

Viele K, Berry S, Neuenschwander B, Amzal B, Chen F, Enas N, Hobbs B, Ibrahim JG, Kinnersley N, Lindborg S. (2014) Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics*, 13(1):41-54.

Wang S-J, Hung HJ. (2013) Adaptive enrichment with subpopulation selection at interim: methodologies, applications and design considerations. *Contemporary clinical trials*, 36(2):673-681.

Wassmer G, Dragalin V. (2015) Designing issues in confirmatory adaptive population enrichment trials. *Journal of biopharmaceutical statistics*, 25(4):651-669.

WHO. (2018) Preterm birth. https://www.who.int/en/news-room/fact-sheets/detail/preterm-birth.

Witt WP, Cheng ER, Wisk LE, Litzelman K, Chatterjee D, Mandell K, Wakeel F. (2014) Preterm birth in the United States: the impact of stressful life events prior to conception and maternal age. *Am J Public Health*, 104 Suppl 1:S73-80.

Witt WP, Cheng ER, Wisk LE, Litzelman K, Chatterjee D, Mandell K, Wakeel FJAjoph. (2014) Preterm birth in the United States: the impact of stressful life events prior to conception and maternal age. 104(S1):S73-S80.

Zhang C, Mayo MS, Gajewski BJ. (2018) Comments on "Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials". *Statistics in medicine*, 37(19):2900-2901.

Appendices

Appendix 2.1 the development of full complete conditional distribution of treatment effectiveness difference between Arm A and Arm B (θg), and of treatment effectiveness of Arm A (γ_g) under different models given subgroup g; DP specification.

Pairwise independent model specification

- Arm A :

Suppose

$$Y_{1g}^{(A)}, Y_{2g}^{(A)}, Y_{3g}^{(A)}, Y_{4g}^{(A)} \text{ .... } Y_{N_g^{(A)}g}^{(A)} \sim N\left(\gamma_g, \sigma^2\right),\ \gamma_g \sim N\left(\mu_g^{(A)}, \tau_g^{(A),2}\right), \sigma^2 \sim IG\left(\frac{\sigma_n}{2},\ \frac{\sigma_\mu^2 \sigma_n}{2}\right),$$

where “A” in the superscript stands for “Arm A”;

g is the index of the subgroup, $g = \{1, 2, 3, \ldots g_n\}$, e.g., $g_n = 4$ refers there are four subgroups and $g_n = 8$ refers there are eight subgroups;

$N_g^{(A)}$ represents for the sample size of the subgroup g from Arm A;

γ_g and σ^2 denote the mean and variance of the normal distribution from which the Arm A subgroup sample is drawn, and we assume all the distributions from which the sample is drawn have the same variance (σ^2);

$\mu_g^{(A)}$ and $\tau_g^{(A),2}$ denote the mean and variance of normal distribution as which γ_g is distributed; σ_n and σ_μ^2 are the fixed parameters of inverse gamma distribution as which σ^2 is distributed.

- Arm B :

Suppose $Y_{1g}^{(B)}, Y_{2g}^{(B)}, Y_{3g}^{(B)}, Y_{4g}^{(B)} \text{ .... } Y_{N_g^{(B)}g}^{(B)} \sim N\left(\gamma_g +\ \theta_g, \sigma^2\right), \theta_g \sim N\left(\mu_g^{(B)}, \tau_g^{(B),2}\right),$

where "B" in the superscript stands for "Arm B ";

$N_g^{(B)}$ represents for the sample size of the subgroup g from Arm B;

g, γ_g and σ^2 have meanings and or are distributed identically to those from Arm A circumstance;

θ_g is the treatment difference between Arm B and Arm A given subgroup g;

$\mu_g^{(B)}$ and $\tau_g^{(B),2}$ denotes the mean and variance of normal distribution as which θ_g is distributed.

-The full joint PDF under the pairwise independent model:

$$\prod_{g=1}^{g_n}\prod_{i=1}^{N_g^{(A)}}\frac{1}{\sqrt{2\pi}\sigma}exp\left(-\frac{\left(Y_{ig}^{(A)}-\gamma_g\right)^2}{2\sigma^2}\right)\prod_{g=1}^{g_n}\prod_{i=1}^{N_g^{(B)}}\frac{1}{\sqrt{2\pi}\sigma}exp\left(-\frac{\left(Y_{ig}^{(B)}-\gamma_g-\theta_g\right)^2}{2\sigma^2}\right)\times$$

$$\prod_{g=1}^{g_n}\frac{1}{\sqrt{2\pi}\tau_g^{(A)}}exp\left(-\frac{\left(\gamma_g-\mu_g^{(A)}\right)^2}{2\tau_g^{(A),2}}\right)\prod_{g=1}^{g_n}\frac{1}{\sqrt{2\pi}\tau_g^{(B)}}exp\left(-\frac{\left(\theta_g-\mu_g^{(B)}\right)^2}{2\,\tau_g^{(B),2}}\right)\left(\frac{\left(\frac{\sigma_\mu^2\sigma_n}{2}\right)^{\frac{\sigma_n}{2}}exp\left(-\frac{\frac{\sigma_\mu^2\sigma_n}{2}}{\sigma^2}\right)}{(\sigma^2)^{\frac{\sigma_n}{2}+1}\Gamma\left(\frac{\sigma_n}{2}\right)}\right).$$

-The full complete conditional distribution of treatment effectiveness difference between Arm A and Arm B given subgroup $\boldsymbol{g}$ ($\boldsymbol{\theta_g}$):

$$P(\theta_g|\vec{\boldsymbol{Y}},\gamma_g,\sigma^2,\mu_g^{(B)},\tau_g^{(B),2})\ \propto exp\left(-\frac{\sum_{i=1}^{N_g^{(B)}}\left(Y_{ig}^{(B)}-\gamma_g-\theta_g\right)^2}{2\sigma^2}\right)exp\ \left(-\frac{\left(\theta_g-\mu_g^{(B)}\right)^2}{2\tau_g^{(B),2}}\right)$$

$$=exp\left(-\frac{N_g^{(B)}\theta_g^2-2\theta_g\sum_{i=1}^{N_g^{(B)}}\left(Y_{ig}^{(B)}-\gamma_g\right)+\sum_{i=1}^{N_g^{(B)}}\left(Y_{ig}^{(B)}-\gamma_g\right)^2}{2\sigma^2}\right)exp\left(-\frac{\theta_g^2-2\theta_g\mu_g^{(B)}+\mu_g^{(B),2}}{2\tau_g^{(B),2}}\right)$$

$$\propto exp\ \left(\frac{N_g^{(B)}\theta_g^2-2\theta_g\sum_{i=1}^{N_g^{(B)}}\left(Y_{ig}^{(B)}-\gamma_g\right)}{2\sigma^2}\right)exp\left(-\frac{\theta_g^2-2\theta_g\mu_g^{(B)}}{2\tau_g^{(B),2}}\right)$$

$$=exp\left(-\frac{\theta_g^2\frac{N_g^{(B)}}{\sigma^2}-2\theta_g\frac{\sum_{i=1}^{N_g^{(B)}}\left(Y_{ig}^{(B)}-\gamma_g\right)}{\sigma^2}}{2}\right)exp\left(-\frac{\theta_g^2\frac{1}{\tau_g^{(B),2}}-2\theta_g\frac{\mu_g^{(B)}}{\tau_g^{(B),2}}}{2}\right)$$

$$= exp\left(-\frac{\theta_g^2\left(\frac{N_g^{(B)}}{\sigma^2}+\frac{1}{\tau_g^{(B),2}}\right)-2\theta_g\left(\frac{\sum_{i=1}^{N_g^{(B)}}\left(Y_{ig}^{(B)}-\gamma_g\right)}{\sigma^2}+\frac{\mu_g^{(B)}}{\tau_g^{(B),2}}\right)}{2}\right),\Rightarrow$$

$$\theta_g|\vec{\boldsymbol{Y}},\gamma_g,\sigma^2,\mu_g^{(B)},\tau_g^{(B),2}\sim N\left(\frac{\frac{\sum_{i=1}^{N_g^{(B)}}\left(Y_{ig}^{(B)}-\gamma_g\right)}{\sigma^2}+\frac{\mu_g^{(B)}}{\tau_g^{(B),2}}}{\frac{N_g^{(B)}}{\sigma^2}+\frac{1}{\tau_g^{(B),2}}},\frac{1}{\frac{N_g^{(B)}}{\sigma^2}+\frac{1}{\tau_g^{(B),2}}}\right)$$

The mean and variance of the normal distribution arrived above can be simplified as

$\frac{\tau_g^{(B),2}N_g^{(B)}\left(\bar{Y}_g^{(B)}-\gamma_g\right)+\sigma^2\mu_g^{(B)}}{N_g^{(B)}\tau_g^{(B),2}+\sigma^2}$ and $\frac{\tau_g^{(B),2}\sigma^2}{N_g^{(B)}\tau_g^{(B),2}+\sigma^2}$.

- The full complete conditional distribution of treatment effectiveness of Arm A given subgroup $\boldsymbol{g}$ $\boldsymbol{(\gamma_g)}$:

$$P\left(\gamma_g\middle|\vec{\boldsymbol{y}},\theta_g,\sigma^2,\mu_g^{(A)},\tau_g^{(A),2}\right)$$

$$\propto exp\left(-\frac{\sum_{i=1}^{N_g^{(A)}}\left(Y_{ig}^{(A)}-\gamma_g\right)^2}{2\sigma^2}\right)exp\left(-\frac{\sum_{i=1}^{N_g^{(B)}}\left(Y_{ig}^{(B)}-\gamma_g-\theta_g\right)^2}{2\sigma^2}\right)exp\left(-\frac{\left(\gamma_g-\mu_g^{(A)}\right)^2}{2\tau_g^{(A),2}}\right)$$

$$\propto exp\left(-\frac{\left(N_g^{(A)}+N_g^{(B)}\right)\gamma_g^2-2\gamma_g\left(N_g^{(A)}\bar{Y}_g^{(A)}+N_g^{(B)}\bar{Y}_g^{(B)}-N_g^{(B)}\theta_g\right)}{2\sigma^2}\right)exp\left(-\frac{\gamma_g^2-2\gamma_g\mu_g^{(A)}}{2\tau_g^{(A),2}}\right)$$

$$= exp\left(-\frac{\gamma_g^2\left(\frac{N_g^{(A)}+N_g^{(B)}}{\sigma^2}\right)-2\gamma_g\left(\frac{N_g^{(A)}\bar{Y}_g^{(A)}+N_g^{(B)}\bar{Y}_g^{(B)}-N_g^{(B)}\theta_g}{\sigma^2}\right)}{2}\right)exp\left(-\frac{\gamma_g^2\frac{1}{\tau_g^{(A),2}}-2\gamma_g\frac{\mu_g^{(A)}}{\tau_g^{(A),2}}}{2}\right)$$

$$= exp\left(-\frac{\gamma_g^2\left(\frac{N_g^{(A)}+N_g^{(B)}}{\sigma^2}+\frac{1}{\tau_g^{(A),2}}\right)-2\gamma_g\left(\frac{N_g^{(A)}\bar{Y}_g^{(A)}+N_g^{(B)}\bar{Y}_g^{(B)}-N_g^{(B)}\theta_g}{\sigma^2}+\frac{\mu_g^{(A)}}{\tau_g^{(A),2}}\right)}{2}\right), \Rightarrow$$

$$\gamma_g|\vec{\boldsymbol{Y}},\theta_g,\sigma^2,\mu_g^{(B)},\tau_g^{(B),2}\sim N\left(\frac{\frac{N_g^{(A)}\bar{Y}_g^{(A)}+N_g^{(B)}\bar{Y}_g^{(B)}-N_g^{(B)}\theta_g}{\sigma^2}+\frac{\mu_g^{(A)}}{\tau_g^{(A),2}}}{\frac{N_g^{(A)}+N_g^{(B)}}{\sigma^2}+\frac{1}{\tau_g^{(A),2}}},\frac{1}{\frac{N_g^{(A)}+N_g^{(B)}}{\sigma^2}+\frac{1}{\tau_g^{(A),2}}}\right)$$

The mean and variance can be simplified as $\frac{\tau_g^{(A),2}\left(N_g^{(A)}\bar{Y}_g^{(A)}+N_g^{(B)}\bar{Y}_g^{(B)}-N_g^{(B)}\theta_g\right)+\sigma^2\mu_g^{(A)}}{\left(N_g^{(A)}+N_g^{(B)}\right)\tau_g^{(A),2}+\sigma^2}$ and

$\frac{\tau_g^{(A),2}\sigma^2}{\left(N_g^{(A)}+N_g^{(B)}\right)\tau_g^{(A),2}+\sigma^2}$.

<u>Hierarchical model specification</u>

- Arm A :

Suppose

$$Y_{1g}^{(A)},Y_{2g}^{(A)},Y_{3g}^{(A)},Y_{4g}^{(A)} \dots.\ Y_{N_g^{(A)}g}^{(A)} \sim N\left(\gamma_g,\sigma^2\right),$$

$$\gamma_g \sim N\left(\mu_\gamma^{(A)},\tau_\gamma^{(A),2}\right),\mu_\gamma^{(A)} \sim N\left(\mu_0,\sigma_0^2\right),\tau_\gamma^{(A),2} \sim IG\left(\frac{\tau_n}{2},\ \frac{\tau_\mu^2\tau_n}{2}\right),$$

where “A” in the superscript still stands for “Arm A”;

$g, N_g^{(A)}, \gamma_g, \sigma^2, \sigma_n$ and σ_μ^2 have the meanings and are distributed identically to those from Arm A circumstance in pairwise independent model;

γ_g has the same meanings as that from Arm A in pairwise independent model but with different distribution;

$\mu_\gamma^{(A)}$ and $\tau_\gamma^{(A),2}$ denotes the mean and variance of normal distribution as which γ_g is distributed;

μ_0 and σ_0^2 denotes the mean and variance of normal distribution as which $\mu_\gamma^{(A)}$ is distributed;

τ_n and τ_μ^2 are the fixed parameters of inverse gamma distribution as which $\tau_\gamma^{(A),2}$ is distributed.

- Arm B :

Suppose

$$Y_{1g}^{(B)}, Y_{2g}^{(B)}, Y_{3g}^{(B)}, Y_{4g}^{(B)} \ .... \ Y_{N_g^{(B)}g}^{(B)} \sim N\left(\gamma_g + \theta_g, \sigma^2\right),$$

$$\theta_g \sim N\left(\mu_\gamma^{(B)}, \tau_\gamma^{(B),2}\right), \mu_\gamma^{(B)} \sim N\left(\mu_0, \sigma_0^2\right), \tau_\gamma^{(B),2} \sim IG\left(\frac{\tau_n}{2}, \ \frac{\tau_\mu^2 \tau_n}{2}\right),$$

where "B" in the superscript still stands for "Arm B";

$g, N_g^{(B)}, \gamma_g, \sigma^2, \sigma_n$ and σ_μ^2 have the meanings and are distributed identically to those from Arm B circumstance in pairwise independent model;

θ_g has the same meanings as that from Arm B in pairwise independent model but with different distribution;

$\mu_\gamma^{(B)}$ and $\tau_\gamma^{(B),2}$ denotes the mean and variance of normal distribution as which θ_g is distributed;

μ_0 and σ_0^2 denotes the mean and variance of normal distribution as which $\mu_\gamma^{(B)}$ is distributed, and

τ_n and τ_μ^2 are the fixed parameters of inverse gamma distribution as which $\tau_\gamma^{(B),2}$ is distributed, which is identical to those from Arm A circumstance in Hierarchical Model.

-The full joint PDF under the Hierarchical Model:

$$\prod_{g=1}^{g_n} \prod_{i=1}^{N_g^{(A)}} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{\left(Y_{ig}^{(A)} - \gamma_g\right)^2}{2\sigma^2}\right) \prod_{g=1}^{g_n} \prod_{i=1}^{N_g^{(B)}} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{\left(Y_{ig}^{(B)} - \gamma_g - \theta_g\right)^2}{2\sigma^2}\right) \times$$

$$\prod_{g=1}^{g_n} \frac{1}{\sqrt{2\pi}\tau_\gamma^{(A)}} exp\left(-\frac{\left(\gamma_g - \mu_\gamma^{(A)}\right)^2}{2\tau_\gamma^{(A),2}}\right) \frac{1}{\sqrt{2\pi}\sigma_0} exp\left(-\frac{\left(\mu_\gamma^{(A)} - \mu_0\right)^2}{2\sigma_0^2}\right) \times$$

$$\prod_{g=1}^{g_n} \frac{1}{\sqrt{2\pi}\tau_\gamma^{(B)}} exp\left(-\frac{\left(\theta_g - \mu_\gamma^{(B)}\right)^2}{2\,\tau_\gamma^{(B),2}}\right) \frac{1}{\sqrt{2\pi}\sigma_0} exp\left(-\frac{\left(\mu_\gamma^{(B)} - \mu_0\right)^2}{2\sigma_0^2}\right) \times$$

$$\left(\frac{\left(\frac{\sigma_\mu^2 \sigma_n}{2}\right)^{\frac{\sigma_n}{2}} exp\left(-\frac{\frac{\sigma_\mu^2 \sigma_n}{2}}{\sigma^2}\right)}{(\sigma^2)^{\frac{\sigma_n}{2}+1} \Gamma\left(\frac{\sigma_n}{2}\right)}\right)\left(\frac{\left(\frac{\tau_\mu^2 \tau_n}{2}\right)^{\frac{\tau_n}{2}} exp\left(-\frac{\frac{\tau_\mu^2 \tau_n}{2}}{\tau_\gamma^{(A),2}}\right)}{\left(\tau_\gamma^{(A),2}\right)^{\frac{\tau_n}{2}+1} \Gamma\left(\frac{\tau_n}{2}\right)}\right)\left(\frac{\left(\frac{\tau_\mu^2 \tau_n}{2}\right)^{\frac{\tau_n}{2}} exp\left(-\frac{\frac{\tau_\mu^2 \tau_n}{2}}{\tau_\gamma^{(B),2}}\right)}{\left(\tau_\gamma^{(B),2}\right)^{\frac{\tau_n}{2}+1} \Gamma\left(\frac{\tau_n}{2}\right)}\right).$$

- The full complete conditional distribution of treatment effectiveness difference between Arm A and Arm B given subgroup $\boldsymbol{g}$ ($\boldsymbol{\theta_g}$):

$P\left(\theta_g \middle| \vec{\boldsymbol{Y}}, \gamma_g, \sigma^2, \mu_\gamma^{(B)}, \tau_\gamma^{(B),2}, \mu_0, \sigma_0^2\right) = P\left(\theta_g \middle| \vec{\boldsymbol{Y}}, \gamma_g, \sigma^2, \mu_\gamma^{(B)}, \tau_\gamma^{(B),2}\right) \propto$

$exp\left(-\frac{\sum_{i=1}^{N_g^{(B)}} \left(Y_{ig}^{(B)} - \gamma_g - \theta_g\right)^2}{2\sigma^2}\right) exp\left(-\frac{\left(\theta_g - \mu_\gamma^{(B)}\right)^2}{2\tau_\gamma^{(B),2}}\right)$, which arrives at the similar expression

$P\left(\theta_g \middle| \vec{\boldsymbol{Y}}, \gamma_g, \sigma^2, \mu_g^{(B)}, \tau_g^{(B),2}\right)$ as in pairwise independent model. Finally,

$\theta_g | \vec{\boldsymbol{Y}}, \gamma_g, \sigma^2, \mu_\gamma^{(B)}, \tau_\gamma^{(B),2}, \mu_0, \sigma_0^2 = \theta_g | \vec{\boldsymbol{Y}}, \gamma_g, \sigma^2, \mu_\gamma^{(B)}, \tau_\gamma^{(B),2} \sim$

$$N\left(\frac{\frac{\sum_{i=1}^{N_g^{(B)}} \left(Y_{ig}^{(B)} - \gamma_g\right)}{\sigma^2} + \frac{\mu_\gamma^{(B)}}{\tau_\gamma^{(B),2}}}{\frac{N_g^{(B)}}{\sigma^2} + \frac{1}{\tau_\gamma^{(B),2}}}, \frac{1}{\frac{N_g^{(B)}}{\sigma^2} + \frac{1}{\tau_\gamma^{(B),2}}}\right)$$

The mean and variance can be simplified as $\frac{\tau_\gamma^{(B),2} \sum_{i=1}^{N_g^{(B)}} \left(Y_{ig}^{(B)} - \gamma_g\right) + \sigma^2 \mu_\gamma^{(B)}}{N_g^{(B)} \tau_\gamma^{(B),2} + \sigma^2}$ and $\frac{\tau_\gamma^{(B),2} \sigma^2}{N_g^{(B)} \tau_\gamma^{(B),2} + \sigma^2}$.

- The full complete conditional distribution of $\boldsymbol{\mu_\gamma^{(B)}}$:

Since $\mu_\gamma^{(B)}$ is considered as random variable in hierarchical model, the full complete conditional distribution of $\mu_\gamma^{(B)}$ is derived below:

$$P\left(\mu_\gamma^{(B)}\middle|\vec{\boldsymbol{\theta}},\tau_\gamma^{(B),2},\mu_0,\sigma_0^2\right) \propto\ exp\left(-\frac{\sum_{g=1}^{g_n}\left(\theta_g-\mu_\gamma^{(B)}\right)^2}{2\tau_\gamma^{(B),2}}\right) exp\ \left(-\frac{\left(\mu_\gamma^{(B)}-\mu_0\right)^2}{2\sigma_0^2}\right)$$

$$\propto exp\ \left(-\frac{g_n\mu_\gamma^{(B),2}-2\mu_\gamma^{(B)}\sum_{g=1}^{g_n}\theta_g}{2\tau_\gamma^{(B),2}}\right)exp\left(-\frac{\mu_\gamma^{(B),2}-2\mu_\gamma^{(B)}\mu_0}{2\sigma_0^2}\right)$$

$$= exp\left(-\frac{\mu_\gamma^{(B),2}\left(\frac{g_n}{\tau_\gamma^{(B),2}}+\frac{1}{\sigma_0^2}\right)+2\mu_\gamma^{(B)}\left(\frac{\sum_{g=1}^{g_n}\theta_g}{\tau_\gamma^{(B),2}}+\frac{\mu_0}{\sigma_0^2}\right)}{2}\right),\Rightarrow$$

$$\mu_\gamma^{(B)}|\vec{\boldsymbol{\theta}},\tau_\gamma^{(B),2},\mu_0,\sigma_0^2\sim N\left(\frac{\frac{\sum_{g=1}^{g_n}\theta_g}{\tau_\gamma^{(B),2}}+\frac{\mu_0}{\sigma_0^2}}{\frac{g_n}{\tau_\gamma^{(B),2}}+\frac{1}{\sigma_0^2}},\frac{1}{\frac{g_n}{\tau_\gamma^{(B),2}}+\frac{1}{\sigma_0^2}}\right)$$

The mean and variance can be simplified as $\frac{\sigma_0^2\sum_{g=1}^{g_n}\theta_g+\tau_\gamma^{(B),2}\mu_0}{g_n\sigma_0^2+\tau_g^{(B),2}}$ and $\frac{\tau_g^{(B),2}\sigma_0^2}{g_n\sigma_0^2+\tau_g^{(B),2}}$.

- The full complete conditional distribution of treatment effectiveness of Arm A given subgroup g ($\boldsymbol{\gamma_g}$):

$$P\left(\gamma_g\middle|\vec{\boldsymbol{Y}},\theta_g,\sigma^2,\mu_\gamma^{(A)},\tau_\gamma^{(A),2},\mu_0,\sigma_0^2\right)=P\left(\gamma_g\middle|\vec{\boldsymbol{Y}},\theta_g,\sigma^2,\mu_\gamma^{(A)},\tau_\gamma^{(A),2}\right)$$

$\propto exp\left(-\frac{\sum_{i=1}^{N_g^{(A)}}\left(Y_{ig}^{(A)}-\gamma_g\right)^2}{2\sigma^2}\right)exp\left(-\frac{\sum_{i=1}^{N_g^{(B)}}\left(Y_{ig}^{(B)}-\gamma_g-\theta_g\right)^2}{2\sigma^2}\right)exp\left(-\frac{\left(\gamma_g-\mu_\gamma^{(A)}\right)^2}{2\tau_\gamma^{(A),2}}\right)$, which arrives at the similar expression as $P\left(\gamma_g\middle|\vec{\boldsymbol{Y}},\theta_g,\sigma^2,\mu_g^{(A)},\tau_g^{(A),2}\right)$ in pairwise independent model. Finally,

$$\gamma_g\middle|\vec{\boldsymbol{Y}},\theta_g,\sigma^2,\mu_\gamma^{(A)},\tau_\gamma^{(A),2},\mu_0,\sigma_0^2=\gamma_g\middle|\vec{\boldsymbol{Y}},\theta_g,\sigma^2,\mu_\gamma^{(A)},\tau_\gamma^{(A),2}\sim$$

$$\left(\frac{\frac{N_g^{(A)}\bar{Y}_g^{(A)}+N_g^{(B)}\bar{Y}_g^{(B)}-N_g^{(B)}\theta_g}{\sigma^2}+\frac{\mu_\gamma^{(A)}}{\tau_\gamma^{(A),2}}}{\frac{N_g^{(A)}+N_g^{(B)}}{\sigma^2}+\frac{1}{\tau_\gamma^{(A),2}}},\frac{1}{\frac{N_g^{(A)}+N_g^{(B)}}{\sigma^2}+\frac{1}{\tau_\gamma^{(A),2}}}\right)$$

The mean and variance can be simplified as $\frac{\tau_\gamma^{(A),2}\left(N_g^{(A)}\bar{Y}_g^{(A)}+N_g^{(B)}\bar{Y}_g^{(B)}-N_g^{(B)}\theta_g\right)+\sigma^2\mu_\gamma^{(A)}}{\left(N_g^{(A)}+N_g^{(B)}\right)\tau_\gamma^{(A),2}+\sigma^2}$ and $\frac{\tau_\gamma^{(A),2}\sigma^2}{\left(N_g^{(A)}+N_g^{(B)}\right)\tau_\gamma^{(A),2}+\sigma^2}$.

- The full complete conditional distribution of $\boldsymbol{\mu_\gamma^{(A)}}$:

Still, we need to get the full complete conditional distribution of $\mu_\gamma^{(A)}$ given $\vec{\boldsymbol{\gamma}}, \tau_\gamma^{(A),2}, \mu_0, \sigma_0^2$. Since $\mu_\gamma^{(A)}$ is served as a random variable in hierarchical model.

$$P\left(\mu_\gamma^{(A)}\middle|\vec{\boldsymbol{\gamma}}, \tau_\gamma^{(A),2}, \mu_0, \sigma_0^2\right) \propto\ exp\left(-\frac{\sum_{g=1}^{g_n}\left(\gamma_g-\mu_\gamma^{(A)}\right)^2}{2\tau_\gamma^{(A),2}}\right) exp\ \left(-\frac{\left(\mu_\gamma^{(A)}-\mu_0\right)^2}{2\sigma_0^2}\right)$$

$$\propto exp\ \left(-\frac{g_n\mu_\gamma^{(A),2}-2\mu_\gamma^{(A)}\sum_{g=1}^{g_n}\gamma_g}{2\tau_\gamma^{(A),2}}\right) exp\left(-\frac{\mu_\gamma^{(A),2}-2\mu_\gamma^{(A)}\mu_0}{2\sigma_0^2}\right)$$

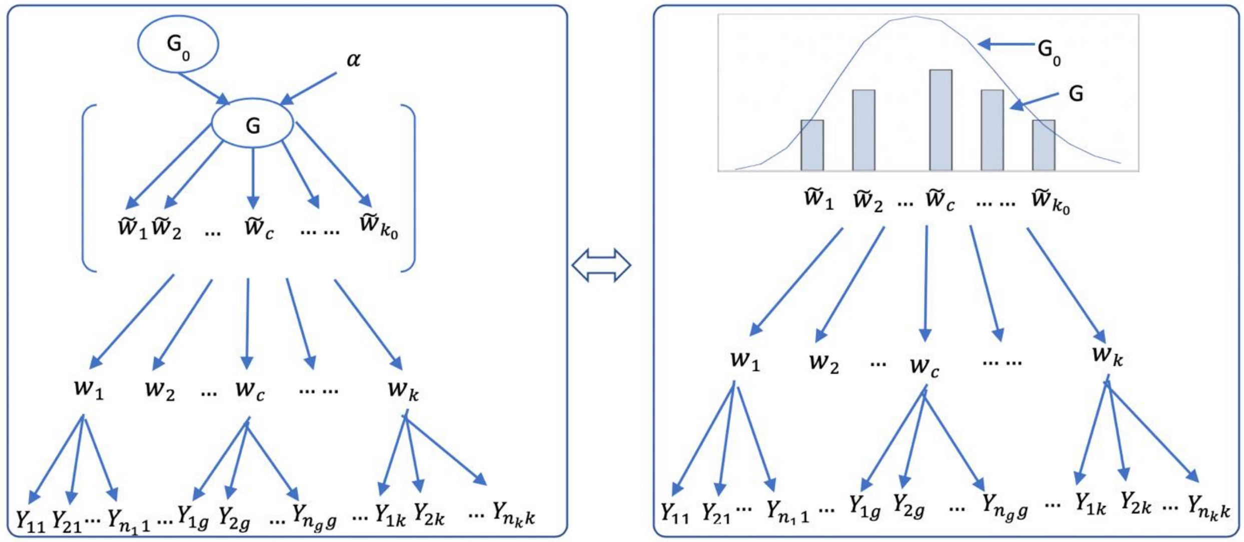
$$= exp\left(-\frac{\mu_\gamma^{(A),2}\left(\frac{g_n}{\tau_\gamma^{(A),2}}+\frac{1}{\sigma_0^2}\right)+2\mu_\gamma^{(B)}\left(\frac{\sum_{g=1}^{g_n}\gamma_g}{\tau_\gamma^{(A),2}}+\frac{\mu_0}{\sigma_0^2}\right)}{2}\right), \Rightarrow$$

$$\mu_\gamma^{(A)}|\vec{\boldsymbol{\gamma}}, \tau_\gamma^{(A),2}, \mu_0, \sigma_0^2 \sim \boldsymbol{N}\left(\frac{\frac{\sum_{g=1}^{g_n}\gamma_g}{\tau_\gamma^{(A),2}}+\frac{\mu_0}{\sigma_0^2}}{\frac{g_n}{\tau_\gamma^{(A),2}}+\frac{1}{\sigma_0^2}}, \frac{1}{\frac{g_n}{\tau_\gamma^{(A),2}}+\frac{1}{\sigma_0^2}}\right)$$

The mean and variance can be simplified as $\frac{\sigma_0^2\sum_{g=1}^{g_n}\gamma_g+\tau_\gamma^{(A),2}\mu_0}{g_n\sigma_0^2+\tau_g^{(A),2}}$ and $\frac{\tau_g^{(A),2}\sigma_0^2}{g_n\sigma_0^2+\tau_g^{(A),2}}$.

DP specification

-Model specification:



G0
α
G
$\tilde{w}_1$ $\tilde{w}_2$ … $\tilde{w}_c$ …… $\tilde{w}_{k_0}$
w_1 w_2 … w_c …… w_k
Y_{11} Y_{21} … $Y_{n_1 1}$ … Y_{1g} Y_{2g} … $Y_{n_g g}$ … Y_{1k} Y_{2k} … $Y_{n_k k}$
G0
G
$\tilde{w}_1$ $\tilde{w}_2$ … $\tilde{w}_c$ …… $\tilde{w}_{k_0}$
w_1 w_2 … w_c …… w_k
Y_{11} Y_{21} … $Y_{n_1 1}$ … Y_{1g} Y_{2g} … $Y_{n_g g}$ … Y_{1k} Y_{2k} … $Y_{n_k k}$

The DP is intuitively introduced by the graph above. Specifically, G_0 is the base distribution. It can be either continuous or discrete. From the perspective of easy understanding and our concrete research circumstance, it presents as Normal distribution in the right part of the graph above. $A_1, A_2 \dots A_r$ are a random partition of the support of G_0. The "Bars" stands for a random discrete distribution, denoting as G, drawn from G_0. G can be considered as the "discrete" form of G_0.

The relationship between G and G_0 is: $G \sim DP\ (\alpha, G_0)$, where α is scaling parameter, $\alpha > 0$. Generally, G is asymptotical to G_0 as $\alpha \rightarrow \infty$; G becomes very discrete (e.g., only several bars stand for G_0) as $\alpha \rightarrow 0$.

$$\left(G(A_1), G(\ A_2) \dots G\left(A_j\right) \dots G(A_r)\right) \sim Dirichlet\left(\alpha G_0(A_1), \alpha G_0(\ A_2) \dots \alpha G_0\left(A_j\right) \dots \alpha G_0(A_r)\right).$$

$$\sum_{j=1}^{r} G\left(A_j\right) = 1, \sum_{j=1}^{r} \alpha G_0\left(A_j\right) = \alpha \sum_{j=1}^{r} G_0\left(A_j\right) = \alpha$$

Based on the moment formula of Dirichlet distribution:

$$E\left(G(A_j)\right) = \frac{\alpha G_0(A_j)}{\sum_{j=1}^{r} \alpha G_0(A_j)} = \frac{\alpha G_0(A_j)}{\alpha} = G_0(A_j)$$

$$Var\left(G(A_j)\right) = \frac{\alpha G_0(A_j)\left(\sum_{j=1}^{r} \alpha G_0(A_j) - \alpha G_0(A_j)\right)}{\left(\sum_{j=1}^{r} \alpha G_0(A_j)\right)^2\left(\sum_{j=1}^{r} \alpha G_0(A_j) + 1\right)} = \frac{\alpha G_0(A_j)\left(\alpha - \alpha G_0(A_j)\right)}{\alpha^2(\alpha+1)}$$

$$= \frac{G_0(A_j)\left(1 - G_0(A_j)\right)}{\alpha + 1}$$

Data generation flow:

Firstly, draw $\widetilde{w}_1, \widetilde{w}_2 \dots \widetilde{w}_c \dots \widetilde{w}_{k_0}$ from G, denote $\widetilde{\boldsymbol{w}} = \left(\widetilde{w}_1, \widetilde{w}_2 \dots \widetilde{w}_c \dots \widetilde{w}_{k_0}\right)$. k_0 can be thought as the number of "original" clusters (Please note that the $\widetilde{w}$ and w are the general notation rather than treatment difference as specified in Section 2.3)

Next, draw the distinctive $w_1, w_2 \dots w_c \dots w_k$ from $\widetilde{\boldsymbol{w}}$. Note that $\widetilde{\boldsymbol{w}}$ is from G, which means w_c is originally from $G. P\left(w_g \in A_j\right) = G\left(A_j\right)$. Denote $\boldsymbol{w} = (w_1, w_2 \dots w_c \dots w_k)$. k is the number of distinctive elements of **w** and k is the number of "real" clusters, $k \leq k_0$.

Finally, draw $\boldsymbol{Y}$ from **w** = $(w_1, w_2 \dots w_c \dots w_k)$.

Overall, the model can be specified as:

$$Y_{ig}|w_c \sim F(w_c)$$

$$w_c \sim G$$

$$G \sim DP\ (\alpha, G_0)$$

In our DACTPerM research, w_c (still referring the general one) is the summation of Arm A treatment effect (γ_g) and treatment effect difference between two arms (θ_g); $G_0 = N\ (\mu_0, \sigma_0^2)$

-Posterior distribution of $\left(\boldsymbol{G}(\boldsymbol{A_1}), \boldsymbol{G}(\ \boldsymbol{A_2}), \dots \boldsymbol{G}\left(\boldsymbol{A_j}\right) \dots \boldsymbol{G}(\boldsymbol{A_r})\right)$.

Let n_j be the number of observed $\widetilde{w}_c$ in A_j, then

$\left(n_1, n_2 \dots n_j \dots n_r\right) \sim Mult\left(k_0, G(A_1), G(\ A_2) \dots G\left(A_j\right) \dots G(A_r)\right)$, where $k_0 = \sum_{j=1}^{r} n_j$.

$$P\left(G(A_1)\dots G\left(A_j\right)\dots G(A_r)\middle|\widetilde{\boldsymbol{w}}\right) = P\left(G(A_1)\dots G\left(A_j\right)\dots G(A_r)\middle|n_1\dots n_j\dots\ n_r\right)$$

$$\propto P\left(n_1\dots n_j\dots\ n_r\middle|k_0, G(A_1)\dots G\left(A_j\right)\dots G(A_r)\right) * P\left(G(A_1)\dots G\left(A_j\right)\dots G(A_r)\right)$$

$$= Mult\left(n_1\dots n_j\dots\ n_r\middle|G(A_1)\dots G\left(A_j\right)\dots G(A_r)\right) * Dirichlet\ \left(\alpha G_0(A_1)\dots\alpha G_0\left(A_j\right)\dots\alpha G_0(A_r)\right)$$

Due to the Dirichlet distribution conjugate to itself with respect to a Multinomial likelihood function, finally:

$$P\left(G(A_1)\dots G\left(A_j\right)\dots G(A_r)\middle|\widetilde{\boldsymbol{w}}\right)\sim\text{Dirichlet}\left(\alpha G_0(A_1)+n_1,\dots,\alpha G_0\left(A_j\right)+n_j,\dots,\alpha G_0(A_r)\right.$$
$$\left.+n_r\right),$$

which indicates that $G|\widetilde{w} \sim DP\left(\alpha+k_0, \frac{\alpha G_0\left(A_j\right)+n_j}{\alpha+k_0}\right)$.

-Predictive distribution of $\widetilde{\mathrm{w}}_{\mathrm{k}_0+1}$.

$$P\left(\widetilde{w}_{k_0+1}\in A_j|\widetilde{\boldsymbol{w}}\right) = \int P\left(\widetilde{w}_{k_0+1}\in A_j, G\left(A_j\right)|\widetilde{\boldsymbol{w}}\right)dG\left(A_j\right)$$

$$= \int P\left(\widetilde{w}_{k_0+1}\in A_j|G\left(A_j\right),\widetilde{\boldsymbol{w}}\right)P\left(G\left(A_j\right)|\widetilde{\boldsymbol{w}}\right)dG\left(A_j\right) = \int P\left(\widetilde{w}_{k_0+1}\in A_j|\widetilde{\boldsymbol{w}}\right)P\left(G\left(A_j\right)|\widetilde{\boldsymbol{w}}\right)dG\left(A_j\right)$$

$$= \int\left(G\left(A_j\right)|\widetilde{\boldsymbol{w}}\right)P\left(G\left(A_j\right)|\widetilde{\boldsymbol{w}}\right)dG\left(A_j\right) = E\left(G\left(A_j\right)|\widetilde{\boldsymbol{w}}\right) = \frac{1}{\alpha+k_0}\left(\alpha G_0\left(A_j\right)+n_j\right)$$

The last step is due to the posterior distribution of $\left(G(A_1), G(\ A_2),\dots G\left(A_j\right)\dots G(A_r)\right)$ is Dirichlet. Finally,

$\widetilde{w}_{k_0+1}\in A_j|\widetilde{\boldsymbol{w}} \sim \frac{1}{\alpha+k_0}\left(\alpha G_0\left(A_j\right)+n_j\right) = \frac{\alpha}{\alpha+k_0}G_0\left(A_j\right) + \frac{n_j}{\alpha+k_0}$, which indicates that $\widetilde{w}_{k_0+1}$ belongs to A_j is a weighted summation of $G_0\left(A_j\right)$ and n_j.

Appendix 2.2: ITP specification for virtual endpoints simulation

The formula used for endpoints simulation based on ITP is

$$Y_{it,g}^{(j)} = \left(\mu_g^{(j)} + S_{i,g}^{(j)} + \varepsilon_{it,g}^{(j)}\right)\left(\frac{1-EXP(k_g^{(j)}t)}{1-EXP(k_g^{(j)}T)}\right),$$

where $Y_{it,g}^{(j)}$ is the endpoint for subject *i* from arm *j* and subgroup g at visit *t*, *j* = [Arm A, Arm B]. $\mu_g^{(j)}$ is the mean value of the final endpoint from arm j and subgroup g, and $\mu_g^{(j)}$ = $\gamma_g$when *j* = Arm A and $\mu_g^{(j)}$ = $\gamma_g + \theta_g$ when *j* = Arm B. $S_{i,g}^{(j)}$ is the specific random effect for subject *i* from arm *j* and subgroup g, $S_{i,g}^{(j)} \sim N\left(0, \tau_g^{(j),2}\right)$, and $\tau_g^{(j),2}$ = $\omega_g^{(j)}\sigma^2$. $\omega_g^{(j)}$ is the fraction of the final response variance. $\varepsilon_{it,g}^{(j)}$ is the residual error for subject i from arm j and subgroup g, $\varepsilon_{it,g}^{(j)} \sim \left(0, \sigma_g^{\prime,(j),2}\right)$, and $\sigma_g^{\prime,(j),2}$ = $\sigma^2 - \tau_g^{(j),2}$ = $\left(1-\omega_g^{(j)}\right)\sigma^2$, $\sigma^2 = \tau_g^{(j),2} + \sigma_g^{\prime,(j),2} = \omega_g^{(j)}\sigma^2 + \left(1-\omega_g^{(j)}\right)\sigma^2$. The variability ($\sigma^2$, denoted as in Appendix A) for a subject divides two parts: per subject component $\left(\omega_g^{(j)}\sigma^2\right)$ and a component $\left(\left(1-\omega_g^{(j)}\right)\sigma^2\right)$ varying between visits. $k_g^{(j)}$ is a shape parameter of the exponential component, and it controls how quickly the response approaching the final one. Generally, smaller $k_g^{(j)}$ value makes the responses within the study to achieve the final endpoint value quickly; we specify the $\omega_g^{(j)}$ and $k_g^{(j)}$ identically for all subgroups of the two arms, respectively; t is the specific visit time that $Y_{it,g}^{(j)}$ is observed; T is the final endpoint is observed.

Appendix 3.1: The derivation of posterior probability of response rate via power prior borrowing method

$$\pi(\theta_{cc}|\boldsymbol{D},\boldsymbol{D_0},\alpha) \propto L(\theta_{cc}|\boldsymbol{D})L(\theta_{hc}|\boldsymbol{D_0})^{\alpha}\pi_0(\theta_{hc})$$

$$= \binom{n_{cc}}{y_{cc}}\theta_{cc}^{y_{cc}}(1-\theta_{cc})^{(n_{cc}-y_{cc})}\binom{n_{hc}}{y_{hc}}(\theta_{hc})^{\alpha y_{hc}}(1-\theta_{hc})^{\alpha*(n_{hc}-y_{hc})}\frac{\Gamma(1)}{\Gamma(0.5)\Gamma(0.5)}{\theta_{hc}}^{(0.5-1)}(1 - \theta_{hc})^{(0.5-1)}$$

$$\propto \theta_{cc}^{y_{cc}}(1-\theta_{cc})^{(n_{cc}-y_{cc})}(\theta_{hc})^{\alpha y_{hc}}((1-\theta_{hc})^{\alpha*(n_{hc}-y_{hc})}{\theta_{hc}}^{(0.5-1)}(1-\theta_{hc})^{(0.5-1)}$$

$$= \theta_{cc}^{y_{cc}}(1-\theta_{cc})^{(n_{cc}-y_{cc})}(\theta_{cc})^{\alpha y_{hc}}(1-\theta_{cc})^{\alpha*(n_{cc}-y_{hc})}{\theta_{cc}}^{(0.5-1)}(1-\theta_{cc})^{(0.5-1)}$$

$$= \theta_{cc}^{(y_{cc}+\alpha y_{hc}+0.5-1)}(1-\theta_{cc})^{((n_{cc}-y_{cc})+\alpha*(n_{cc}-y_{hc})+0.5-1)}$$

$$\Rightarrow\ \theta_{cc} \sim Beta\big((y_{cc}+\alpha y_{hc}+0.5,(n_{cc}-y_{cc})+\big(\alpha*(n_{cc}-y_{hc})\big)+0.5\big)$$

Appendix 3.2: Stan code of commensurate prior

```
data {
  int<lower=0> H; //Indicate how many studies
  real k;
  int Y[H];        //Responder from a specific study
  int N[H];        //Sample size from a  specific study
  int group[H];  //Vector of the studies
}

parameters {
  real <lower = 0> kappa;
  real <lower =0, upper =1> thetahc;
  real <lower =0, upper =1> thetacc;
  }

model {
      target +=   beta_lpdf(thetacc |kappa* thetahc, kappa*(1- thetahc));
      target +=  (group[1] == 1)*(binomial_lpmf(Y[1]|N[1], thetahc));
      target +=  (group[2] == 2)*(binomial_lpmf(Y[2]|N[2], thetacc));
      kappa ~ gamma(k, 1);
      thetahc ~ beta(0.5,0.5);
  }
```

Appendix 3.3: threshold type values under different borrowing methods and historical control type

Table threshold type values under different borrowing methods and historical control type

Method	HC Type	Threshold Type	HC(0.1)	HC(0.2)	HC(0.3)	HC(0.4)	HC(0.5)
Power Prior	Observation	Global	0.998	0.9975	0.9928	0.9805	0.9787
		Local	0.94689	0.9649	0.9612	0.9625	0.96444
		Regional	0.980625	0.98128	0.981001	0.97748	0.964776
	Simulation	Global	0.9964	0.9958	0.992	0.9859	0.97755
		Local	0.97281	0.97746	0.97791	0.97521	0.97696
		Regional	0.988175	0.98718	0.98565	0.98515	0.976625
Commensurate Prior (K = 1)	Observation	Global	0.99704	0.9946	0.99278	0.9914	0.9908
		Local	0.99704	0.98531	0.98085	0.9785	0.9769
		Regional	0.9971	0.99459	0.98253	0.98058	0.9772
	Simulation	Global	0.99628	0.99333	0.99158	0.98955	0.9877
		Local	0.99628	0.9853	0.98064	0.97859	0.97833
		Regional	0.99645	0.9947	0.9832	0.98006	0.97833
Commensurate Prior (K = 50)	Observation	Global	0.99945	0.9978	0.99336	0.98316	0.96135
		Local	0.95391	0.95424	0.96046	0.95823	0.96135
		Regional	0.98378	0.98236	0.9819	0.97853	0.96135
	Simulation	Global	0.99963	0.99863	0.99576	0.98831	0.97304
		Local	0.97443	0.97439	0.97196	0.97273	0.97304
		Regional	0.98939	0.98885	0.98948	0.98764	0.973
Commensurate Prior (K = 100)	Observation	Global	0.99968	0.99855	0.99495	0.98511	0.96346
		Local	0.9499	0.95216	0.9533	0.95581	0.96346
		Regional	0.98415	0.98354	0.98163	0.97873	0.96345
	Simulation	Global	0.99979	0.99913	0.99675	0.99035	0.97445
		Local	0.97596	0.97621	0.97471	0.97464	0.97445
		Regional	0.99039	0.99075	0.99166	0.98935	0.97445
Full Borrowing	Observation	Global	0.99985	0.99913	0.99615	0.987	0.952
		Local	0.9377	0.95288	0.9522	0.9474	0.952
		Regional	0.983925	0.985	0.987851	0.97898	0.962275
	Simulation	Global	0.9999	0.99949	0.9975	0.99305	0.97836
		Local	0.97981	0.97799	0.97539	0.97841	0.97836
		Regional	0.991201	0.99298	0.99305	0.98913	0.978025
No Borrowing	NA	NA	0.977925	0.97775	0.977201	0.974	0.977925