Parametric and Non-Parametric Models in Health Research: Analysis and Design

By

Yang Lei

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

_____

Chairperson Byron J. Gajewski, Ph.D.

_____

Matthew S. Mayo, Ph.D.

_____

Susan Carlson, Ph.D.

_____

Hung-Wen (Henry) Yeh, Ph.D

_____

Jo A. Wick, Ph.D.

Date defended: March 2, 2016

The Dissertation Committee for Yang Lei certifies that this is the approved version of the following dissertation

Parametric and Non-Parametric Models in Health Research: Analysis and Design

_____

Chairperson Byron J. Gajewski, Ph.D.

Date approved: March 2, 2016

ABSTRACT

Personalized medicine is emerging in both clinical practice and clinical trials. “Precision” medicine not only promises improved safety and efficacy but also lowered cost in clinical practice and clinical trials. In 2015, President Obama launched the Precision Medicine Initiative. This initiative requires close collaboration among clinicians, researchers, and biostatisticians. Enrichment design is an important strategy for increasing study efficiency in personalized medicine. Enrichment clinical trial designs involve identifying high-risk patients and choosing patients most likely to respond to treatment. In this dissertation, we have developed and applied parametric and non-parametric models to the following specific problems: 1) identifying high risk patients using Classification and Regression Trees (CART) model; 2) using Bayesian distributional approach and finite mixture normal model to improve trial efficiency in a rare endpoint scenario; 3) using dynamic linear normal model in enrichment trial designs with ordinal risk subgroups. The topics we discussed in this dissertation form a self-contained system within the enrichment clinical trial design structure. Identifying high risk patients and efficient statistical models are two major components in enrichment designs. However, the application of the models we discussed is far beyond the scope in this dissertation. Using CART to identify high risk subpopulations can overcome the incapacity of logistic regression models in revealing unknown interaction effects. A distributional approach using finite mixture normal model provides a flexible model design to fit strongly skewed data. Dynamic linear normal model in the enrichment trial design shows to be more efficient and robust compared to previously studied designs because it locally smoothes the trend. All these methods help us to accurately identify target patients and treat patients efficiently.

ACKNOWLEDGEMENT

First and foremost I want to extend my special appreciation and gratitude to my major advisor Dr. Byron J. Gajewski for the wonderful opportunity to join his research team. I appreciate all his contributions of time, ideas, and encouragements that make my Ph.D. experience productive and stimulating. The joy and enthusiasm he has for his research was contagious and motivational for me. I am thankful for the excellent example he has provided as a successful researcher and professor. I also appreciate his support in my career pursuit and advice on how to achieve a work-life balance. It is his wisdom and excellent supervision that make my Ph.D. journey a great joy.

I would thank my wonderful committee members, Dr. Matthew S. Mayo, Dr. Susan Carlson (Department of Dietetics and Nutrition), Dr. Hung-Wen (Henry) Yeh, and Dr. Jo Wick. Dr. Mayo has provided great support for my research and professional development. His advice is always pragmatic and broadens my horizons. His inspiring questions make every research topic extremely interesting. As the chair of the department, he provides many great opportunities for students to grow along with the department. These experiences are priceless. Dr. Susan Carlson provided me the great opportunity to involve in her research projects and helped me to understand my research topics comprehensively. I especially enjoyed the chances to attend KUDOS meetings and was greatly impressed by the wonderful team. Dr. Hung-Wen (Henry) Yeh always provides hit-the-key advices and insightful instructions. I also appreciate all the encouragement Dr. Yeh gave me when I had struggles in my study. Dr. Jo Wick always supports me with her passion, openness, and kindness. Her dedication in the statistical community services is something I need to improve in my future life. All my committee members' instructions and rigorous attitude toward research will benefit me in my whole life.

I want to thank my colleagues: Yu (Joyce) Jiang, Lili Garrard, Janelle Noel, Wei (Will) Jiang, Milan Bimali, Yi Zhong, and Xueyi Chen for their friendship and help in my study. All the discussions and group studies are unforgettable.

I want to thank all the professors and staff in the biostatistics department for their instructions and support in my study, research, and daily work.

At last, I would like to thank my dear parents and my husband for their support and my daughter for being a wonderful child and bringing a lot of happiness to my family. I feel extremely blessed to have all of you in my life!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 : Introduction

Modern medicine has evolved from broad-spectrum medical care to targeted therapeutics. Patient populations are heterogeneous. Characteristics vary between individuals, such as demographics, life style, environments, genetic variants, and etc. These varied characteristics can potentially modify the treatment effects on different individuals or subsets of patient populations. On one hand, people with some characteristics are more susceptible to certain disease. It is important to identify the target populations. The first step of quality care is to clearly define the target population according to variables such as age, gender, genetics, specific medical data, and etc. Identification of patients at higher risk of a certain disease or disease stage can help us to use resources efficiently and improve patients' quality of life. On the other hand, quality care requires identifying the most appropriate treatment strategies for the different populations. Personalized medicine, which aims to match patient subpopulation to the most beneficial treatment, is more ethical and efficient. FDA has been pushing for personalized medicine for a long time. In January 2015, President Obama launched the Precision Medicine Initiative, including establishing a national database of the genetic and other data of one million people in the United Sates. This is a new research effort to tailor treatment and prevention strategies to individuals' unique characteristics (White House, 2015). Targeted therapeutics enables physicians to select treatments that improve patients' health status and reduce exposure to adverse effects.

The concept of personalized medicine not only promises to improve safety and efficacy in clinical practice, but also to lower health care cost through early-detection, prevention, accurate risk assessments and efficient care delivery (Jakka & Rossbach, 2013). Personalized medicine also has a potential to lower costs in clinical trials. Cost containment in clinical trials is another prominent concern in both private and public sectors. The average cost of clinical trials

across all therapeutic areas was around $4 million for Phase I trials, $13 million for Phase II trials, and $20 million for Phase III trials (U.S. Department of Health and Human Services, 2014). Phase III trials is the most expensive component among the pre-approval trial stages, yet the combined success rate at Phase III and submission has fallen to about 50% in recently years (Arrowsmith, 2011). Enrichment on subpopulations that may be more responsive to treatments can improve the chance of trial success. In 2012 FDA issued a draft guidance to facilitate enrichment designs (U.S. Food and Drug Administration, 2012). The purposes of enrichment designs include:1) decreasing heterogeneity; 2) identifying high-risk patients; and 3) choosing patients most likely to respond to treatment (U.S. Food and Drug Administration, 2012).

Statistical model choice in the health care arena need to accommodate the evolution of personalized medicine. By definition, a statistical model is generated from observed or experimental data and is a simplification of reality. Statistical models can be used for description, prediction, or causal analysis (Maathuis, 2007). These usages are not necessarily mutually exclusive (Maathuis, 2007). What good are models and what models are good (Schnerider, 1993)? These are two fundamental questions in statistical research. Does the model describe the reality concisely? Does the model make predictions accurately? Does the model have strong capability to identify causal relationships? These naturally become the criteria to determine whether the model is good or not.

This dissertation is trying to explore statistical models, using both parametric and non-parametric approaches, under the personalized medicine framework. This dissertation is composed of three individual studies: the first study discussed methods to identify high-risk subpopulations; the second study focused on selecting efficient statistical models in clinical trials using a fixed Bayesian design; and the third study explored methods to identify the most

beneficial subpopulation in clinical trials, under a Bayesian enrichment design framework. These three studies form a self-contained research topic within the personalized medicine paradigm. The intrinsic relationship of these three studies can be described by the following vendor diagram:

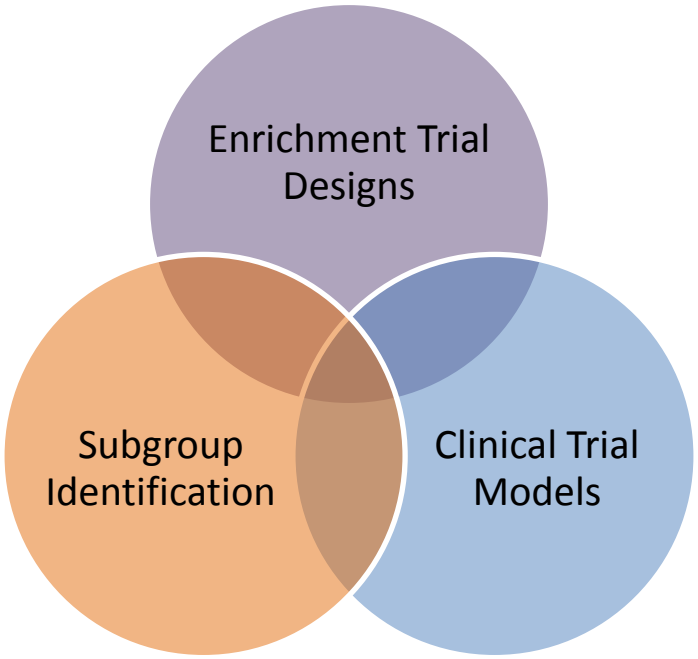

Figure 1.1: The relationship of the three individual studies

In the following chapters, parametric or non-parametric models will be developed and evaluated. In chapter 2, traditional logistic regression model and classification and regression tree (CART) model are compared, and applied to a study to identify high-risk populations in alternative tobacco products use. Logistic regression models are the most commonly used method to identify important factors for binary responding variables. Logistic regression model was developed by David Cox in 1958 (Cox, 1958). Dr. Frank E. Harrell's book "Regression Modeling Strategies: with Application to Linear Models, Logistic Regression, and Survival

Analysis" (Harrell, 2001) is the mostly cited book in logistic regression model selection. In Chapter 2, a model selection strategy recommended by Dr. Harrell is applied to narrow down potential models and then AIC is used as the criterion to determine the final model. Data is divided into independent training and validation samples. The final model is built upon the training data. A hold-out validation is performed to analyze the goodness-of-fit of the final regression model, and to check whether the model's predictive ability deteriorates substantially when applies to the validation data. Classification and regression trees (CART) model is another strategy to identify target subpopulations. CART is a data-driven, machine learning method that does not assume any specific form of the model such as individual variables, interaction relationships, and etc. In problems that no priori specified interaction term exists, CART is informative, efficient, and straightforward. This chapter compares and discusses the characteristics of these two methods. All analysis is done using SAS version 9.3 and SAS Enterprise Miner version 12.3.

In the third chapter, a model selection problem is discussed in a fixed Bayesian clinical trial design framework. The specific issue is that clinicians are interested in a binary outcome while the data is collected in a continuous form. Preterm birth (gestation age <37 weeks) and earliest preterm birth (gestation age <34 weeks) are clinically important outcomes while gestation age is collected in a continuous form. Docosahexaenoic acid (DHA) supplementation is a provocative strategy to reduce early preterm delivery (Carlson et al., 2013). Preterm birth and earliest preterm birth are rare events so the distribution of gestational age has substantially negative skewness. The traditional method is dichotomizing gestational age and then using binomial distribution for the dichotomized outcomes. It is widely accepted that pre-dichotomizing loses information and results in reduced power. Distributional method is another

way to gain inference on dichotomized outcomes while retaining statistical power from a continuous distribution (Peacock, Sauzet, Ewings, & Kerry, 2012). Gestational age data has substantially negative skewness therefore a transformation as log(C-x), where x is gestational age data and C is a constant great than x, is a reasonable transformation that normalizes the data (Hoaglin, Mosteller, & Tukey, 1983). However, this transformation could introduce a big bias. A three-component finite mixture normal model is proposed in this chapter because this distribution fits the gestational age data well. In this chapter, dichotomized model, logarithmic transformation model, and finite mixture normal model are compared through simulations. In simulation studies, we can tune and obtain desirable operating characteristics such as Type I Error rate, and evaluate models through Mean Squared Error (MSE) and power. The three models are applied to two completed clinical trials and the results are compared by checking their bias and standard deviations. In this chapter, simulations and analysis are done using Openbugs and R 3.1.1.

Chapter 4 discusses a topic that is a combination of identifying target populations and efficient clinical trial designs, namely enrichment trial designs. Classifying patient populations into different risk levels not only helps us to understand the target population, but also introduces opportunities to improve the trial efficiency, thus lower the costs of clinical trials. This chapter starts with a scenario that subgroup populations can be classified and ordered by their risk levels. This chapter continues to examine the effect of DHA supplementation on preterm birth rate, with subgroup effects specifically discussed. To improve the power in enrichment clinical trials, this chapter applies two approaches: using informative priors and selecting powerful models. Informative priors are obtained through meta-analysis that includes nine clinical trials across the world. This chapter compares four different statistical models: 1) logistic model which is not

flexible and assumes a constant relationship between log odds and risk levels; 2) independent model in which no information borrowing across subgroups but the model still borrows information from the informative prior distribution; 3) hierarchical model which borrows information from informative priors and across groups, assuming exchangeability; and 4) dynamic linear model which borrows information from informative priors and across groups, not assuming exchangeability. Overall Type I error rate is calibrated in simulations and then subgroups' Type I error rate and power, including the power to capture the most benefited group, are compared. Simulations are done using R 3.2.2 and Openbugs.

The dissertation concludes in Chapter 5 with summary and future work.

Chapter 2 : An Application in Identifying High-risk Populations in Alternative Tobacco Product Use Utilizing Logistic Regression and CART: a Heuristic Comparison

Yang Lei, Nikki Nollen, Jasjit S. Ahluwahlia, Qing Yu, and Matthew S. Mayo

Lei Y, Nollen N, Ahluwahlia JS, Yu Q, Mayo MS. An Application in Identifying High-risk Populations in Alternative Tobacco Product Use Utilizing Logistic Regression and CART: a Heuristic Comparison. *BMC Public Health* 2015; 15:341.

Abstract

Background: Other forms of tobacco use are increasing in prevalence, yet most tobacco control efforts are aimed at cigarettes. In light of this, it is important to identify individuals who are using both cigarettes and alternative tobacco products (ATPs). Most previous studies have used regression models. We conducted a traditional logistic regression model and a classification and regression tree (CART) model to illustrate and discuss the added advantages of using CART in the setting of identifying high-risk subgroups of ATP users among cigarettes smokers.

Methods: The data were collected from an online cross-sectional survey administered by Survey Sampling International between July 5, 2012 and August 15, 2012. Eligible participants self-identified as current smokers, African American, White, or Latino (of any race), were English-speaking, and were at least 25 years old. The study sample included 2,376 participants and was divided into independent training and validation samples for a hold out validation. Logistic regression and CART models were used to examine the important predictors of cigarettes + ATP users.

Results: The logistic regression model identified nine important factors: gender, age, race, nicotine dependence, buying cigarettes or borrowing, whether the price of cigarettes influences the brand purchased, whether the participants set limits on cigarettes per day, alcohol use scores, and discrimination frequencies. The C-index of the logistic regression model was 0.74, indicating good discriminatory capability. The model performed well in the validation cohort also with good discrimination (c-index=0.73) and excellent calibration (R-square=0.96 in the calibration regression). The parsimonious CART model identified gender, age, alcohol use score, race, and discrimination frequencies to be the most important factors. It also revealed interesting

partial interactions. The c-index is 0.70 for the training sample and 0.69 for the validation sample. The misclassification rate was 0.342 for the training sample and 0.346 for the validation sample. The CART model was easier to interpret and discovered target populations that possess clinical significance.

Conclusion: This study suggests that the non-parametric CART model is parsimonious, potentially easier to interpret, and provides additional information in identifying the subgroups at high risk of ATP use among cigarette smokers.

Keywords: Survey sampling, stratified samples, logistic regression, CART, partial interaction

2.1 Background

Recent years have witnessed increased tobacco control policies at both the state and national level (American Nonsmokers' Rights Foundation, 2013; Congress, 2009; Orzechowski & Walker, 2011). Most of these efforts are aimed at cigarette smoking (American Nonsmokers' Rights Foundation, 2013). The net effects of these policies include decreased cigarettes consumption, as well as a shift in the type of tobacco products used (CDC, 2012; Kasza, Bandal-Travers, & O'Connor, 2014). The use of alternative forms of tobacco products (ATPs), such as large cigars, little cigars, cigarillos, pipes, hand-rolled cigarettes, smokeless tobacco, and hookahs, are increasing in prevalence (Campbell, Bozec, McGrath, & Barrett, 2012; McGrath, Temporale, Bozec, & Barrett, 2011). About 8%-38% of U.S. daily smokers and as many as 44% of non-daily smokers (smoke on some but not all days) are ATP users (Backinger et al., 2008; Bombard, Pederson, Nelson, & Malarcher, 2007; Campbell et al., 2012; Kasza et al., 2014; McGrath et al., 2011; Popova & Ling, 2013), defined as anyone who uses cigarettes and alternative forms of tobacco. These tobacco products have been promoted as less addictive and less harmful than cigarettes (Jolly, 2008; Page & Evans, 2003). Nevertheless, data suggest that use of these products could be associated with higher nicotine dependence and may contribute to increased risks for diseases caused by tobacco, such as cancer and heart disease (Djordjevic & Doran, 2009).

It is of utmost importance to identify individuals who are at high risk of using both cigarettes and ATPs. Research subjects in previous studies have been predominately White (Backinger et al., 2008; Bombard, Pederson, Koval, & O'Hegarty, 2009; Bombard et al., 2007; Richardson, Xiao, & Vallone, 2012) and most existing studies have used traditional regression

approaches to identify important factors associated with ATP use. Although regression methods can test a priori specified interaction effects, it lacks the ability to capture unspecified, complex inter-relationships across factors. Classification and Regression Trees model (CART) can address these limitations by revealing unspecified inter-relationships through an easily interpretable tree diagram. Few studies have applied CART modeling to tobacco research (Piper, Loh, Smith, Japuntich, & Baker, 2011; Swan, Javitz, Jack, Curry, & McAfee, 2004). In this paper we used data from a cross-sectional survey of smokers and conducted the most commonly used logistic regression method and relatively underused CART method, and described the strength and limitations of these two statistical approaches in identifying cigarette smokers at highest-risk for ATP use.

2.2 Methods

2.2.1 Study population

The data was collected through a cross-sectional survey administered through an online panel survey service, Survey Sampling International (SSI), between July 5, 2012 and August 15, 2012. Ethical approval was granted by the University of Minnesota Institutional Review Board. Participants were presented with a written informed consent page prior to completing the screener. Only participants who indicated their consent were directed to the study questions. Eligible participants self-identified as current smokers, African American, White, or Latino (of any race), were English-speaking, and were at least 25 years old. The study sample contained 2,376 participants balanced by the three racial/ethnic groups across smoking frequencies (daily and nondaily smoking): 794 African Americans, 786 Latinos, and 796 whites. Among them, 1,220 participants (51.35%) were cigarettes + ATP users who used both cigarettes

and other tobacco products and 1,156 (48.65%) were cigarettes-only users. Variable domains in this study included: demographics, tobacco characteristics, cost concerns, harm reduction efforts, and psychosocial variables. There was minimal missing data, about 4.3% subjects were missing one variable (income), and therefore imputation was not necessary. Chi-square tests were used to test the unadjusted effects of categorical variables and T-tests were used to test continuous variables (Table 2.1).

2.2.2 Training and validation data sets

The large sample size allowed for the use of a hold-out validation to obtain independent training and validation data sets (Harrell, 2001; Larson, 1931; Mosteller & Wallace, 1963; Refaeilzadeh, Tang, & H., 2008; Ruggeri, Kenett, & Faltin, 2008). The data was partitioned by random sampling, stratifying by cigarettes + ATP use and race/ethnicity to ensure the balance we designed. Training sample contained 1,584 participants (two thirds of the sample) and was used to derive the model. The remaining data contained 792 participants (one third of the sample) and were used to evaluate the predictive ability of the final model. The training and validation samples were compared to ensure the differences between the two were negligible (Table 2.2).

2.2.3 Analysis

Logistic regression. Logistic regression is a traditional way to identify important factors for binary outcomes. The Akaike Information Criterion (AIC) is widely recommended as a model selection criterion (Harrell, 2001). To avoid the technical difficulty of comparing AICs from all possible variable combinations, we followed a model selection strategy recommended by Frank E. Harrell to trim the potential models (Harrell, 2001) and then picked the minimal AIC model from the potential models as the final model. The selection process started with all

potential factors. Predicted values from the logistic regression were then regressed on all covariates, with the model explaining 100% of the variance. Backward selection based on R^2 was used to select a parsimonious set of variables. The contribution of each covariate in the multivariable model was ranked, and variables with the smallest contribution to the model were sequentially eliminated. This iterative process continued until further variable elimination led to a greater than 5% loss in model prediction, as compared with the initial model. The remaining covariates comprised the parsimonious model and explained over 95% of the variance of the full model. Finally, we compared AIC values of neighborhood models around the model we obtained in the last step and the minimum AIC model was selected as the final model. This selection strategy supports inclusion of only variables that provide incremental prognostic value, avoids over-fitting, and maximizes the potential usefulness of the model. Besides this model selection strategy, we examined backward selections based on p-value with 0.15 as the threshold to enter and 0.05 as the threshold to stay in the model. Both approaches identified the same model.

Predicted values using the model estimates from the training cohort were generated for the validation cohort and the c-index was then calculated based on the proportion of concordance. The predicted values were ranked and cut into deciles. The calibration plot was graphed comparing the average predicted probabilities with the observed average probabilities. A calibration regression on observed mean probabilities was performed using predicted mean probabilities to check the strength of correlation between the predicted and the observed average probabilities across deciles.

Classification and Regression Tree (CART) Model. Although the logistic regression model provides knowledge of important profile characteristics, it lacks the ability to identify unknown, and therefore, unspecified interaction effects. The interpretation of parameter

estimates is based on the fact of controlling for all other covariates. To address these problems, we built Classification and Regression Tree models (CART) in SAS Enterprise Miner version 12.3 (Gordon, 2013; Loh, 2011). CART is a nonparametric method that identifies mutually exclusive and exhaustive subgroups. Members within each subgroup share the same characteristics that influence the probability of belonging to the interested response group (Lemon, Roy, Clark, Friedmann, & Rakowski, 2003). CART produces a model structure that resembles an upside-down tree. The tree starts with the parent node, and the parent node contains the entire population. The CART algorithm examines all possible independent variables according to a predetermined splitting rule and divides the parent node into two child nodes; the child nodes can be further divided into more child nodes. There are many splitting rules, and they all begin with defining the impurity of a node (Lemon et al., 2003). The impurity function measures the extent of difference/similarity for a node containing data points from possible different classes. A node that has no impurity would have no variability (e.g. all cigarettes-only smokers, or all cigarettes + ATP smokers). The highest impurity is achieved when p(k|t)=0.5, where p(k|t) is defined as the conditional probability of belonging to class k given in node t. Although the impurity functions may vary, all splitting rules select the split that has the largest difference between the impurity of the parent node and a weighted average of the impurity of the two child nodes (Lemon et al., 2003). The Gini splitting rule was recommended most for binary outcomes (Gordon, 2013). This rule maximizes the following improvement of impurity function (Timofeev, 2004):

$$\underset{x_j \le x_j^R, j=1,\dots,M}{\arg max} \left[-\sum_{k=1}^{K} p^2(k|t_p) + P_l \sum_{k=1}^{K} p^2(k|t_l) + P_r \sum_{k=1}^{K} p^2\,(k|t_r) \right]$$

p(k|t): conditional probability of dependent variable= k given node t

subscript p: parent node

subscript r: child right node

subscript l: child left node

P_l: probability in the left child node

P_r: probability in the right child node (note: $P_l + P_r = 1$)

x_j^R: best splitting value of variable x_j

M: number of potential independent variables

K: level of dependent variables. For binary outcomes, K=2

The larger the value of the improvement in impurity function, the greater difference between the two child nodes with respect to the prevalence of the dependent measure. The CART procedure selects the independent variable and the splitting cutoff of the continuous independent variable to maximize the improvement at each step. The tree grows as child nodes are divided into more child nodes. The terminal nodes are where predictions and inferences are made.

It is clear that different samples would produce different trees. One common way to assess how different the trees could be is using training and validation samples. To facilitate comparisons, the same set of training and validation samples were used in logistic regression model and CART model. In CART model, misclassification rates from both the training sample and the validation sample were compared to ensure the model is stable.

The maximum tree with the minimum misclassification error was examined and the misclassification error graph showed that it contained insignificant nodes, which reduced the misclassification error marginally but increased the complexity greatly. A popular stopping strategy was applied by predefining the minimum number of points in the terminal node to control the size of the tree (Lemon et al., 2003). The minimum node size was set to be 10% of the training sample size or about 150 subjects in our study. Models were assessed to identify a parsimonious tree that produces non-trivial results with acceptable misclassification rates.

2.3 Results

2.3.1 Logistic regression model.

The final model consisted of nine variables (Table 2.3). Males had the strongest association with being a cigarettes + ATP user vs. cigarettes-only user (adjusted OR 2.66, 95% CI 2.12 – 3.33). African Americans and Latino were more likely to be cigarettes + ATP users compared to whites (adjusted OR 1.58, 95% CI 1.21 – 2.07 and adjusted OR 1.52, 95% CI 1.16 – 1.99, respectively). Individuals with higher nicotine dependence were more likely to be cigarettes + ATP users (adjusted OR 1.51, 95% CI 1.20 – 1.90). Participants who buy their cigarettes were less likely to be cigarettes + ATP users compared to those who borrow cigarettes from others (adjusted OR 0.617, 95% CI 0.49 – 0.78). Individuals who were more sensitive to the price of cigarettes were more likely to be cigarettes + ATP users (adjusted OR 1.43, 95% CI 1.14 -1.79). Individuals who set limit on cigarettes per day were more likely to be cigarettes + ATP users (adjusted OR 1.30, 95% CI 1.04 – 1.62). Individuals with higher alcohol scores were more likely to be cigarettes + ATP users (adjusted OR 1.10, 95% CI 1.064-1.145). Older people were less likely to use cigarettes + ATPs (adjusted OR 0.97, 95% CI 0.96-0.98). Higher

discrimination scores were associated with higher probability of using cigarettes + ATPs (adjusted OR 1.03, 95% CI 1.01 – 1.05). The C-index of the final model was 0.74, indicating good discriminatory capacity (Figure 2.1).

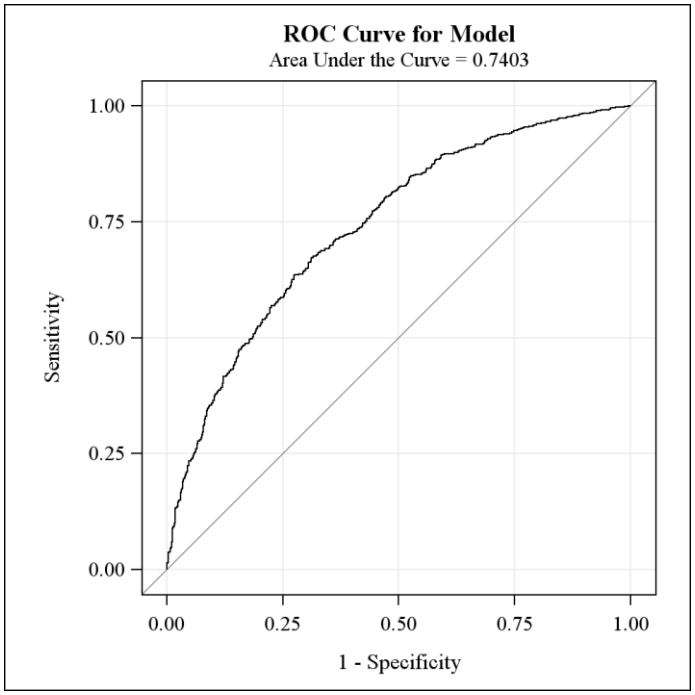

Figure 2.1: Results from logistic regression on the training sample: ROC curve

2.3.2 Model validation

Participants were similar in terms of all profile characteristics (Table 2.2), except that participants in the validation cohort smoked about 1 cigarette per day less than the training cohort (10 vs. 9, p=0.009).The model performed well in the validation cohort with good discrimination (c-index=0.73) and excellent calibration with an intercept of 0.018 (p-value for difference from 0 = 0.65) and a slope=0.96 (p-value for difference from 1= 0.58). The R-square for the calibration regression was 0.96 and the Pearson correlation coefficient was 0.98 (p-value<0.0001) (Figure 2.2).

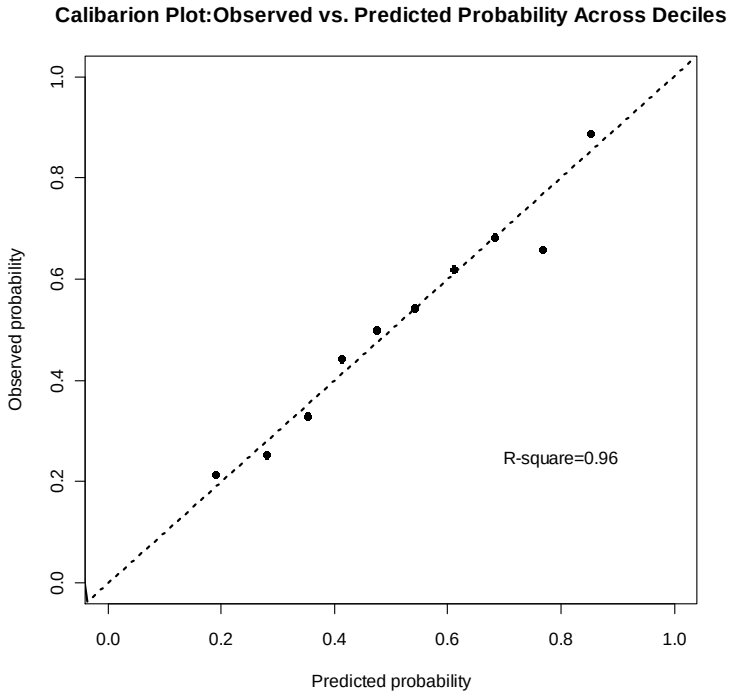

Figure 2.2: Results of calibration from validation sample: calibration plot

2.3.3 Classification and Regression Tree (CART) Model.

Figure 3 shows the final tree results using the stopping rule of minimum node size no less than 150. The same independent training and validation samples were used as in the logistic regression. The misclassification rate was 0.342 for the training sample and 0.346 for the validation sample. The C-index was 0.70 for the training sample and 0.69 for the validation sample.

Males were more likely to be cigarettes + ATP users, especially when they were moderate to heavy drinkers (alcohol use score> 2). A male with a 3 or higher alcohol score had 73.5% probability of being a cigarettes + ATP user. Females were less likely to be cigarettes + ATP users, especially when they were older. Female participants aged 46 or older had a 29.0%

probability of being cigarettes + ATP users. Among females age 45 years or younger, Latino and African Americans were more likely to be cigarettes + ATP users compared to whites. 37.2% of White females aged 45 years or younger were cigarettes + ATP users. Latino and African American females aged 45 or younger, who also experienced greater discrimination were more likely to be cigarettes + ATP users, about 62.2% probability if their discrimination score was greater than 6 (Figure 2.3). Interestingly, age, race, and discrimination effects that impacted female participants did not play important roles for males. Alcohol scores increased the risk of cigarettes + ATP use for males but were not important for females. These indicated informative interaction patterns to examine the profile characteristics of cigarettes + ATP users.

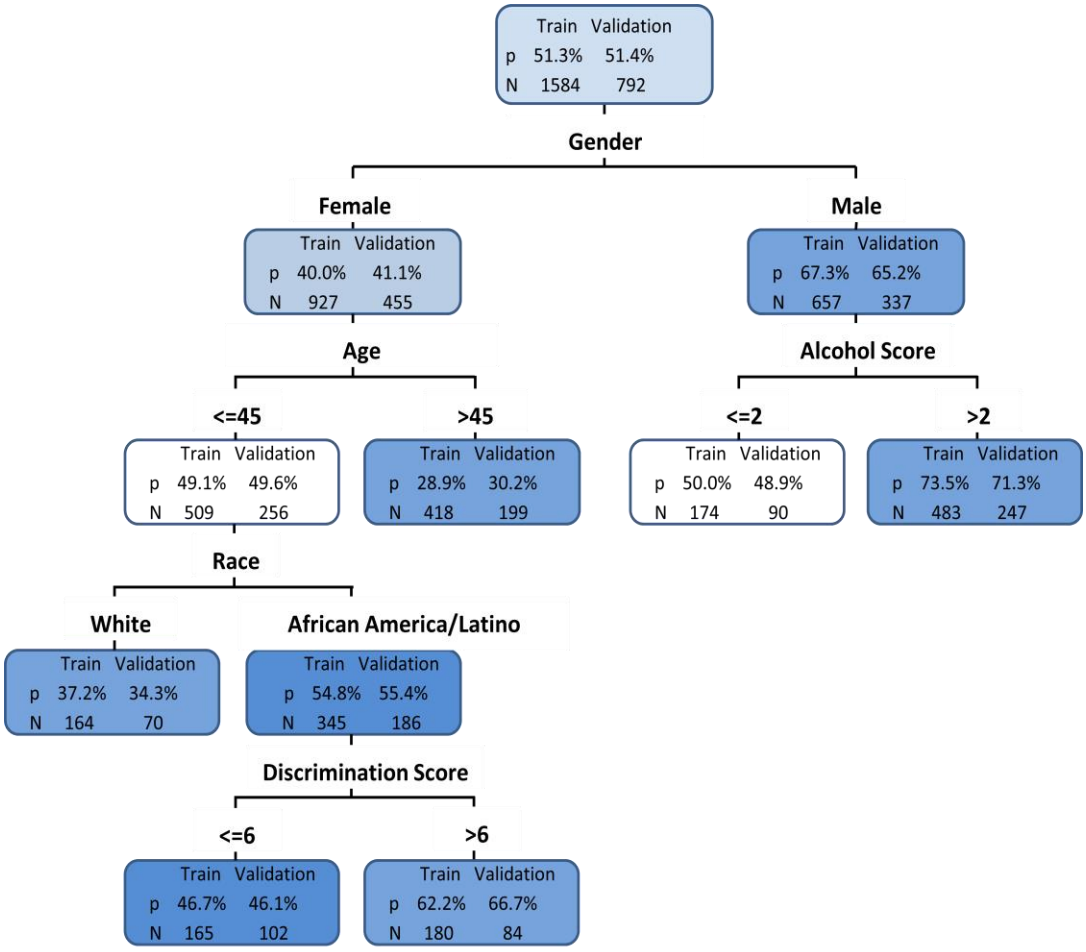

Figure 2.3: Classification and Regression Tree model for predicting Cig + ATP users

2.4 Limitations

A hold-out validation strategy was used in this study to obtain independent training and validation datasets. The reduced data can result in an enlarged variance. Although this method is reasonable in this study because the sample size is large, other validation strategies, such as k-fold cross validation, which uses overlapped training data, may achieve more accurate performance estimation. We used a method suggested by Harrell (2001) (Harrell, 2001) to trim potential models and then compared AIC of these potential models to obtain the final model. Other model selection strategies, such as LASSO and ridge regression were not compared with this method.

2.5 Discussion

The CART model identified the five most important factors: gender, alcohol scores, age, race, and discrimination scores. The logistic regression model identified nine variables: the same five as the CART model, and additionally, whether the participant buys or borrows cigarettes from others, whether the participant limits cigarettes per day, price influences, and nicotine dependence. Therefore, the logistic regression model expanded the variable pool from the CART model.

The logistic regression model results in higher C-index than the CART model (0.74 versus 0.70 for the training sample and 0.73 versus 0.69 for the validation sample). However, the C-index from the CART model was not directly comparable to that in the logistic regression model because the classifiers varied across different subgroups in the CART model due to partial interaction effects. On the other hand, logistic regression models lack the ability to identify unspecified, complex inter-relationships between factors. In studies where interaction effects are

unclear, it is impractical to test all potential interaction effects in logistic regression models. However, there might be potential inter-relationships, especially among demographic, psychosocial, and economic factors. Even if the logistic regression model achieves good model fit, we could still miss interaction effects that are significant to clinical practice. CART analysis is efficient to address these problems and it is easy to perform with available statistical software. It has great flexibility of building a model that can be easily interpreted through pictorial illustration, without pulling in too much complexity. CART can be considered as complementary to logistic regression models and the result from CART revealed clearly classified high-risk populations of ATP use among cigarette smokers.

2.6 Conclusions

The growing trend of ATP use could ultimately cut down the effect of tobacco control efforts that we have seen in recent years. Compared to the traditional logistic regression model, our CART model is more straightforward in classifying individuals at high risk of using cigarettes + ATPs. This model identified fewer factors associated with cigarettes + ATP use and revealed partial interactions that are not easy to find in logistic regression, thus provided clearer direction for identification and treatment in clinical practice. In general, the CART methodology can be used to classify high risk or at need groups for identification for treatment protocols including behavioral interventions.

Table 2.1: Univariate Differences between Smokers who Use Cigarettes in Combination with Alternative Tobacco Product (Cigarettes + ATP) Compared to those who use Cigarettes Only

	Cigarettes + ATP (n=1,220)	Cigarettes Only (n=1,156)	p value
Demographics			
Male	662 (27.9%)	332 (14.0%)	< 0.001
Age (±SD)	40.24 ± 11.64	45.85 ± 12.62	< 0.001
Race			< 0.001
African American	436 (18.4%)	358 (15.1%)	
Latino	455 (19.1%)	331 (13.9%)	
White	329 (13.8%)	467 (19.7%)	
Education, % college graduate or higher	474 (19.9%)	364 (15.3%)	< 0.001
Income, % < $1800/month	480 (20.2%)	463 (19.5%)	0.725
Tobacco Characteristics			
Smoking status (%)			< 0.001
Nondaily	673 (28.3%)	528 (22.2%)	
Daily light (1-10 cpd)	259 (10.9%)	319 (13.4%)	
Daily heavy (11+ cpd)	288 (12.1%)	309 (13.0%)	
Menthol smoker	737 (31.0%)	623 (26.2%)	0.001
Cigarettes per day, mean (±SD)	9.30 ± 8.70	10.14 ± 8.52	0.017
Time to first cigarette, % within 30 minutes of waking	720 (30.3%)	629 (26.5%)	0.024
24 hour quit attempts in last 12 months, mean (±SD)	5.50 ± 9.53	5.94 ± 11.79	0.451
Cost			
Price of cigs influenced them to smoke less, % yes	726 (30.6%)	644 (27.1%)	0.061
Price of cigs influenced where they buy cigs, % yes	840 (35.4%)	826 (34.8%)	0.166
Price of cigs influenced the brand they buy, % yes	590 (24.8%)	455 (19.1%)	< 0.001

	Cigarettes + ATP (n=1,220)	Cigarettes Only (n=1,156)	p value
Buy versus borrow cigs, % buy all cigs they smoke	683 (28.7%)	824 (34.7%)	< 0.001
Harm Reduction			
Trying to cut down on cigs smoke, % yes	862 (36.3%)	818 (34.4%)	0.955
Limit cpd to decrease health risk, % yes	596 (25.1%)	505 (21.3%)	0.012
Limit smoking in last year to decrease health risks, % always or often	360 (15.2%)	356 (15.0%)	0.494
Psychosocial			
Depression score, mean (±SD)[a]	2.14 ± 1.83	1.80 ± 1.84	< 0.001
Alcohol score, mean (±SD)[b]	4.64 ± 3.10	3.30 ± 2.98	< 0.001
Discrimination score, mean (±SD)[c]	8.28 ± 6.72	5.85 ± 5.66	< 0.001

[a] Scores range from 0-6 with scores of 3 or higher indicating possible depressive symptoms

[b] Scores range from 0-12 with scores of $\geq$4 for men and $\geq$3 for women indicating possible alcohol misuse

[c] Scores range from 0-25 with higher scores indicating greater frequency of discrimination in daily life

Table 2.2: Univariate Differences between training sample and validation sample

	Training (n = 1584)	Validation (n = 792)	P value
Demographics			
Male	657 (27.7%)	337 (14.2%)	0.617
Age (±SD)	42.94 ± 12.39	43.03 ± 12.5	0.880
Race			0.997
African American	530 (22.3%)	264 (11.1%)	
Latino	524 (22.1%)	262 (11.0%)	
White	530 (22.3%)	266 (11.2%)	
Education, % college graduate or higher	550 (23.1%)	288 (12.1%)	0.430
Income, % < $1800/month	614 (25.8%)	329 (13.8%)	0.192
Tobacco Characteristics			
Smoking status (%)			0.263
Nondaily	799 (33.6%)	402 (16.9%)	
Daily light (1-10 cpd)	373 (15.7%)	205 (8.6%)	
Daily heavy (11+ cpd)	412 (17.3)	185 (7.8%)	
Menthol smoker	899 (37.8%)	461 (19.4%)	0.500
Cigarettes per day, mean (±SD)	10.03 ± 9.03	9.06 ± 7.69	0.009
Time to first cigarette, % within 30 minutes of waking	900 (37.9%)	449 (18.9%)	0.953
24 hour quit attempts in last 12 months, mean (±SD)	5.54 ± 9.87	6.00 ± 11.93	0.454
Cost			
Price of cigs influenced them to smoke less, % yes	920 (38.7%)	450 (18.9%)	0.557
Price of cigs influenced where they buy cigs, % yes	1100 (46.3%)	566 (23.8%)	0.311
Price of cigs influenced the brand they buy, % yes	685 (28.8%)	360 (15.2%)	0.306
Buy versus borrow cigs, % buy all cigs they smoke	1004 (42.3%)	503 (21.2%)	0.952

	Training (n = 1584)	Validation (n = 792)	P value
Harm Reduction			
Trying to cut down on cigs smoke, % yes	1119 (47.1%)	561 (23.6%)	0.924
Limit cpd to decrease health risk, % yes	730 (30.7%)	371 (15.6%)	0.727
Limit smoking in last year to decrease health risks, % always or often	476 (20.0%)	240 (10.1%)	0.899
Psychosocial			
Depression score, mean (±SD)[a]	1.99 ± 1.86	1.96 ± 1.82	0.683
Alcohol score, mean (±SD)[b]	4.02 ± 3.16	3.93 ± 3.03	0.494
Discrimination score, mean (±SD)[c]	7.03 ± 6.30	7.23 ± 6.44	0.460

[a] Scores range from 0-6 with scores of 3 or higher indicating possible depressive symptoms

[b] Scores range from 0-12 with scores of $\geq$4 for men and $\geq$3 for women indicating possible alcohol misuse

[c] Scores range from 0-25 with higher scores indicating greater frequency of discrimination in daily life

Table 2.3: Results from logistic regression on the training cohort: estimates and Odds Ratios

Parameter	Estimate	Odds Ratio	95% CL for OR	P-value
Intercept	-0.2617	NA	NA	0.3497
Age	-0.0265	0.974	(0.964, 0.983)	<.0001
Male	0.9766	2.655	(2.118, 3.329)	<.0001
Buy vs. Borrow	-0.4832	0.617	(0.486, 0.783)	<.0001
Alcohol	0.0986	1.104	(1.064, 1.145)	<.0001
Price influenced the brand they buy	0.3579	1.430	(1.144, 1.788)	0.0017
African American vs. white	0.4576	1.580	(1.208, 2.066)	0.0008
Latino vs. white	0.4170	1.517	(1.155, 1.994)	0.0028
Discrimination	0.0259	1.026	(1.007, 1.045)	0.0065
Time to first cig less than 30 min	0.4100	1.507	(1.197, 1.897)	0.0005
Limit cigarettes per day	0.2612	1.299	(1.041, 1.619)	0.0203

Chapter 3 : Comparison of Dichotomized and Distributional approaches in a DHA Supplementation Clinical Trial Design: a Fixed Bayesian Design

Yang Lei, Susan Carlson, Lisa Yelland, Maria Makrides, Robert Gibson, & Byron J. Gajewski

(Submitted to Journal of Applied Statistics)

Abstract

This research was motivated by our goal to design an efficient clinical trial to compare two doses of docosahexaenoic acid supplementation for reducing the rate of earliest preterm births and/or preterm births in a fixed Bayesian design framework. Dichotomizing continuous gestational age data and analyzing the data using a classic binomial distribution will result in a loss of information and reduced power. A distributional approach is an improved strategy for dichotomizing continuous data while retaining statistical power from the continuous distribution. However, appropriate distributions that fit the data properly, particularly in the tails, must be chosen, especially when the data are skewed. A recent study proposed log transformation and use of a normal distribution for the transformed gestational age data. We propose a three-component normal mixture model and introduce separate treatment effects at different components of gestational age. We evaluate operating characteristics of clinical trial designs comparing this mixture model with a beta-binomial model and a normal model applied to the log transformed data through simulation. We also apply these three methods to data from two completed clinical trials from the USA and Australia. Finite mixture models are shown to have favorable properties in preterm births analysis but limited benefit for earliest preterm births analysis. Normal models on log transformed data have the largest bias and are not as efficient as finite mixture models. Therefore we recommend finite mixture model for preterm births study. Either finite mixture model or beta-binomial model is acceptable for earliest preterm births study.

Keyword: Bayesian, Normal mixture model, simulation, Dichotomization, preterm birth

3.1Introduction

In many circumstances, clinical researchers are interested in studying categorized outcomes using cutoff points despite continuous measurements being collected. It has been widely accepted that dichotomizing continuous data prior to analysis results in a loss of information and reduced power (Altman & Royston, 2006; Deyi, Kosinski, & Snapinn, 1998; Peacock et al., 2012). A distributional approach can be used to dichotomize continuous data while retaining the statistical power from the continuous distribution (Peacock et al., 2012). Peacock et al. (2012) described the use of the distributional method and showed the good performance of this parametric approach under standard normal distributional assumptions (Peacock et al., 2012). Sauzet et al. (2015) further discussed the distributional approach when the outcome is skewed and proposed a skew-normal distributional method for dichotomization (Sauzet, Ofuya, & Peacock, 2015). They used a logarithm transformation to normalize negatively skewed gestational age data and then applied the skew-normal distributional method under the Frequentist framework. They acknowledged that no satisfactory transformation is available for gestational age data (Sauzet et al., 2015). Mixture models with different components might be a better choice for skewed outcomes such as gestational age, because they allow for greater flexibility in modeling heterogeneous populations (McLachlan & Peel, 2000), which largely explains the skewness of gestational age data.

Our research was motivated by our goal to design an efficient clinical trial to compare two doses of docosahexaenoic acid (DHA) supplementation for reducing the rate of earliest preterm births (ePTB, gestation age<34 weeks) and/or preterm births (PTB, gestational age<37 weeks). Both endpoints have been evaluated in past studies (Makrides et al., 2010). The United States currently has a PTB rate of 11.4% (House, 2014) and babies born preterm are at increased

risk of immediate life-threatening health problems, as well as long-term complications and developmental delays (Gajewski, Reese, Colombo, & Carlson, 2016). Among preterm infants, those babies who are born the earliest (<34 weeks) are at greatest risk of complications. Although the overall PTB rates have decreased over time, the ePTB rates in the U.S. have decreased little since 1990 and the overall ePTB rates in the US for 2012 were 3.4% (Martin, Hamilton, Osterman, Curtin, & Mathers, 2013). These births impact overall infant mortality the most and result in much higher hospital costs than uncomplicated births (Russell et al., 2007). Docosahexaenoic acid (DHA) supplementation potentially provides a high yield, low risk provocative strategy to reduce early preterm delivery (Gajewski et al., 2016). We designed a Phase III clinical trial (randomized to low or high dose DHA, double-blind) to examine the efficacy of 1000 mg/day DHA supplementation to reduce the probability of earliest preterm births and/or preterm births compared to 200 mg/day, an amount recommended by the FAO/WHO for pregnant and lactating women and currently in many prenatal supplements. Our goal was to identify a powerful design that would provide an efficient estimate of the treatment effect.

Gestational age (GA) data will be measured in completed weeks/days and collected in a continuous form. The two clinically important endpoints of interest are: ePTB (GA<34 weeks) and PTB (GA<37 weeks). The traditional analysis approach is to dichotomize the continuous gestational age data using these cutoff points and to compare the probabilities of binary outcomes, using a chi-square test for example. Distributional methods compare the proportions below the cutoff points in continuous distributions (Gajewski et al., 2016; Peacock et al., 2012). Sauzet et al. (2015) proposed a skew-normal method and used normal distribution on the logarithmic transformed data (Sauzet et al., 2015). We propose a three-component normal

mixture model and apply the distributional approach directly. The aim of this study is to compare these three statistical methods under a fixed Bayesian design framework for a very rare endpoint (ePTB) and a less rare endpoint (PTB).

The remainder of this article is arranged as follows. In section 2, we describe three statistical models using the pre-dichotomizing and distributional methods separately. In section 3, we provide the simulation details under a fixed Bayesian clinical trial design framework and compare these three statistical methods in several realistic outcome scenarios. In section 4, we apply these three methods to data from two completed clinical trials, one in the USA and one in Australia. The results from the real data analysis are examined and compared. In section 5, we discuss the observations from the simulations and real data analysis and further investigate the rationale of these observations. In section 6, we discuss the limitations of this study. In section 7, we draw conclusions from our analysis and give suggestions to future studies.

3.2 Statistical models

Let $Y_j = (Y_{j1,\ldots}, Y_{jn_j})$ denote the continuous data of gestational age, where j denotes the treatment group assignment (j=c for participants in the control group and j=t for participants in the treatment group) and n_j denotes the sample size in the jth treatment group in a two-armed randomized clinical trial design. Let p_j denote the probability of ePTB or PTB in the jth treatment group.

The first method considered involves dichotomizing the data prior to modeling. We propose a beta-binomial model to simplify a Bayesian inference of $P(p_c > p_t|\text{data})$, denoting the posterior probability that control has a higher ePTB/PTB rate than treatment. Because the endpoints considered are rare, using a uniform prior or a beta (1,1) prior might induce non-

negligible bias. We therefore assume a very weak prior of p_j as beta (0.01, 0.01). Furthermore, the posterior mode is close to a classical Frequentist approach (i.e., Maximum Likelihood Estimator). Let $X_j = \sum_{i=1}^{n_j} I(Y_{ji} < 34 \text{ or } 37)$, where $I(x < y) = \begin{cases} 1, x < y \\ 0, x \geq y \end{cases}$ and n_j is the sample size in the jth treatment group (i=1,..., nj). The distribution of X_j is assumed to follow a binomial distribution: $X_j|p_j \sim \text{binomial}(n_j, p_j)$. The posterior distribution of $p_j|X_j \sim \text{beta}\left(X_j + 0.01, n_j - X_j + 0.01\right)$.

The second method considered is a distributional approach, where we will apply the transformation recommended by Sauzet et al. (2015) (Sauzet et al., 2015). First we take a logarithmic transformation of (45-GA) to normalize the data because we expect GA in weeks to be <45 and is negatively skewed, $Z_{ji} = \log(45 - Y_{ji}), i = 1, \ldots, n_j$, and then assume $Z_{ji} \underset{\sim}{i.i.d} N(\mu_j, \sigma_j^2)$. Since the logarithmic transformation is a continuous and monotonic transformation, this does not affect the proportion below a cut-point (Sauzet et al., 2015). The proportions of GA below 34 and 37 are translated into the proportions greater than log(45-34) =2.3979 and log(45-37) =2.0794 in the normal distribution $N(\mu_j, \sigma_j^2)$. We use non-informative conjugate priors for the parameters in the normal distribution: $N(0, 100^2)$ for μ_j and Gamma (0.001, 0.001) for $\frac{1}{\sigma_j^2}$. The posterior probability of ePTB or PTB $(p_j|Z_j)$ is calculated as $p_j|Z_j = \int_{2.3979 \text{ or } 2.0794}^{\infty} \phi(y|\mu_j, \sigma_j^2)\, dy$, where $\phi(y|\mu_j, \sigma_j^2)$ is a normal density function with posterior variance $\sigma_j^2|Z_j \sim IG(0.001 + \frac{n_j}{2} - \frac{1}{2}, 0.001 + \frac{1}{2}\sum_{i=1}^{n_j}(Z_{ji} - \overline{Z_j})^2)$, and posterior mean $\mu_j|Z_j, \sigma_j^2 \sim \text{Normal}\left(\frac{100^2 n_j \overline{Z}_j}{100^2 n_j + \sigma_j^2}, \frac{100^2 \sigma_j^2}{100^2 n_j + \sigma_j^2}\right)$.

The third method considered is another distributional approach using the finite normal mixture model. Peacock et al. (2012) showed the good performance of the parametric approach under traditional normal distributions (Peacock et al., 2012). We extend this approach here and propose a finite mixture normal model to allow for population heterogeneity. In this method, we apply a three-component normal mixture model derived from the North Carolina Detailed Birth Record (NCDBR) database with 336,129 observations in the final analysis: a three-component mixture of N(39.59, 0.96), N(38.26, 2.48), and N(33.29, 13.23) (Schwartz, Gelfand, & Miranda, 2010). The 95% CIs for the parameter estimates in this model show these estimates are reliable in this registry data. The first component has a mean of 39.59 (39.58, 39.61), and variance of 0.96 (0.95, 0.97). The second component has a mean of 38.26 (38.20, 38.32) and variance of 2.48 (2.42, 2.54). The third component has a mean of 33.29 (33.07, 33.51) and variance of 13.23 (12.78, 13.67) (Schwartz et al., 2010). Although we used fixed parameter estimates from a U.S. registry data, this model has unprecedented advantages in gestational age data analysis or clinical trial design, even for a different population. Firstly, the parameter estimates are derived from a huge registry data thus is representative and has generalizablility. Secondly, a three-component mixture normal model has its own flexibility to model similar but not exactly the same gestational age data from a different population by allowing various component weights. Thirdly, the three components are realistic and interpretable. The three components represent low, medium, and high-risk groups for PTB separately. We assume a unity prior for Δ_j (j=c,t), the mixture weights in the jth treatment group, and the three-component normal mixture model can be written as: $f(Y_{ji}|\Delta_j)=\Delta_{1j}\phi(Y_{ji}|39.59, 0.96) + \Delta_{2j}\phi(Y_{ji}|38.26, 2.48)+\Delta_{3j}\phi(Y_{ji}|33.29, 13.23)$, where $\phi\ (y|\mu, \sigma^2)$ denotes the density of y in a normal distribution with mean μ and variance σ^2, and $\Delta_{1j}, \Delta_{2j}, \text{and } \Delta_{3j}$ denote the mixture weights in the jth treatment group, with $\Delta_{1j} + \Delta_{2j} +$

$\Delta_{3j} = 1$. In this method, the posterior probability of ePTB or PTB ($p_j|Y_j$) is calculated as: $p_j|Y_j = \int_{-\infty}^{34 \text{ or } 37} f(y|\Delta_j)\, dy$. A more general approach would be to let each component's mean and variance be freely modeled. However, we found our approach was flexible and appropriate for our focus on the lower tail. Furthermore, fixing the components avoids some of the identifiability issues found in general mixture models (McLachlan & Peel, 2000).

3.3 Fixed Bayesian clinical trial design and simulation study

A previous Phase III trial comparing 600 mg DHA per day and placebo, Kansas University DHA Outcome Study [KUDOS], found an 85% reduction in ePTB with DHA supplementation (Carlson et al., 2013). Another Australian trial, DHA to Optimize Mother Infant Outcome [DOMInO] trial, which compared 800 mg DHA per day and placebo, found a 50% reduction in ePTB with DHA supplementation (Makrides et al., 2010). In both trials, ePTB was a secondary outcome (Carlson et al., 2013; Makrides et al., 2010). The primary aim of the current proposed Phase III randomized, double-blind trial is to test the hypothesis that ePTB and/or PTB is reduced by 1000 mg of DHA per day compared to 200 mg DHA per day. We performed a simulation study based on realistic response scenarios to investigate the operating characteristics of this fixed Bayesian clinical trial design.

3.3.1 Simulation methods

We simulated gestational age data using different true values of mixture weights (Δ_0) with resulting probabilities of ePTB or PTB close to probability scenarios we observed from our clinical trials (Carlson et al., 2013). In the beta-binomial model, we used simulation to generate the posterior distribution of $p_j|X_j \sim \text{beta}\,(X_j + 0.01, n - X_j + 0.01\ )$ for both treatment and control groups and calculated the probability of $p_c|X_c > p_t|X_t$. In the finite mixture model, we

used Markov Chain Monte Carlo (MCMC) to generate posterior distributions of Δ_j and the posterior probability $p_j|Y_j$ was calculated as: $p_j|Y_j = \int_{-\infty}^{34 \text{ or } 37} f(y|\Delta_j)\, dy$. In the logarithmic transformation method, we used Gibbs sampling to generate posterior distributions of μ_{jp} and σ_{jp}^2. The posterior probability $p_j|Z_j$ was calculated as $\int_{2.3979 \text{ or } 2.0794}^{\infty} \phi(y|\mu_{jp}, \sigma_{jp}^2)\, dy$. If $Pr(p_c > p_t|\text{data}) > \delta$, we counted this as a trial success. The posterior mean of $p_j|\text{data}$, $\widehat{p_j} = E(p_j|\text{data})$, was saved for each simulation in each of the three models. In all models, the expected estimated probability of ePTB or PTB, $E(\widehat{p_j})$ was calculated as the average of $\widehat{p_j}$ across simulations. V_j, the sample variance of $\widehat{p_j}$, was calculated as $\frac{\sum_{j=1}^{S}(\widehat{p_j} - E(\widehat{p_j}))^2}{S-1}$ for each treatment group, where S was the number of simulations. The MSE of $E(\widehat{p_j})$ was calculated as $\text{bias}^2 + \text{sample variance} = (E(\widehat{p_j}) - P_{j,0})^2 + V_j$, with $P_{j,0}$ denoting the true probability of ePTB or PTB in the jth treatment group.

To mimic situations for ePTB in future trials, we simulated 5 scenarios with varying treatment effects: no effect (3 vs. 3%, difference=0), very small (3 vs. 2%, difference=1%), small (3 vs. 1%, difference=2%), medium (3 vs. 0.5%, difference=2.5%) and large (4 vs. 1%, difference=3%) based on our previous clinical trial results (Carlson et al., 2013). To mimic situations for PTB in future trials, we simulated another 5 scenarios: no treatment effect (8 vs. 8%, difference=0), very small (8 vs. 7%, difference=1%), small (8 vs. 6%, difference=2%), medium (8 vs. 5%, difference=3%) and large (8 vs. 4%, difference=4%) based on results from our previous clinical trial (Carlson et al., 2013). In the null scenarios where the treatment effect was 0, we identified the δ values which made the average success rate across simulations approximately equal to 0.05, $P(Pr(p_c > p_t|\text{data}) > \delta) \approx 0.05$. δ values can vary in different

statistical methods. This ensured the type I error rate was about 5%. In other scenarios, $P(Pr(p_c > p_t|data) > \delta)$ was used to calculate the power of the tests.

We compared the simulated trial operating characteristics, (bias, power and MSE) across the three models for both ePTB and PTB. These were based upon 1000 simulations and 600 subjects in each group because our designed trial has a sample size around 1200. The δ value was 0.95 for both ePTB and PTB simulations, in the beta-binomial model and the finite mixture model. In the logarithmic transformation model, δ was 0.999 for ePTB simulations and 0.997 for PTB simulations. All methods were implemented in R 3.1.1 and Openbugs.

3.3.2 Simulation results

In the simulation study of probability of ePTB (<34 weeks), the beta-binomial model had lower bias compared to the finite mixture model and the logarithmic transformation model in all scenarios. The MSE in the finite mixture model was consistently lower than in the beta-binomial model and logarithmic transformation model in the control group and slightly higher than in the beta-binomial model in the last three scenarios in the treatment group (Table 3.1). Figure 3.1 shows the comparisons of bias, variance, MSE and power across the three models. In the null scenario, the type I error rate was 0.048 in the beta-binomial model, 0.054 in the finite mixture normal model, and 0.053 in the logarithmic transformation model (Table 3.1). The power for the finite mixture model was slightly higher than the beta-binomial model in other scenarios, but the difference was small (Figure 3.1). The logarithmic transformation model had the largest bias and lowest power (Table 3.1).

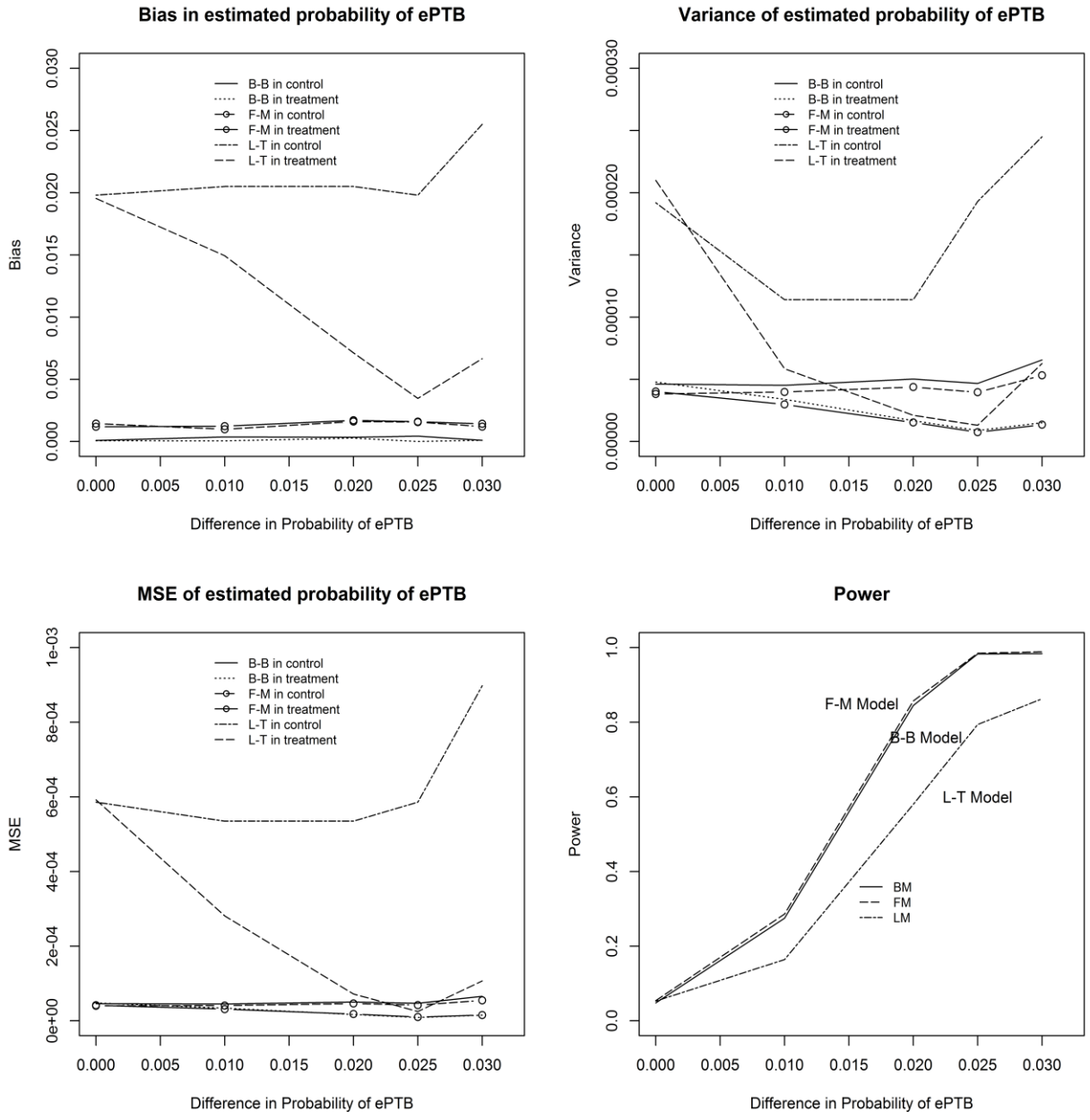

Figure 3.1: Simulated Trial Operating Characteristics for Probability of ePTB (GA<34 weeks)

In the simulation study of probability of PTB (<37 weeks), the beta-binomial model continued to have lower bias compared to the finite mixture normal model and the logarithmic transformation model. The difference in MSE between the finite normal mixture model and the

beta-binomial model was larger than that in the ePTB simulations (Table 3.1 and Table 3.2). The logarithmic transformation model again had the largest bias and largest MSE (Table 3.2). In the null treatment effect scenario, the type I error rate was 0.054 in the beta-binomial model, 0.05 in the finite mixture normal model, and 0.051 in the logarithmic transformation model (Table 3.2). The power for the finite mixture model was higher than the beta-binomial model in other scenarios, with differences as large as 7.5% when the true effect was likely (8 vs. 5%) (Table 3.2). Figure 3.2 shows the comparisons of bias, variance, MSE and power across the three models.

These simulation results demonstrated that although the bias from the finite mixture method was slightly larger than that from the pre-dichotomizing method, the parameter estimates from the finite mixture method had desirable properties such as lower MSE and lower variance. In ePTB simulation, the finite mixture model did not appear to be more desirable than the beta-binomial model. However, the advantages of the finite mixture model became apparent in PTB analysis. The logarithmic transformation method has the largest bias and highest MSE. In a word, the logarithmic transformation model appeared to be inferior to the finite mixture model.

3.4 Application to real data

To illustrate the use of the three models in real data, we reanalyzed the gestational age data from an Australia based clinical trial and a USA based clinical trial.

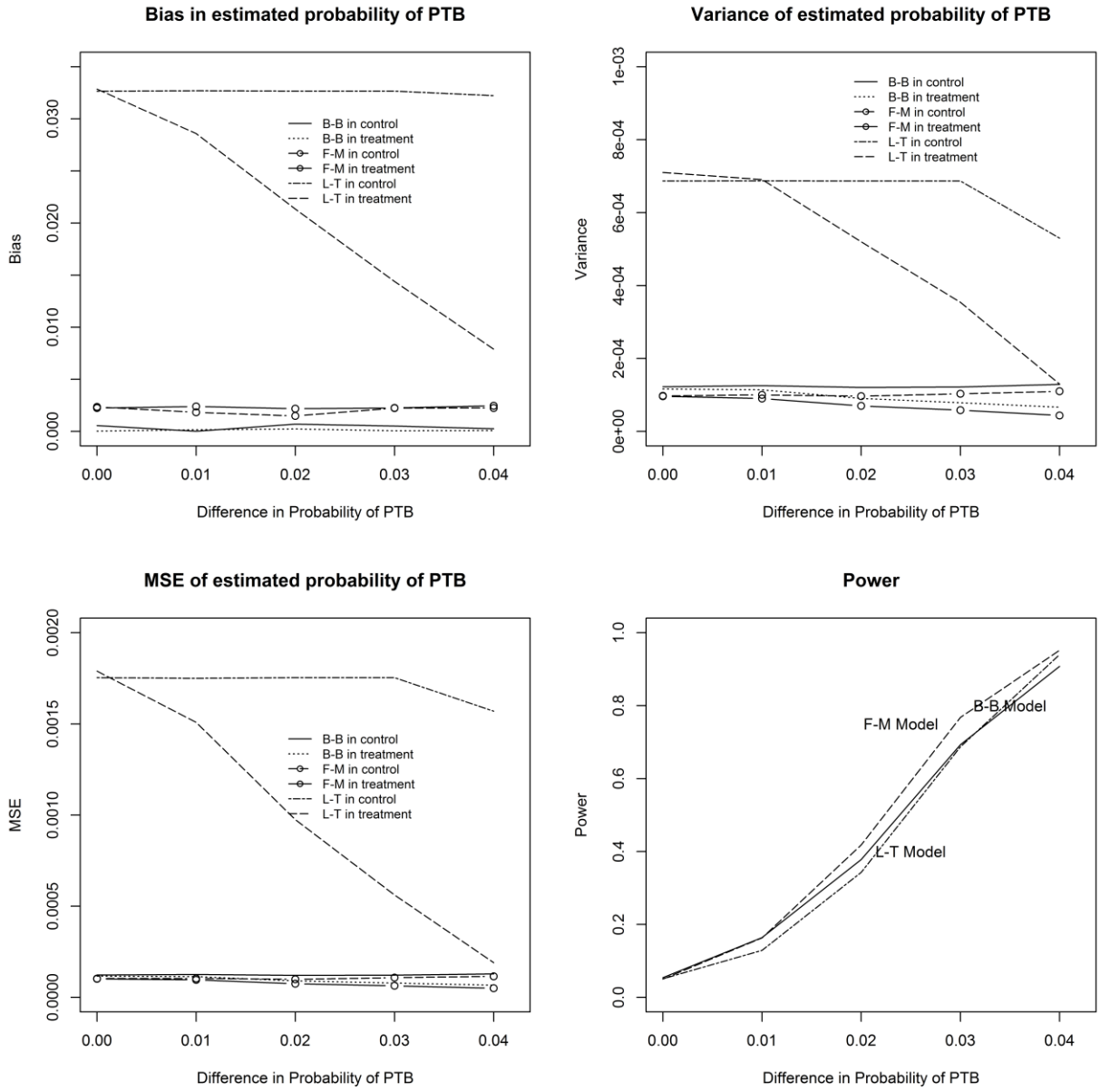

Figure 3.2: Simulated Trial Operating Characteristics for Probability of PTB (GA<37 weeks)

3.4.1 DOMInO trial

The DOMInO trial was a double-blind, multicenter, randomized controlled trial conducted in five Australian maternity hospitals. The trial included 2399 women who were less than 21 weeks' gestation with singleton pregnancies and who were recruited between October 31,

2005, and January 11, 2008 (Makrides et al., 2010). This study compared fish oil capsules (providing 800 mg/d of DHA) or matched vegetable oil capsules without DHA. Gestational age data were available for 2367 (1183 in control and 1184 in treatment) participants in this study.

We looked at the posterior summary statistics of the posterior component probabilities in the control and treatment groups from the finite mixture model (Table 3.3). Compared to the control group, the posterior probability of the first component (low risk of PTB) increased from 0.783 to 0.813 and the posterior probability of the third component (high risk of PTB) decreased from 0.04 to 0.022. The posterior probability of the second component decreased a little from 0.177 to 0.165. Convergence diagnostics were checked to ensure the convergence of posterior samples.

In Table 3.4, we show the calculated and estimated probability of ePTB and PTB and the standard deviation of the estimated probabilities. In this analysis, we found the benefits of the finite mixture model were not clear in ePTB but the standard deviation was slightly smaller in the finite mixture model in PTB analysis. The estimated proportions for the log transformation model are quite different to the raw data (Table 3.4).

3.4.2 KUDOS trial

KUDOS was a Phase III, randomized, double-blind, placebo-controlled clinical trial involving 299 women (Carlson et al., 2013). This study compared participants in the placebo group ($n_1 = 145$) and participants who received 600 mg/day DHA ($n_2 = 154$) in the second and third trimester during pregnancy from 2001 to 2006 in the University of Kansas Hospital (Carlson et al., 2013).

The posterior summary statistics of the mixture weights were summarized in Table 5. Compared to the DOMInO trial, the difference in the three component probabilities between treatment and control groups was much larger (Table 3.3 and Table 3.5). The mixture weight of the third component (high risk of PTB) decreased dramatically from 0.089 in the control group to 0.029 in the intervention group. Convergence diagnostics were checked to ensure the convergence of posterior samples.

In Table 3.6, we show the calculated and estimated probability of ePTB and PTB and the standard deviation of the estimated probabilities. Again in this analysis, we found the advantages of the finite mixture model compared to beta-binomial model were not very clear for ePTB but the standard deviation was smaller in the finite mixture model for PTB. Both the DOMInO and KUDOS data were consistent with the simulation studies and show that the benefits of the finite mixture model are evident for PTB but questionable for ePTB. The logarithmic transformation model produced quite different results compared to the other two models, which may be due to the bias in this method observed in the simulation study.

3.5 Discussion

We aimed to investigate the properties of pre-dichotomizing and distributional approaches using a three-component normal mixture model and a logarithmic transformation model. The three-component normal mixture model has been demonstrated to be identifiable and superior to two-component mixture models while avoiding the poor mixing in models with four or more components (Schwartz et al., 2010). The Bayesian framework provides us with a convenient tool to compare distributional approaches and the pre-dichotomizing method.

In the simulation study, we used a weak beta prior for the beta-binomial model to ensure the bias was negligible and the estimates were close to the Frequentist approach. As a result, the bias from the finite mixture model was greater than that from the beta-binomial model. However, the finite mixture model had lower variance in all scenarios. In the ePTB analysis where the endpoint was very rare, the power of the finite mixture model was only slightly higher than the beta-binomial model and the benefits of the finite mixture model were relatively small. The benefits of the finite mixture model were more apparent in the PTB analysis where the endpoint was less rare. In this case, the variance and hence the MSE were much lower and the power was higher in the finite mixture model compared with the other methods. The logarithmic transformation model had the largest bias and MSE.

In real data analysis, both DOMInO and KUDOS trial data demonstrated the advantages of the finite mixture model in PTB analysis. The finite mixture model had lower standard deviation compared to the beta-binomial model for PTB in both datasets. The logarithmic transformation model produced quite different results in both analyses. These findings confirmed previous findings that the logarithmic transformation was not satisfactory for GA data (Sauzet et al., 2015).

Further investigating the three-component mixture model facilitates understanding of our observations in the simulation study and real data analysis. The three mixture components are: N(39.59, 0.96), N(38.26, 2.48), and N(33.29, 13.23). The mixture weights are about 70-80% for the first component, 10-20% for the second component, and less than 10% for the third component. The three components have different means and standard deviations (heteroscedastic). Therefore it is not straightforward to describe the exhibition of the mixture distribution. However, we can still speculate the mixture exhibition from the three mixture

components and the mixture weights. The first two components have close means and different standard deviations. Distribution mixing these two will display high kurtosis with a sharper peak and heavier tails than a single distribution (Gridgman, 1970). The third component is sufficiently separated from the first two components. The difference in the means between the second and third components is greater than two times the standard deviation of the second component. Mixing of these two could form a bimodal distribution (Schilling, Watkins, & Watkins, 2002). Since the mixture weights of the first two components are dominant and the standard deviation of the third component is large, the exhibition could have a long left tail with a small peak on the tail.

Based on the exploration of the finite mixture model, we can obtain an intuitive explanation of our observations. In the ePTB analysis we used GA<34 as a cutoff. Given the exhibition of the mixture model, the area below 34 was mainly captured by the third component of the distribution. In the PTB analysis we used GA<37 as a cutoff and the area below 37 was comprised of the second and the third components, while the influence of the first component was trivial. Therefore in the ePTB analysis, the finite mixture model did not appear to be much better than the beta-binomial model in terms of power because most of the information we needed to make inference on the probability of ePTB was captured by one mode in a bimodal exhibition. In the PTB analysis, the information to make inference on the probability of PTB was captured by two components and the finite mixture model captured the information from the trend of the two components and retained the power from the continuous distribution. Gestation age analysis is a single example in real life where we care about dichotomized outcomes while continuous data are collected. This study showed the cutoff value and the exhibition of the distribution were important to understand the mechanism of gaining power from a continuous

distribution. This conclusion can be generalized to other studies in which the outcome is dichotomized while data are collected in a continuous form.

3.6 Limitations

There are a few assumptions we have made to pursue this study. Firstly, we used the parameter estimates of the normal components from the North Carolina Detailed Birth Record (NCDBR) database and applied them to different populations. We assumed these component parameters were valid in different populations and they appeared to be fine in this study as the estimated probabilities are quite close to the true data. Although the finite mixture model has certain flexibility to allow component weights to vary, the parameter estimates or even the formation of the mixture model could change in other populations if the population is extremely different. Secondly, we assumed there was no measurement error in the gestational age data. Gestational age data were obtained from medical records but we do not have a technique to test the measurement error in the current study. If the measurement error was large, it could blur the boundary of ePTB and PTB.

3.7 Conclusion

In studies where endpoints are collected as continuous variables but clinicians are interested in studying dichotomized outcomes, a pre-dichotomizing or distributional approach could be used for analysis. In general, a distributional approach that fits the data well retains information and power from the continuous distribution, while a dichotomizing method is close to the traditional Frequentist approach and may result in less bias. The benefits of a distributional method depend on model fit, cutoff values, and the exhibition of the continuous distribution. Meticulous investigation of the distributions is necessary, especially in rare endpoint analysis

where retaining statistical power is more important. In our clinical trial designs for gestational age data, we recommend the finite mixture normal model if the endpoint is PTB (<37 weeks) since this is a more powerful design and think either finite mixture normal model or beta-binomial model is acceptable if the endpoint is ePTB (<34 weeks) since the power from these two designs are close.

Table 3.1: Simulated Trial Operating Characteristics for Probability of ePTB (GA<34 weeks)

Scenarios	Method	Bias		MSE×10^5		Power
		Control	Tx	Control	Tx	
No difference (3 vs. 3%)	B-B	.00009	.00006	4.62	4.78	.048
	F-M	.00144	.00119	4.02	4.17	.054
	L-T	.01983	.01954	58.56	59.21	.053
Very Small (3 vs. 2%)	B-B	.00037	.00005	4.53	3.39	.275
	F-M	.00099	.00123	4.08	3.13	.286
	L-T	.02051	.01493	53.46	28.14	.164
Small (3 vs. 1%)	B-B	.00034	.00028	5.03	1.69	.845
	F-M	.00160	.00171	4.63	1.82	.857
	L-T	.02051	.00713	53.46	7.19	.58
Medium (3 vs. 0.5%)	B-B	.00043	.00004	4.69	0.89	.983
	F-M	.00156	.00160	4.20	1.00	.985
	L-T	.01983	.00347	58.58	2.51	.794
Large (4 vs. 1%)	B-B	.00010	.00010	6.57	1.51	.984
	F-M	.00120	.00143	5.47	1.55	.989
	L-T	.02554	.00667	89.75	10.72	.863

Power: average success rate across simulations, $P(\Pr(p_c > p_t|data) > \delta)$, $\delta = 0.95$ for Beta-binomial and finite mixture model, $\delta = 0.999$ for logarithmic transformation model
B-B: Beta-Binomial model
F-M: Finite Mixture model
L-T: Logarithmic Transformation model

Table 3.2: Simulated Trial Operating Characteristics for Probability of PTB (GA<37 weeks)

Scenarios		Bias		MSE×10^5		Power
	Method	Control	Tx	Control	Tx	
No difference (8 vs. 8%)	B-B	.00056	.00003	12.3	11.6	.054
	F-M	.00235	.00224	10.3	10.1	.05
	L-T	.03265	.03286	175	178	.051
Very Small (8 vs. 7%)	B-B	.00002	.00016	12.6	11.4	.164
	F-M	.00184	.00239	10.4	9.59	.163
	L-T	.03266	.02859	175	151	.129
Small (8 vs. 6%)	B-B	.00070	.00024	12.1	9.05	.378
	F-M	.00149	.00218	9.90	7.47	.418
	L-T	.03266	.02134	175	98	.343
Medium (8 vs. 5%)	B-B	.00051	.00006	12.2	7.86	.693
	F-M	.00224	.00225	10.8	6.33	.768
	L-T	.03266	.01439	175	56	.687
Large (8 vs. 4%)	B-B	.00025	.00008	12.9	6.64	.908
	F-M	.00224	.00246	11.5	4.98	.952
	L-T	.03223	.00789	157	19	.94

Power: average success rate across simulations, $P(\Pr(p_c > p_t|data) > \delta)$, $\delta = 0.95$ for Beta-binomial and finite mixture model, $\delta = 0.997$ for logarithmic transformation model
B-B: Beta-Binomial model
F-M: Finite Mixture model
L-T: Logarithmic Transformation model

Table 3.3: Posterior statistics for mixture weights in finite mixture model in DOMInO trial (10000 simulations)

	Control		Treatment	
	mean	std	mean	std
Δ_1	.783	.022	.813	.021
Δ_2	.177	.023	.165	.021
Δ_3	.040	.007	.022	.006

Δ_1: posterior probability of component 1, N(39.59, 0.96)
Δ_2: posterior probability of component 2, N(38.26, 2.48)
Δ_3: posterior probability of component 3, N(33.29, 13.23)

Table 3.4: Domino Data analysis: calculated and estimated probability and standard deviation (10000 simulations)

	Outcome	Data		Estimated Prob. And SD			
		Pc0	Pt0	Pc	Pt	SDc	SDt
B-B	<34 wks	.023	.011	.023	.011	.004	.003
	<37 wks	.072	.055	.072	.055	.008	.007
F-M	<34 wks	.023	.011	.024	.013	.004	.003
	<37 wks	.072	.055	.075	.057	.007	.006
L-T	<34 wks	.023	.011	.009	.004	.001	.001
	<37 wks	.072	.055	.099	.068	.007	.007

B-B: Beta-Binomial model
F-M: Finite Mixture model
L-T: Logarithmic Transformation model
Outcome: probability of GA less than a certain amount of time
Pc0: the calculated probability in the data in control group
Pt0: the calculated probability in the data in treatment group
Pc: the estimated probability in the control group
Pt: the estimated probability in the treatment group
SDc: standard deviation in the control group
SDt: standard deviation in the treatment group

Table 3.5: Posterior statistics for mixture weights in finite mixture model in KUDOS trial (10000 simulations)

	Control		Treatment	
	mean	std	mean	std
Δ_1	.838	.048	.775	.060
Δ_2	.073	.048	.196	.063
Δ_3	.089	.027	.029	.018

Δ_1: posterior probability of component 1, N(39.59, 0.96)
Δ_2: posterior probability of component 2, N(38.26, 2.48)
Δ_3: posterior probability of component 3, N(33.29, 13.23)

Table 3.6: KUDOS Data analysis: calculated and estimated probability and standard deviation (10000 simulations)

	Outcome	Data		Beta-Binomial Model			
		Pc0	Pt0	Pc	Pt	SDc	SDt
B-B	<34 wks	.048	.007	.048	.007	.018	.007
	<37 wks	.09	.065	.09	.065	.024	.02
F-M	<34 wks	.048	.007	.052	.018	.016	.011
	<37 wks	.09	.065	.094	.069	.022	.017
L-T	<34 wks	.048	.007	.016	.004	.006	.002
	<37 wks	.09	.065	.136	.069	.023	.016

B-B: Beta-Binomial model
F-M: Finite Mixture model
L-T: Logarithmic Transformation model
Outcome: probability of GA less than a certain amount of time
Pc0: the calculated probability in the data in control group
Pt0: the calculated probability in the data in treatment group
Pc: the estimated probability in the control group
Pt: the estimated probability in the treatment group
SDc: standard deviation in the control group
SDt: standard deviation in the treatment group

Chapter 4 : Personalized Medicine Enrichment Design for DHA Supplementation Clinical Trial

Yang Lei, Matthew S. Mayo, Susan E. Carlson, & Byron J. Gajewski

(Submitted to Clinical Trials)

Abstract

Background Personalized medicine aims to match patient subpopulation to the most beneficial treatment. The purpose of this study is to design a prospective clinical trial in which we hope to achieve the highest level of confirmation in identifying and making treatment recommendations for subgroups. This study was motivated by our goal to identify subgroups in a DHA (docosahexaenoic acid) supplementation trial to reduce preterm birth (gestational age<37 weeks) rate.

Methods We used four subgroups with 250 subjects in each group as an example and ordered the subgroups' risk levels in the control arm (placebo or current standard of care). The experimental arm had four groups of same sample sizes with changed risks. We simulated operating characteristics to ensure that overall Type I error rate was close to 0.05 in designs with four different models: naïve logistic, independent, hierarchical, and dynamic linear models. We obtained an informative prior distribution through a meta-analysis that included nine clinical trial studies across the world. We then carried out simulations and sensitivity analysis to examine the subgroup power of the four models and compared the results to a chi-square test.

Results We examined a large overall effect scenario and a small overall effect scenario, and within each scenario, three situations in which the resulting rates in experimental arm were linear, flat, or nonlinear, to mimic situations that subgroups benefited differently. The logistic model was excluded because it was not flexible and induced large Type I error rate in certain subgroup analysis. In the large overall effect setting, dynamic linear model increased the power of the most affected group by 2.9% - 3.3% compared to hierarchical model, 4.3% - 8.6 % compared to independent model, and 7% - 10% compared to chi-square test. Dynamic linear model outperformed the other models in most other subgroup analysis. In the small overall effect

setting, dynamic linear model increased the power of the most affected group by 2% compared to hierarchical model, 1.6 % compared to independent model, and 14.2% compared to chi-square test when the resulting rates are linear. It increased the power of the most affected group by 14% compared to hierarchical model, 13.2 % compared to independent model, and 16.1% compared to chi-square test when the resulting rates are flat. It was outperformed by hierarchical model or chi-square test when the resulting rates are non-linear. Dynamic linear model remained robust and powerful in other subgroup analysis.

Conclusions Compared to independent and hierarchical models, dynamic linear model tends to be relatively robust and powerful when the control arm has ordinal risk groups.

Keywords: enrichment design, subgroup analysis, overall Type I error, power

4.1 Introduction

An important trend in treatment paradigm is personalized medicine, which is aimed to match patients to the most beneficial treatments. Patient populations are heterogeneous even in the same study. Characteristics vary between individuals, such as demographics, life style, genetic variants, etc. These varied characteristics can potentially modify the treatment effects on different individuals or subsets of patient populations. It is important to distinguish the subgroups that benefit the most and subgroups that don't benefit or might even unnecessarily be exposed to a hazardous side effect (Simon & Simon, 2013) . Our research was motivated by our goal to design a clinical trial to identify subgroups in a trial to supplement pregnant women with docosahexaenoic acid (DHA) to reduce the rate of preterm births (PTB, gestational age<37 weeks). This is the first step in an enrichment design where a certain subgroup will be identified and the succeeding steps will distinguish the treatment effect within the selected (enriched) subpopulation (Fedorov & Liu, 2007).

Berry et al. (2013) discussed three clinical trial designs assuming four groups of patients under an adaptive framework: Simon's Optimal Two-Stage design, a Bayesian independent design, and a Bayesian hierarchical design (Berry, Broglio, Groshen, & Berry, 2013). They showed that the hierarchical model could provide additional power and reduction in sample size compared to other two methods but acknowledged that hierarchical modeling could make finding a single effective group more difficult, if there was only one (Berry et al., 2013). We followed their four-group design in this study. The four-group design is an example for illustration and can be generalized to different settings. The hierarchical model does not require the entities to be related (Berry et al., 2013). In practice it is common to classify subjects into different risk levels of subpopulations. We classified the four groups according to their risk levels assuming they are

receiving placebo or standard of care (control arm). Our study extended Berry et al. (2013) and aimed to identify a more efficient design from a prospective perspective to achieve the highest level of confirmation in identifying and making recommendations for subgroups (Ruberg & Shen, 2015), given the fact that the risk groups can be ordered in the control arm.

Two major considerations on subgroup analysis in clinical trial designs are: preserving Type I error and improving power (Simon & Simon, 2013). Testing each hypothesis in a multi-group study inflates the overall Type I error rate. Multiplicity adjustment is required to preserve the overall Type I error rate (Alosh et al., 2015). We calibrated the operating characteristics in simulations to ensure the overall Type I error rate was close to 0.05 (one-sided) in all designs that used different statistical models. Approaches to improve statistical power in subgroup analysis include: using available information from previous studies (Alosh et al., 2015) and borrowing information across subgroups (Berry et al., 2013). We did a meta-analysis that contained data from nine DHA supplementation trials across the world to obtain informative priors. Then informative priors were applied to three different models: independent model, hierarchical model, and dynamic linear model. A naïve logistic model was studied but uninformative priors were used because no prior information was available for the parameters in this model. All methods were compared to the chi-square test to see the benefit of each model in the trial design.

The remainder of this article is arranged as follows. In section 2, we obtained informative prior distributions through meta-analysis based upon nine previous clinical trials across the world. In section 3, we described the four statistical models in trial designs. In section 4, we explained the computation methods and software we used. In section 5, we discussed the results from the simulations. In section 6, we performed a sensitivity analysis and discussed potential

concerns. In section 7, we draw conclusions from our analysis and give suggestions to future studies.

4.2 Prior Distributions

In subgroup analysis, a prior distribution is assumed for subgroup-specific treatment effect (Alosh et al., 2015). There are advantages and disadvantages of using both non-informative and informative priors. When historical data for the control arm are consistent with current study data, using informative priors constructed from previous complete trials can improve testing power and generate robust results (Chen et al., 2015). We performed a meta-analysis based on nine completed DHA supplementation trials that contain preterm birth data across the world. Five of them were included in a review study conducted by the Cochrane Collaboration: Denmark 1992, England 1995, Europe 2000, Netherlands 1994, and USA 2003 (Makrides, Duley, & Olsen, 2006). Besides these we included four other trials: KUDOS (Kansas University Hospital) 2013 (Carlson et al., 2013), DOMInO (Australia) 2010 (Makrides et al., 2010), Mexico 2015 (Ramakrishnan et al., 2015), and NICHD (USA) 2010 (Harper et al., 2010). The data from these studies are summarized in Table 4.1.

We used a hierarchical model with relatively weak priors to obtain future prior distributions. Let P_{ij} denote proportion of preterm birth in the i^{th} study (i=1,…9) and j^{th} arm (j=0,1; 0=control, 1= experimental). We modeled $\theta_{ij} = \log\left(\frac{P_{ij}}{1-P_{ij}}\right) \sim N(\mu, \tau)$, where τ is the precision (1/variance), with relatively weak hyper priors: $\mu \sim N(-2, 0.5)$ and $\tau \sim Gamma\ (1, 1)$. Future priors for μ and τ were derived from the averaged methods of moment estimators in the posterior distributions from the experimental arm and control arm. The posterior mean for μ is -1.872 in the control arm and -1.944 in the experimental arm. The posterior mean for τ is 1.273 in the

control arm and 1.287 in the experimental arm. The standard deviation for τ is 0.5874 in the control arm, and 0.6015 in the experimental arm. The methods of moment estimator for μ is calculated as $\frac{(-1.872-1.944)}{2} = -1.91$. The estimator for τ is calculated as $\frac{1.273+1.287}{2} = 1.28$. The α and β estimators in the distribution of $\tau \sim Gamma\ (\alpha, \beta)$ are derived from equations: $\frac{\alpha}{\beta} = 1.28$ and $\frac{\alpha}{\beta^2} = \frac{0.5874^2+0.6015^2}{2} = 0.3534$. In this way we obtained informative priors $\mu \sim N(-1.91, 1.28)$ and $\tau \sim Gamma(4.6361, 3.622)$. Based on these prior distributions, the median of the proportion is 12.8%. This estimation is reasonable and consistent, as the current preterm birth rate in the US is about 11.4% (House, 2014). We applied these informative priors to simulations in the following trial designs.

4.3 Statistical Models in Trial Designs

We used four risk groups for illustration but the conclusion applies more generally. We assumed equal sample size of 250 subjects in each subgroup in both control and experimental arm. Two scenario settings were considered. The first scenario represents large overall treatment effect. The overall PTB rates are 8% vs. 4% in control and experimental arm respectively. The second scenario represents small overall treatment effect. The overall PTB rates are 8% vs. 6% in control and experimental arm respectively. These percentages are consistent with the results from our previous DHA supplementation trial (Carlson et al., 2013). The control arm in both scenarios has the same structure with ordinal PTB rates in the four subgroups (4%, 6%, 10%, and, 12% respectively). Within each scenario, we designed three different situations where the resulting rates in the experimental arm are linear, flat, or nonlinear (Table 4.2). We compared four trial designs and a chi-square test with Bonferroni correction for Type I error rate.

4.3.1 Naïve logistic Model

This design was originated from a naïve assumption that the PTB proportions can be modeled by a logistic regression based on their baseline risk level.

$$\log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_0 + \beta_1 \times j, \qquad i = 0,1; j = 1,2,3,4$$

This model assumed the difference between one risk level and the next level is a constant at the logit scale. Since no previous research studied the relationship between the preterm birth rates and risk levels, we used non-informative priors for β_0 and β_1.

$$\beta_0 \sim N(0, 100^2)$$

$$\beta_1 \sim N(0, 100^2)$$

4.3.2 Independent Model

We followed Berry et al. (2013) and examined an independent model (Berry et al., 2013). In this design, we presented a Bayesian model with no borrowing from subgroups but we did borrow information from previous studies by applying the informative priors obtained through meta-analysis. We modeled the rate in each subgroup within each arm separately through:

$$\theta_{ij} = \log\left(\frac{P_{ij}}{1-P_{ij}}\right), i = 0,1; j = 1,2,3,4$$

The prior distribution for θ_{ij} is:

$$\theta_{ij} \sim N(-1.91, 1.28)$$

This informative prior results in mean proportion close to 12.8% but it can generate a proportion ranges from about 0.8% to 60%.

4.3.3 Hierarchical Model

This is another design that was examined in Berry et al. (2013) (Berry et al., 2013). This design integrates the heterogeneous information from each subgroup. The hierarchical model assumes the four groups are exchangeable and allowed borrowing information across the four groups. In this design we model θ_{ij} with a normal distribution with unknown mean and precision

$$\theta_{ij} \sim N(\mu_i, \tau_i)\ i = 0,1; j = 1,2,3,4$$

By introducing a hierarchy to model the unknown mean and precision, the design borrows information from previous studies and the current data across the four groups.

$$\mu_i \sim N(-1.91, 1.28)$$

$$\tau_i \sim Gamma\ (4.6361, 3.622)$$

Bigger τ indicates more pooling and information borrowing across the groups and smaller τ represents less pooling, or more heterogeneity across the groups. Our priors from meta-analysis show significant heterogeneity from the data in previously conducted clinical trials and we think it is reasonable to remain the heterogeneity in our simulation analysis to apply to a general population.

4.3.4 Dynamic Linear Model

Dynamic linear model is another model that has an intrinsic hierarchical structure. Unlike the hierarchical model that we discussed in the previous section, the dynamic linear model does not assume exchangeability of the four groups and borrows more information from adjacent groups. The motivation is that this model might be more efficient since the four groups have ordinal risk levels at baseline so the correlation between adjacent groups might be larger. Therefore the dynamic linear model might capture the locally smooth trend better. In this design, the first group has a prior we obtained through meta-analysis, and the other groups have a

hierarchical structure, with a common precision and a mean related to the neighborhood means (Leininger, Reese, FEllingham, & Grimshaw, 2010).

For the first group, we have:

$$\theta_{i1} \sim N(-1.91, 1.28); i = 0,1$$

For the other groups,

$$\theta_{ij} \sim N\left(\theta_{i,j-1}, \tau_i\right);\ i = 0,1;\ j = 2,3,4$$

and

$$\tau_i \sim Gamma\ (4.6361, 3.622).$$

The second group directly borrows information from the first and the third. The third group directly borrows information from the second and the fourth. The first group directly borrows information from previous studies and the second group. However, since the borrowing process is dynamic, the groups not adjacent directly impact each other through an indirect borrowing mechanism. This structure allows groups to borrow more information from adjacent groups which locally smoothes the trend.

4.4 Computation

In each scenario, 1000 simulated trials were used. We assumed the experimental and control arms each had four subgroups and each subgroup had 250 patients. The PTB rates in the four groups under the control arm were ordered and represented the risk levels assuming subjects were receiving placebo or standard of care. The PTB rates in the four groups under the experimental arm represented the risk levels after treatment, e.g., DHA supplementation. Data were simulated from binomial distribution based upon the proportions in each scenario using R 3.2.2. All Bayesian computations were performed using OpenBUGS from within R 3.2.2.

OpenBUGS is powerful and flexible because it includes a system that determines an appropriate MCMC for analyzing a specified model without requiring a closed analytic form.

4.5 Results

Trial success was defined as the posterior probability that the PTB rate in the control arm is bigger than that in the experimental arm is greater than a cutoff value δ: $\Pr(P_c > P_t|data) > \delta$. In simulations the power function is defined as the average trial success rate across simulations. In the null scenario both the control arm and the experimental arm have overall PTB rates of 8% and the average success rate is the Type I error rate. Since we have four subgroups, the overall Type I error rate is $1 - \prod_{i=1}^{4}(1 - \alpha_i)$, where α_i is the Type I error rate for the i^{th} group. In the null scenario, both the control and experimental arms have the same structure in their subgroups. The four risk groups have PTB rates of 4%, 6%, 10%, and 12% in both arms. We tuned the δ value in each method separately to ensure the overall Type I error rate is close to 0.05. The δ values for the logistic model, independent model, hierarchical model, and dynamic model are 0.993, 0.985, 0.985, and 0.98 respectively.

In the first scenario we examined the situation where the overall PTB rate is 8% in the control arm and 4% in the experimental arm, indicating a large overall treatment effect. Within this setting, we tried different structure of PTB rates in the experimental arm to mimic different subgroup effects (Table 4.2). First we simulated a situation where the resulting rate in the experimental arm is quite linear in four subgroups: 2%, 3%, 5%, and 6% correspondingly. In this situation, each subgroup experienced a 50% reduction in PTB rate. Second we simulated a situation where the resulting rate in the experimental arm is flat in all four subgroups: all 4%. In this situation, there is no effect in the first subgroup and the last group has the biggest treatment effect. Since the first group has no effect at all, the average success rate we obtained for this

subgroup is a Type I error rate for this subgroup analysis. Thirdly we simulated a situation where the resulting rates in subgroups in the experimental arm are non-linear: 1%, 6%, 3%, and 6% correspondingly. We created a situation where the second subgroup has no treatment effect while the other three groups have treatment effects. Therefore the average success rate we obtained from the second subgroup is the Type I error rate for this group. We noticed that the naïve logistic model introduced exploded Type I error rate in this situation while the other three models had reasonable Type I error rates. The Type I error rates in the second subgroup under this situation for logistic, independent, hierarchical, and dynamic linear models are 0.732, 0.006, 0.011, and 0.055 respectively. This shows the logistic model is not flexible in modeling situations where subgroup effects are not linear. We excluded logistic model in our following comparisons due to this inflexibility and the unintended explosion in Type I error rate in subgroup analysis.

In the comparison of the other three models, we focused on power in subgroup analysis. Dynamic linear model has the highest power in subgroup analysis in all situations except for the first subgroup in the non-linear situation (Table 4.3). In the non-linear scenario, the proportions in the second group are 0.06 for both control and treatment arm. In this case, dynamic linear model has a slightly higher Type I error rate than other models but the error rate is still acceptable (Table 4.3). The dynamic linear model has the highest power in capturing the most affected group in all three situations. This is important because it represents the capability to identify or confirm the most beneficial subgroup. In the situation where the rates in the four groups of the experimental arm are linear, the power to capture the most affected subgroup in independent, hierarchical, and dynamic linear models is 0.548, 0.578, and 0.595 respectively. A regular chi-square test with Bonferroni adjustment has a power of 0.541 to capture the most

affected subgroup. In the situation where the rates in the four groups in the experimental arm are flat, the power to capture the most affected subgroup in independent, hierarchical, and dynamic linear models is 0.879, 0.891, and 0.917 respectively. The chi-square test has a power of 0.857 to capture the most affected subgroup. In the situation where the rates in the four groups in the experimental arm are non-linear, the power to capture the most affected subgroup in independent, hierarchical, and dynamic linear models is 0.848, 0.861, and 0.889 respectively. The chi-square test has a power of 0.827 in capturing the most affected subgroup. In a nutshell, the dynamic linear model increases power to capture the most affected subgroup compared to the other three methods where the overall treatment effect is large (Table 4.3). In addition to providing the power to capture the most affected group, the dynamic linear model appears to be powerful and robust in other subgroup analysis (Figure 4.1- Figure 4.3).

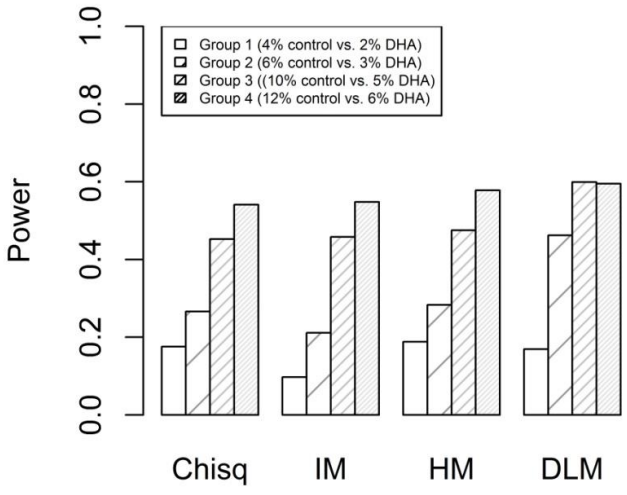

Figure 4.1: Power analysis for linear subgroup effects in treatment arm in scenario 1

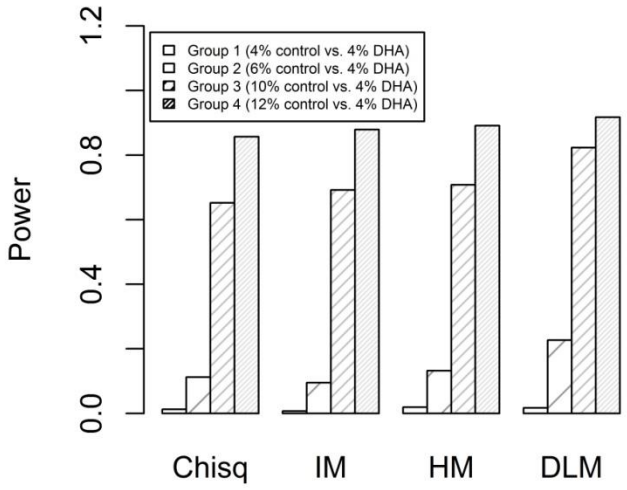

Figure 4.2: Power analysis for flat subgroup effects in treatment arm in scenario 1

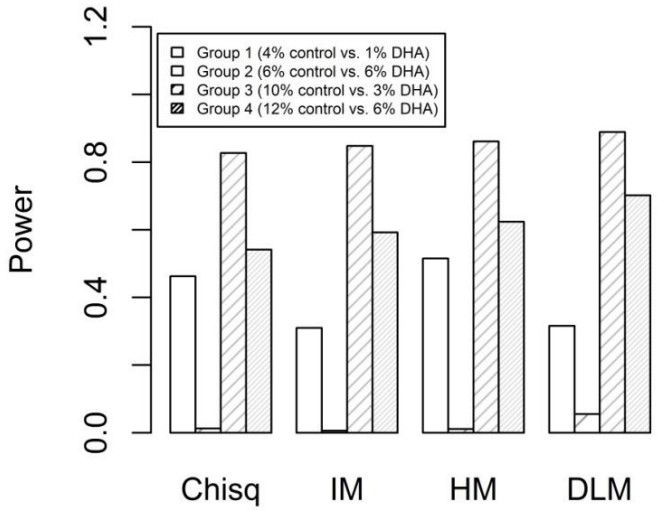

Figure 4.3: Power analysis for nonlinear subgroup effects in treatment arm in scenario 1

In the second scenario setting we examined the situations where the overall PTB rate is 8% in the control arm and 6% in the experimental, indicating a small overall treatment effect. Again, we tried different structure of PTB rates in the experimental arm to mimic different subgroup effects (Table 4.2). In the first situation the rates in the four groups in the experimental arm are linear or ordinal: 4%, 5%, 7%, and 8% respectively. In the second situation the rates in the four groups in the experimental arm are flat: all 6%. In this situation, there is no effect in the second subgroup and a negative effect in the first subgroup. In the last situation the rates in the four groups in the experimental arm are non-linear: 1%, 6%, 6%, and 11% respectively. We created a situation where the second subgroup had no treatment effect while the other three groups had treatment effects, but the effects varied among the three groups. Therefore the average success rate we obtained from the second subgroup is the Type I error rate for this group. The Type I error rates in the second subgroup under this situation for logistic, independent, hierarchical, and dynamic linear models are 0.487, 0.009, 0.013, and 0.031 respectively. The naïve logistic model exploded the Type I error rate in this situation while the other three models had reasonable Type I error rate. Again this shows the inflexibility of the logistic model when the group rates are not linear.

The results of subgroup power analysis were summarized in Table 4.4. In the situation where the experimental arm subgroup rates are linear, the power to capture the most affected subgroup in independent, hierarchical, and dynamic linear models is 0.254, 0.253, and 0.258 respectively. A regular chi-square test with Bonferroni adjustment has a power of 0.226 to capture the most affected subgroup. In the situation where the treatment arm group rates are flat, the power to capture the most affected subgroup in independent, hierarchical, and dynamic linear models is 0.555, 0.551, and 0.628 respectively. The chi-square test has a power of 0.541 to

capture the most affected subgroup. In the second group analysis in the flat situation, the experimental effect is negative. Independent, hierarchical, and dynamic linear models perform well but the chi-square test cannot capture this trend unless we discern the results from comparing the control and experimental arms (Table 4.4). In the situation where the experimental arm group rates are non-linear, the power to capture the most affected subgroup in independent, hierarchical, and dynamic linear models is 0.32, 0.461, and 0.304 respectively. The chi-square test has a power of 0.462 to capture the most effected subgroup. Therefore the chi-square test and the hierarchical model perform well to capture the most affected group when the resulting subgroup rates are nonlinear (Table 4.4). When the resulting subgroup rates are quite linear or flat, the dynamic linear model still outperforms the other three methods. In other subgroup analysis, the dynamic linear model appears to be robust and powerful (Figure 4.4 – Figure 4.6).

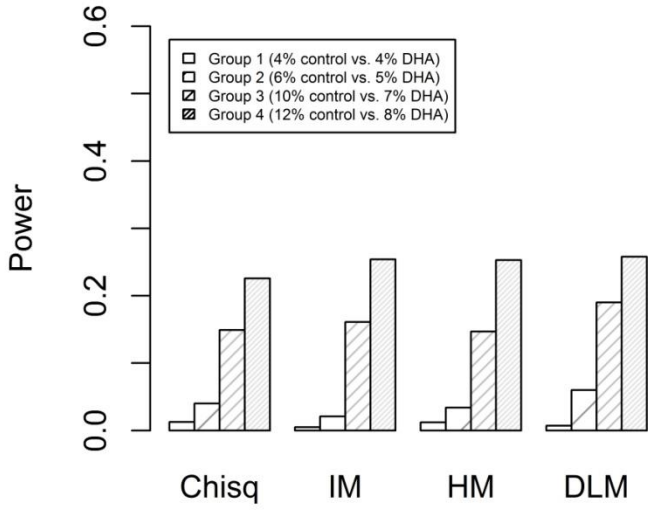

Figure 4.4: Power analysis for linear subgroup effects in treatment arm in scenario 2

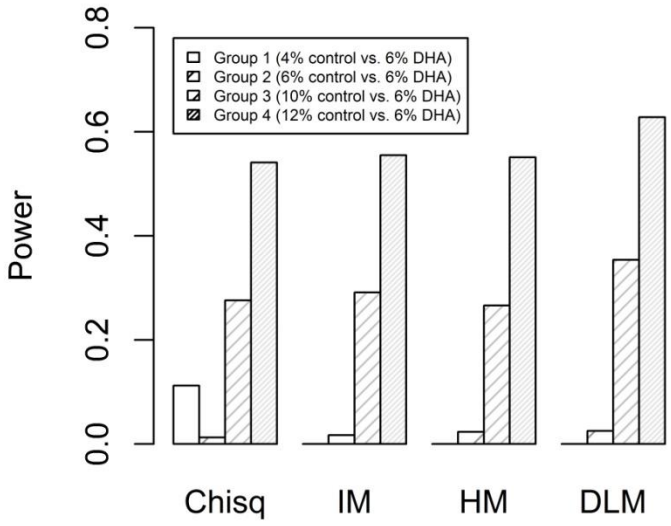

Figure 4.5: Power analysis for flat subgroup effects in treatment arm in scenario 2

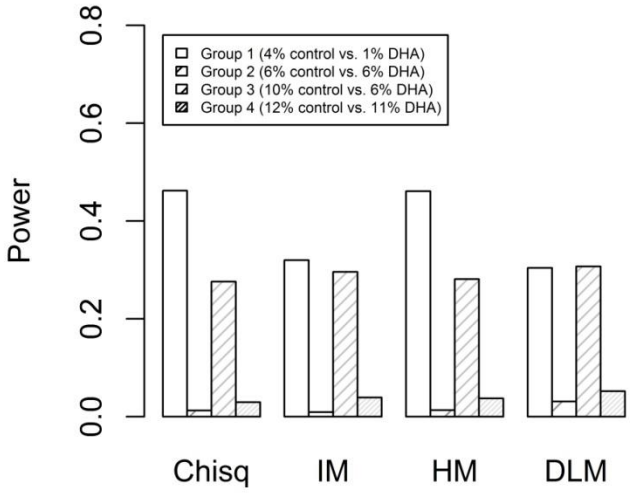

Figure 4.6: Power analysis for nonlinear subgroup effects in treatment arm in scenario 2

4.6 Sensitivity Analysis and Discussion

We did a sensitivity analysis using 500 subjects per group. The comparisons between the statistical models remained similar with increased capability to confirm futility or success of subgroups (Table 4.3 and Table 4.4). We assumed ordinal risk subgroups in the control arm have equal sample sizes. The general population may have unequal-sized strata for similar ordinal risk levels. However, in clinical trial designs, it is still possible to selectively include participants to create balanced numbers in each subgroup. The main advantage is that when sample size is predetermined and equal in subgroups, the most affected subgroups have the highest treatment effects. Predetermined subgroup sample sizes decrease the risk of insufficient statistical power at the end of study (Padmanabhan, 2014). At the current stage of subgroup analysis, it is not necessary to meet the power requirement in general statistical analysis. Once the most affected subgroups are identified through efficient designs, we can "enrich" the interested subgroups, i.e., recruit more subjects from the interested subgroup populations and stop recruiting certain subgroups that are futile. The final statistical analysis will be based upon all stages' recruitment.

We used unanimous informative priors in our statistical models. It is desirable to use subgroup specific priors if previous data are available. If the subgroup data are not consistent with the informative priors, the result could be decreased power.

4.7 Conclusions

In clinical trial designs with subgroup analysis, it is important to preserve a low Type I error rate and improve power to capture the most affected group. Informative priors are one way to increase power. When informative priors from historical data are consistent with current study data, they represent a more powerful mechanism. Another way to increase power is through design selection. Designs that used independent and hierarchical models have been discussed in

previous studies (Berry et al., 2013). Other efficient models may exist if we have ordinal risk groups. We compared the dynamic linear model, independent model, hierarchical model and logistic model. All these models can be tuned to have desirable operating characteristics in terms of overall Type I error rate. When we performed subgroup analysis, the logistic model turned out to be inflexible and exploded the Type I error in certain subgroups analysis. The main comparisons were executed among the other three methods. The dynamic linear model outperformed the other models in most situations with various structures of subgroup effects. We conclude that the dynamic linear model is relatively robust and efficient. This study shows that when the subgroups have certain structure, more efficient designs may exist and can lead to cost savings in clinical trials.

Table 4.1: Number of preterm babies and sample sizes in completed trials

	Treatment		Control	
Study	preterm birth	Total	preterm birth	Total
Denmark 1992	9	266	15	267
England 1995	22	113	19	119
Europe 2000	152	394	167	403
Netherlands 1994	8	32	10	31
USA 2003	14	142	17	149
KUDOS 2013	12	154	13	147
DOMINO 2010	88	1202	67	1197
Mexico 2015	32	365	30	365
NICHD 2010	82	434	83	418

Table 4.2: Preterm birth rates in subgroups in simulated scenarios

Scenario/risk group	Group 1	Group 2	Group 3	Group 4
Control arm				
8% (control arm)	4%	6%	10%	12%
Scenario 1: Treatment arm				
4% (linear)	2%	3%	5%	6%
4% (flat)	4%	4%	4%	4%
4% (nonlinear)	1%	6%	3%	6%
Scenario 2: Treatment arm				
6% (linear)	4%	5%	7%	8%
6% (flat)	6%	6%	6%	6%
6% (nonlinear)	1%	6%	6%	11%

Table 4.3: Power in subgroup analysis when the overall treatment effect is large (8% vs. 4%)

			Control	Treatment	Power			
	Scenarios	Group	True P	True P	DLM	IM	HM	Chi-sq
250 subjects/group	Linear	Group1	4%	2%	0.169	0.097	0.188	0.176
		Group2	6%	3%	0.462	0.211	0.283	0.266
		Group3	10%	5%	0.599	0.458	0.475	0.452
		Group4	12%	6%	0.595	0.548	0.578	0.541
	Flat	Group1	4%	4%	0.017	0.007	0.019	0.013
		Group2	6%	4%	0.227	0.095	0.132	0.112
		Group3	10%	4%	0.823	0.692	0.708	0.652
		Group4	12%	4%	0.917	0.879	0.891	0.857
	Non-linear	Group1	4%	1%	0.316	0.31	0.515	0.463
		Group2	6%	6%	0.055	0.006	0.011	0.013
		Group3	10%	3%	0.889	0.848	0.861	0.827
		Group4	12%	6%	0.702	0.592	0.624	0.541
500 subjects/group	Linear	Group1	4%	2%	0.401	0.3	0.426	0.349
		Group2	6%	3%	0.681	0.496	0.596	0.519
		Group3	10%	5%	0.878	0.811	0.823	0.777
		Group4	12%	6%	0.889	0.866	0.887	0.860
	Flat	Group1	4%	4%	0.023	0.012	0.029	0.013
		Group2	6%	4%	0.323	0.188	0.272	0.214
		Group3	10%	4%	0.98	0.958	0.962	0.932
		Group4	12%	4%	0.997	0.997	0.997	0.993
	Non-linear	Group1	4%	1%	0.769	0.753	0.867	0.788
		Group2	6%	6%	0.039	0.015	0.026	0.013
		Group3	10%	3%	0.998	0.995	0.995	0.988
		Group4	12%	6%	0.916	0.883	0.903	0.860

Table 4.4: Power in subgroup analysis when the overall treatment effect is small (8% vs. 6%)

			Control	Treatment	Power (250 subjects/group)			
			True P	True P	DLM	IM	HM	Chi-sq
250 subjects/group	Linear	Group1	4%	4%	0.007	0.005	0.012	0.0125
		Group2	6%	5%	0.06	0.021	0.034	0.0399
		Group3	10%	7%	0.19	0.161	0.147	0.1491
		Group4	12%	8%	0.258	0.254	0.253	0.2259
	Flat	Group1	4%	6%	0	0	0	0.112*
		Group2	6%	6%	0.025	0.017	0.023	0.0125
		Group3	10%	6%	0.354	0.291	0.266	0.276
		Group4	12%	6%	0.628	0.555	0.551	0.541
	Non-linear	Group1	4%	1%	0.304	0.32	0.461	0.462
		Group2	6%	6%	0.031	0.009	0.013	0.0125
		Group3	10%	6%	0.307	0.296	0.281	0.2761
		Group4	12%	11%	0.052	0.039	0.037	0.0293
500 subjects/group	Linear	Group1	4%	4%	0.019	0.014	0.024	0.013
		Group2	6%	5%	0.085	0.055	0.08	0.061
		Group3	10%	7%	0.385	0.342	0.337	0.294
		Group4	12%	8%	0.528	0.481	0.512	0.447
	Flat	Group1	4%	6%	0	0	0	0.214*
		Group2	6%	6%	0.013	0.012	0.017	0.013
		Group3	10%	6%	0.635	0.565	0.559	0.536
		Group4	12%	6%	0.902	0.876	0.886	0.86
	Non-linear	Group1	4%	1%	0.764	0.748	0.84	0.788
		Group2	6%	6%	0.031	0.009	0.019	0.013
		Group3	10%	6%	0.593	0.584	0.585	0.536
		Group4	12%	11%	0.06	0.052	0.056	0.04

* The power calculated from the one-sided Chi-square test cannot distinguish the direction of treatment effect

Chapter 5 : Summary and Future Directions

Personalized medicine is emerging in both clinical practice and clinical trials. Recognition that individual variability needs to be taken into account has driven the huge interest in 'precision' medicine. The motive is to identify the individuals who benefit from the treatment and who do not benefit from the treatment or even suffer from hazardous effects. On the other hand, personalized medicine has a potential to lower the skyrocketing health care cost by reducing unnecessary treatment and harmful effects. This initiative requires close collaboration among clinicians, researchers and biostatisticians. In 2015, President Obama launched the Precision Medicine Initiative, including establishing a national database of the genetic and other data of one million people in the United Sates. We are in the era of "big data". The contributions that biostatisticians can make in the health care arena are far beyond what we have witnessed in the past decades.

Personalized medicine contains efforts from the following aspects: identify the target population and direct the most beneficial treatment to the individuals. This dissertation provides statistical suggestions in pursuing these goals. In contrast to traditional logistic regression, this dissertation discussed the use of machine learning strategies and application in identifying target population in a binary responding variable setting. Given the fast development of computing power, data driven and machine learning methods are more feasible than before. With the establishment of gigantic databases, we can expect the rising use of data mining methods. CART model is one of the data mining methods. It is user-friendly and can provide straightforward interpretations in health care research. Clinical trials are a key research tool for advancing medical knowledge as well as patient care, yet they come with a staggering price tag. Efficient clinical trial designs help to lower the costs. This dissertation discussed statistical methods to improve the trial design efficiency. In many situations, clinicians are interested in a rare binary

event while continuous data are collected. To avoid losing power in pre-dichotomizing, this dissertation proposes distributional method. This dissertation further explores different distributional methods to identify a less biased and more powerful method to accommodate the rare event circumstance. For gestational age data, we proposed a three-component mixture normal model. Although this model increases the bias a little bit compared to binomial model, it promises lower variance, lower MSE and higher power in simulation studies. This model also outperformed the distributional model that uses logarithmic transformation. In personalized medicine, subgroup variation is the focus of the research. An efficient clinical trial design not only recognizes the subgroup differences, but also utilizes statistical models that borrow information across groups. Through borrowing information from previous studies and across groups, this dissertation discussed methods to increase the power to capture the most affected group as well as the power of other subgroups. We propose a dynamic linear normal model in this dissertation which promises a reliable and efficient design when we have ordinal risk groups.

This dissertation has motivated some topics for future research. In this dissertation, the methods used to classify target population and the methods to improve statistical power in trials with classified subgroups are discussed in two different application studies. It is very nature to use the methods we discussed in Chapter 2, i.e., logistic regression models and CART to classify subgroups in clinical trial designs. In DHA supplementation trials, we could use existing data, including clinical trials data or registry data to classify patient subpopulations. The classification based on real data may inspire different subgroups, e.g., different risk levels etc. It will be interesting to see whether these will impact on our model selection. Secondly, what we have discussed in Chapter 4 is the first stage in an adaptive enrichment trial design. After the interim analysis, we can adaptively enrich certain groups with potential to succeed and drop groups with

no potential to succeed in next stage recruitment. If the design involves multiple stages, we will reconsider how to preserve Type I error rate in the clinical trial designs. We will also consider sample size calculation to ensure power is guaranteed at the final stage analysis. Thirdly, we could consider distributional approaches in enrichment trials designs to see if it further boosts the power of subgroup analysis.

References

Alosh, M., Fritsch, K., Huque, M., Mahjoob, K., Pennello, G., Rothmann, M., . . . Yue, L. (2015). Statistical consideration on subgroup analysis in clinical trials. *Statistics in Biopharmaceutical Research, 7*(4), 286-303.

Altman, D. G., & Royston, P. (2006). The cost of dichotomising variables. *BMJ, 332*(7549), 1080.

American Nonsmokers' Rights Foundation. (2013). Overview list -- how many smokefree laws. Retrieved August 19, 2013, from http://www.no-smoke.org/pdf/mediaordlist.pdf

Arrowsmith, J. (2011). Trial watch: Phase III and submission failures: 2007-2010. *Nature Reviews Drug Discovery, 10*(2), 87.

Backinger, C. L., Fagan, P., O'Connell, M. E., Grana, R., Lawrence, D., Bishop, J. A., & Gibson, J. T. (2008). Use of other tobacco products among U.S. adult cigarette smokers: prevalence, trends and correlates. *Addict Behav, 33*(3), 472-489.

Berry, S. M., Broglio, K. R., Groshen, S., & Berry, D. A. (2013). Bayesian hierarchical modeling of patient subpopulations: efficient designs of Phase II oncology clinical trials. *Clinical Trials, 10*, 720-734.

Bombard, J. M., Pederson, L. L., Koval, J. J., & O'Hegarty, M. (2009). How are lifetime polytobacco users different than current cigarette-only users? Results from a Canadian young adult population. *Addict Behav, 34*(12), 1069-1072.

Bombard, J. M., Pederson, L. L., Nelson, D. E., & Malarcher, A. M. (2007). Are smokers only using cigarettes? Exploring current polytobacco use among an adult population. *Addict Behav, 32*(10), 2411-2419.

Campbell, M. L., Bozec, L. J., McGrath, D., & Barrett, S. P. (2012). Alcohol and tobacco co-use in nondaily smokers: an inevitable phenomenon? *Drug Alcohol Rev, 31*(4), 447-450.

Carlson, S., Colombo, J., Gajewski, B., Gustafson, K., Mundy, D., Yeast, J., . . . Shaddy, D. (2013). Docosahexaenoic acid supplementation and pregnancy outcomes. *The American Journal of Clinical Nutrition, 97*(4), 808-815.

CDC. (2012). Current cigarette smoking among adults--United States, 2011. *MMWR, 61*(44).

Chen, Z., Liu, A., Qu, Y., Tang, L., Ting, N., & Tsong, Y. (2015). *Applied statistics in biomedicine and clinical trial designs.* New York: Springer.

Congress. (2009, June 22, 2009). Family Smoking Prevention and Tobacco Control Act, Public Law 111-31, 123 Stat. 1776. Retrieved Jan 14, 2014, from http://www.gpo.gov/fdsys/pkg/PLAW-111publ31/pdf/PLAW-111publ31.pdf

Cox, D. (1958). The regression analysis of binary sequences (with discussion). *J Roy Stat Soc, 20*, 215-242.

Deyi, B. A., Kosinski, A. S., & Snapinn, S. M. (1998). Power considerations when a continuous outcome variable is dichotomized. *Journal of Biopharmaceutical Statistics, 8*(2), 337-352.

Djordjevic, M. V., & Doran, K. A. (2009). Nicotine content and delivery across tobacco products. *Handb Exp Pharmacol*(192), 61-82.

Fedorov, V. V., & Liu, T. (2007). Enrichment design. *Wiley Encyclopedia of Clinical Trials*.

Gajewski, B. J., Reese, C. S., Colombo, J. A., & Carlson, S. E. (2016). Commensurate priors on a finite mixture model for incorporating repository data in clinical trials. *In press.*

Gordon, L. (2013). *Using Classification and Regression Trees (CART) in SAS Enterprise Miner For Application in Public Health*. Paper presented at the SAS Global Forum. http://support.sas.com/resources/papers/proceedings13/089-2013.pdf

Gridgman, N. T. (1970). A comparison of two methods of analysis of mixtures of normal distributions. *Technometrics, 12*, 223-239.

Harper, M., Thom, E., Klebanoff, M. A., Thorp, J., Sorokin, Y., Varner, M. W., . . . Network, E. K. S. N. I. o. C. H. a. H. D. M.-F. M. U. (2010). Omega-3 fatty acid supplementation to prevent recurrent preterm birth: a randomized controlled trial. *Obstet Gynecol, 115*(201), 234-242.

Harrell, F. E. (2001). *Regression modeling strategies: with application to linear models, logistic regression, and survival analysis*. New York: Springer-Verlag.

Hoaglin, D., Mosteller, F., & Tukey, J. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

House, J. N. (2014). Preterm birth rate still too high: Q & A with Jennifer L. Howse, PhD, the March of Dimes. Retrieved Feb 2, 2016, from http://www.rwjf.org/en/culture-of-health/2014/11/u-s-_preterm_birthr.html

Jakka, S., & Rossbach, M. (2013). An economic perspecitive on personalized medicine. *The HUGO Journal, 7*, 1.

Jolly, D. (2008). Exploring the use of little cigars by students at a historically black university. *Prev Chronic Dis, 5*(3), A82. doi: A82 [pii]

Kasza, K., Bandal-Travers, M., & O'Connor, R. (2014). Cigarette smoker's use of unconventional tobacco products and associations with quitting activity: findings from the ITC-4 U.S. cohort. *Nicotine Tob Res, 16*(6), 672-681.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *J. Educat. Psychol, 22*(1), 45-55.

Leininger, T. J., Reese, S. C., FEllingham, G. W., & Grimshaw, S. D. (2010). An adaptive Bayesian approach to continuous dose-response modeling. Retrieved Dec 9, 2015, from https://stat.duke.edu/~tjl13/papers/TJL_MSthesis.pdf

Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *The Society of Behavioral Medicine, 26*(3), 172-181.

Loh, W. Y. (2011). Classification and Regression Trees. *Discovery*(1), 14-23.

Maathuis, M. (2007). Role of statistical models. Retrieved Dec 29, 2015, from https://stat.ethz.ch/~maathuis/teaching/stat423/handouts/Chapter1.pdf

Makrides, M., Duley, L., & Olsen, S. F. (2006). Marine oil, and other prostaglandin precursor, supplementation for pregnancy uncomplicated by pre-eclampsia or intrauterine growth restriction. *Cochrane Database Syst Rev, 19*(3), CD003402.

Makrides, M., Gibson, R. A., Mcphee, A. J., Yelland, L., Quinlivan, J., Ryan, P., & Team, D. I. (2010). Effect of DHA supplementation during pregnancy on maternal depression and neurodevelopment of young childern: a randomized controlled trial. *JAMA, 304*(15), 1675-1683.

Martin, J., Hamilton, B. E., Osterman, M. J., Curtin, S. C., & Mathers, T. J. (2013). Birth: Final data for 2012. National vital statistics reports, 2013. (Vol. 62). Hyattsville, MD: National Center for Health Statistics.

McGrath, D. S., Temporale, K. L., Bozec, L. J., & Barrett, S. P. (2011). Polytobacco use in non-daily smokers: an issue requiring greater attention. *Prev Med, 53*(4-5), 353-354.

McLachlan, G., & Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem. *J Am. Stat. Assoc., 1*(58), 275-309.

Orzechowski, W., & Walker, R. C. (2011). The tax burden on tobacco (Vol. 46). Arlington, VA: Orzechowski and Walker Economic Consulting Firm.

Padmanabhan, S. (2014). *Handbook of Pharmacogenomics and Stratified Medicine.* (1st ed.). Waltham, Massachusetts: Academic Press.

Page, J., & Evans, S. (2003). Cigars, cigarillos, and youth: Emergent patterns in subcultural complexes. *Journal of Ethnicity in Substance Abuse, 2*, 63-76.

Peacock, J., Sauzet, O., Ewings, S., & Kerry, S. (2012). Dichotomising continuous data while retaining statistical power using a distributional approach. *Statistics in Medicine, 31*(26), 3089-3103.

Piper, M. E., Loh, W. Y., Smith, S. S., Japuntich, S. J., & Baker, T. B. (2011). Using decision tree analysis to identify risk factors for relapse to smoking. *Subst Use Misuse, 46*(4), 492-510.

Popova, L., & Ling, P. (2013). Alternative tobacco product use and smoking cessation: a national study. *Am J Public Health, 103*(5), 923-930.

Ramakrishnan, U., Stinger, A., DiGirolamo, A. M., Martorell, R., Neufeld, L. M., Rivera, J. A., . . . Wang, M. (2015). Prenatal docosahexaenoic acid supplementation and offspring development at 18 months: randomized controlled trial. *PLoS ONE, 10*(8), e0120065.

Refaeilzadeh, P., Tang, L., & H., L. (2008, 6/11). Cross Validation. Retrieved Dec 18, 2013, from http://www.cse.iitb.ac.in/~tarung/smt/papers_ppt/ency-cross-validation.pdf

Richardson, A., Xiao, H., & Vallone, D. M. (2012). Primary and dual users of cigars and cigarettes: profiles, tobacco use patterns and relevance to policy. *Nicotine Tob Res, 14*(8), 927-932.

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

Ruggeri, F., Kenett, R., & Faltin, F. (2008) *Encyclopedia of statistics in Quality and Reliability* (Vol. 1, pp. 315-323): Wiley.

Russell, R. B., Green, N. S., Steiner, C. A., Meikle, S., Howse, J. L., Poschman, K., . . . Petrini, J. R. (2007). Cost of hospitalization for preterm and low birth weight infants in the United States. *Pediatrics, 120*(1), e1-e9.

Sauzet, O., Ofuya, M., & Peacock, J. L. (2015). Dichotomisation using a distributional approach when the outcome is skewed. *MBC Medical Research Methodology, 15*, 40.

Schilling, M. F., Watkins, A. E., & Watkins, W. (2002). Is human height bimodal? *The American Statistician, 56*(3), 223-229.

Schnerider, F. (1993). *Fistributed Systems* (S. Mullender Ed. 2nd ed.). Boston: Addison-Wesley.

Schwartz, S. L., Gelfand, A. E., & Miranda, M. L. (2010). Joint bayesian analysis of birthweight and censored gestational age using finite mixture models. *Statistics in Medicine, 29*(16), 1710-1723.

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

Swan, G. E., Javitz, H. S., Jack, L. M., Curry, S. J., & McAfee, T. (2004). Heterogeneity in 12-month outcome among female and male smokers. *Addiction, 99*(2), 237-250.

Timofeev, R. (2004). *Classification and regression trees (CART) theory and applications.* (Master of Art), Humboldt University, Berlin. Retrieved from http://edoc.hu-berlin.de/master/timofeev-roman-2004-12-20/PDF/timofeev.pdf

U.S. Department of Health and Human Services. (2014). Examination of clinical trial costs and barriers for drug development. Retrieved Dec 25, 2015, from https://aspe.hhs.gov/report/examination-clinical-trial-costs-and-barriers-drug-development

U.S. Food and Drug Administration. (2012). Guidance for industry: enrichment strategies for clinical trials to support approval of human drug and biological products. Retrieved Dec 24, 2015, from http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm332181.pdf

White House. (2015). Fact sheet: New patient-focused commitments to advance the presidnet's precision medicine initiative. Retrieved Dec 23, 2015, from https://www.whitehouse.gov/the-press-office/2015/07/08/fact-sheet-new-patient-focused-commitments-advance-president%E2%80%99s-precision