

Copyright
by
Yisi Wang
2015

The Report Committee for Yisi Wang
certifies that this is the approved version of the following report:

**A Comparison of Multiple Approaches to Subgroup
Analysis in Clinical Trial**

APPROVED BY

SUPERVISING COMMITTEE:

Peter Mueller, Supervisor

Mike Daniels

**A Comparison of Multiple Approaches to Subgroup
Analysis in Clinical Trial**

by

Yisi Wang, B.S.

REPORT

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN STATISTICS

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2015

A Comparison of Multiple Approaches to Subgroup Analysis in Clinical Trial

Yisi Wang, M.S.STAT.

The University of Texas at Austin, 2015

Supervisor: Peter Mueller

In randomized clinical trials, medical researchers are interested to determine the effectiveness of a new treatment not only in the overall population but also to some subgroups with possible enhanced treatment effects. However, subgroup analysis may become problematic due to the issue of multiplicities, data dredging etc. Accounting for these issue, we summarized some guidelines on the use and interpretation of subgroup analysis. We reviewed three approaches to subgroup analysis, a traditional Bayesian regression with interaction terms, the 'Virtual Twins' methods and a Bayesian model selection approach. The advantage and disadvantage of these three approaches are discussed.

Table of Contents

Abstract	iv
List of Tables	vi
List of Figures	vii
Chapter 1. Introduction	1
Chapter 2. Simulation Example	4
2.1 Simulated Data	4
2.2 Evaluation Criteria	5
Chapter 3. Traditional Bayesian	8
3.1 Bayesian Model Construction	8
3.2 Summarizing Results	9
Chapter 4. Virtual Twins method	11
4.1 Virtual Twins Strategy	11
4.1.1 Step 1: Apply random forests to data	12
4.1.2 Step 2: Estimate a Regression or Classification Tree	12
4.2 Evaluating subgroup regions	13
Chapter 5. Bayesian Model Selection	15
5.1 Bayesian Models Construction for Overall and Subgroup Effects	16
5.2 Probability Distribution for Models	18
5.3 Bayesian Model Selection Procedure	20
Chapter 6. Discussion	23
Bibliography	25

List of Tables

5.1	Subgroup models in Model space \mathbb{M}_x	18
5.2	Probabilities for Models in Model space \mathbb{M}_x	19

List of Figures

4.1	Virtual Twins Regression Tree in Simulation Study Scenario 3	13
-----	--	----

Chapter 1

Introduction

In randomized clinical trials, medical researchers are interested to investigate the effectiveness of a new treatment. Ideally, a clinical trials are designed to determine the effectiveness of the treatment respect to the overall population. However, the patient populations is often heterogeneous. The subjects of a clinical trial may vary by age, sex, severity and duration of the diseases. It is plausible that the new treatment is especially effective for some subgroups of the population. It is also likely that the new treatment is not effective or not effective in certain subgroups. In clinical trials, the subjects are often recruited based on a number of covariates or baseline variables, such as age or gender. The question arises as to find the subgroups defined by the covariates if they exist. This kind of analysis is referred to as subgroup analyses.

A subgrouping is a partition of the set of all patients into subgroups. It is usually determined by a measurable covariate. However, analyzing many subgroups can greatly increase Type I error rates, which may lead to invalid conclusion. There is an extensive literature about the dangers of subgroup analyses.

Berry (1990) defines two approaches for related problems. Type I approaches adjust significance levels to account for multiple analyses. Type II approach calculate nominal or unadjusted significance levels.

Berry points out two related Type I concerns about subgroup analyses: multiplicities and unplanned analyses (or data dredging). Multiplicities refer to the effect of examining many subgroups, that is testing multiple subgroups condition on the hypothesis of no subgroup effects. If there is no overall effect observed, there are always some subgroups that show significance just by chance when conducting multiple usual statistical test. Adjusting significance levels for multiple tests helps alleviate the multiplicity problems. Adjusting significance levels has the same effects as shrinking observations towards common mean in Bayesian methods. Both approaches make it harder to reject the null hypothesis of the mean treatment effect. However, adjusting significance levels for multiple tests may be inconsistent with the likelihood principle and the scientific method. Data dredging refers to a process where the subgroups or covariates are not specified before the initial analysis. People with different prior knowledge may draw completely different conclusion from the same observed data. There is a trade off between prior knowledge and the level of statistical significance. Some with strong prior information may only require moderate statistical significance. Whereas, one with weak prior information require stronger significance. Some statisticians insist that one should not test any hypothesis that is not pre-planned. However, it is hard for a scientist to proceed with any research without briefly taking a look at the data in reality.

In summary, there is no easy solution that accounts for all the issues related to subgroup analysis.

In order to avoid misuse and misunderstanding of subgroup analysis, some guidelines are introduced on the use and interpretation of subgroup analysis. Most of guidelines are covering the following points. Subgroups should be pre-specified, supported by biological reasoning or previous studies' result or base on pre-randomization characteristics; Analysis should be only applied to a small number of important subgroups; Significant level must be adjusted for multiple testing; all subgroups must be reported separating the prior planned group from the post hoc groups explicitly.

In this paper, we will introduce a simulation study set up and discuss evaluation criteria in Chapter 2. In Chapter 3-5, we will discuss three approaches for subgroup analysis. The traditional Bayesian regression with treatment and subgroup interaction will be introduced in Chapter 3. In Chapter 4, we will introduce the Virtual Twins methods for searching for enhanced treatment effect in subgroups. In Chapter 5, we will introduce a Bayesian model selection approach to identify and report subgroups. In Chapter 6, we discuss the advantages and disadvantages of the three approaches and conclude the discussion about subgroup analysis.

Chapter 2

Simulation Example

2.1 Simulated Data

In order to better demonstrate the difference between models, a simulation study is designed with linear combination of possible main treatment effects and subgroup effects. Let $n = 100$ denote the number of patients in the study, $i = 1 \dots n$. Let $p = 3$ denote the total number of candidate covariates. Let x_k be a covariate, $k = 1, \dots, 10$. Assume x_{ik} is binary, that is, $x_{ik} \in \{0, 1\}$. covariate values x_{ik} were generate by $p(x_{ik} = 1) = 0.5$. Let $T_i \in \{0, 1\}$ denote the binary treatment indicator with probability $p(T_i = 1) = 0.5$, which represents a randomized study). The Outcome was modeled with a linear combination of treatment covariate and candidate covariates as following:

$$y_i = x_i' \beta + t_i \alpha + x_i x_i' \gamma + \epsilon_i$$
$$\epsilon_i \sim N(0, 1)$$

β is not of particular interest in the subgroup analysis problem. Here we restrict $\beta = (1, 1, -1)$

We design 6 different scenarios of assumptions of γ and α .

1. Scenario : $\alpha = 0, \gamma = 0$
 H_0 is true, no treatment effect
2. Scenario : $\alpha = 1, \gamma = 0$
 H_1 is true, a strong overall effect
3. Scenario : $\alpha = 0, \gamma = (1, 0, 0)$
One winning subgroup; no overall effect
4. Scenario : $\alpha = 0.25, \gamma = (1, 0, 0)$
A small overall effect and a strong subgroup effect
5. Scenario : $\alpha = 0, \gamma = (0.25, 1, 0)$
A moderate and a strong subgroup effect
6. Scenario : $\alpha = 0.25, \gamma = (0.25, 0, 0)$
A small overall effect and a strong subgroup effect

2.2 Evaluation Criteria

Statistical estimation approximates the unknown parameters based on measured data. Point estimation of an unknown parameter is calculated based on the sample data and serves the purpose of an best guess. In frequentist inference, the bias of an estimator is a crucial criteria to evaluate the optimality of an estimator. The bias of an estimator is the expected value of the difference between the estimator and the true value of the parameter. An true value of the parameter was implicitly assumed when considering the bias of an estimator.

An estimator with zero bias is called unbiased estimator.

$$E[(\hat{\mu}(y) - \mu)] = 0$$

In Bayesian statistics, there is no single truth about an unknown parameter. The truth of a parameter is expressed in terms of degrees of belief or a probability distribution. The posterior distribution of the unknown parameter is calculated after updating the belief using observed data. The posterior mean or posterior median are robust point estimators of the unknown parameter.

In clinical trials, estimating an unknown quantity is not the only interest. Scientist also want to test the correctness of an assumption about a population parameter based on the observed data. Statistical hypothesis testing is also a major area in statistical inference. A test is called statistically significant if the null hypothesis is proven to be unlikely given the observed data. The process of testing null hypothesis and alternative hypothesis are closely related to two types of errors type 1 error and type 2 error. Type 1 error is that null hypothesis is rejected based on the rejection criteria when the null hypothesis is true. Type 2 error is that null hypothesis fails to be rejected based on the rejection criteria when null hypothesis is false. Type 1 and type 2 error also refer to the terms as false positives and false negatives in binary classification in medical testing. In simple hypothesis testing case, the type 1 error usually equal to the significance level of a test α . Type 2 error denoted by β is closed related to the the power of a test, explicitly $1 - \beta$.

In subgroup analysis setting, we assume the null hypothesis (H_0) is that

there is no overall treatment effects and the alternative hypothesis (H_a) is that there is an overall treatment effects. As an result, the subgroup effects (H_i) lays in between null hypothesis and alternative hypothesis. We can define the following error rates as the operating characteristics.

TIE: Type I Error = $P(H_0|H_0)$

FNR: False Negative Rate = $P(H_0|H_a)$

FSR: False Subgroup Rate = $P(H_i, i \neq 0, i \neq a|H_a)$

TSR: True Subgroup Rate = $P(H_i|H_i, i \neq 0, i \neq a)$

FPR: False Positive Rate = $P(H_a|H_i, i \neq 0, i \neq a)$

Here $P(A|B)$ denotes the probability under repeat experimentation of reporting A where B is the simulated truth. We will use these operating characteristics to evaluate the models that we will introduce in the later chapters.

In Bayesian statistics, Bayes factor can be used as an alternative to frequentist hypothesis testing. Bayes factor is often used in multiple testing to determine which model is more reflected by the observed data. It is calculated as the ratio of the marginal likelihood of the data given the models, which represent the possibility that the observed data are produced under given model.

$$B = \frac{p(y|M_1)}{p(y|M_2)} = \frac{\int p(y|\theta, M_1)p(\theta|M_1)}{\int p(y|\theta, M_2)p(\theta|M_2)}$$

However, Bayes factor in general are hard to calculated and the marginal likelihood of the data under model has to be proper.

Chapter 3

Traditional Bayesian

3.1 Bayesian Model Construction

Dixon and Simon(1991) introduced a Bayesian method to exam treatment effects based on the levels of covariates that define subset. Two treatment and binary covariates was used for illustration purpose. We will demonstrate the method using a linear model as following:

$$y_i = t_i\alpha + x_i\beta + \gamma t_i x_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

This linear model includes overall treatment effects $t_i\alpha$, covariate based effects $x_i\beta$ and the treatment and covariates interaction term $\gamma t_i x_i$ representing subgroup effects. Let θ denote a $2p + 1$ a linear combination of (α, β, γ) , which are the parameters corresponding to (t, x, tx) .

Dixon and Simon(1990) assume that $\gamma \sim N(0, \xi^2 I)$ and imposed a modified Jeffreys prior for variance parameter ξ^2 . A flat informative prior for α and β was used. To simplify the model, we use a multivariate normal prior for the vector γ . Define $\hat{\theta}$ an estimate of θ obtained by the maximum partial likelihood method. Then, the model can be described using the following assumption:

1. The maximum likelihood estimate θ has approximately normal distribution. $\hat{\theta} \sim N(\theta, C)$ where C is a known covariance matrix.
2. A prior for θ follows a multivariate normal distribution

with above assumptions, the posterior distribution of θ is also normal:

$$\theta|\hat{\theta} \sim N(Bb, B) \tag{3.1}$$

where $B^{-1} = C^{-1} + D^{-1}$ and $b = C^{-1}\hat{\theta}$

With flat prior for α and β , D^{-1} becomes a diagonal matrix with main diagonal as

$$\begin{aligned} & \left(\frac{1}{\sigma_\alpha^2}, \frac{1}{\sigma_{\beta_1}^2}, \frac{1}{\sigma_{\beta_2}^2}, \dots, \frac{1}{\sigma_{\beta_p}^2}, \frac{1}{d_1}, \frac{1}{d_2}, \dots, \frac{1}{d_p} \right) \\ & = (0_{p+1}, \frac{1}{d_1}, \frac{1}{d_2}, \dots, \frac{1}{d_p}) \end{aligned}$$

where p is the number of covariates, 0_{p+1} is a vector of $p + 1$ zeros. d_i corresponds to the prior variance of the i th treatment-by-covariate interaction effect.

Define a $2p + 1$ vector W as linear combined vector of (t, x, tx) . matrix C was derived from maximum partial likelihood as $C = (W^T W)^{-1} \sigma^2$.

3.2 Summarizing Results

It is controversial on how to choose a quantity to summarize results in Bayesian inference problems. Dixon and Simon use the marginal 95% highest

posterior density interval. The 95% highest posterior density joint region is also a reasonable quantity to use. The idea of searching for a region with high joint probability help to avoid multiplicity effects. In our case, the target posterior distribution follows a multivariate normal distribution. Thus the marginal variance of the parameter of interest lay on the diagonal of posterior covariance matrix.

Chapter 4

Virtual Twins method

The traditional Bayesian method is not very efficient when the dimension of X gets larger. Foster, Taylor and Ruberg developed a standard, predetermined strategy, referred to as 'Virtual Twins', to find subgroups enhanced treatment effect. The method is particularly applicable for two situation: (1) a new treatment is shown overall slightly better than the standard treatment but there is not enough strong evidence to be widely adopted and (2) in some situation, the new treatment may appear better, and the new treatment might be especially effective to some subgroups of patients.

4.1 Virtual Twins Strategy

The same problem set up and the same notations follows from the last chapter. The method is aimed to partition the covariate space into two regions, a small subset of X s A and complementary set A^c . Enhanced treatment effect was identified in region A comparing to region A^c . This approach assume there are two possible outcomes for each patient (one under treatment and one under control group). We are interested in the difference between these two outcome.

4.1.1 Step 1: Apply random forests to data

A random forest is an ensemble predictor constructed by multiple regression or classification trees. The random forest can be view as a black box predictor that take covariate values as and gives desired estimators as output for that input. For our purpose, the input covariate values are x_i, t_i and possible terms such as $x_i I(t_i = 0)$ and $x_i I(t_i = 1)$. The output is an estimator of y_i . If the subject i is in actual treatment group j , let \hat{Y}_i^j denote the out-of-bag estimate from the random forest, whereas let \hat{Y}_i^{1-j} denote the estimate applying the random forest with the other treatment group $1 - j$. Define $Z_i = \hat{Y}_i^1 - \hat{Y}_i^0$, which is a measurement of the treatment effects of subject i . R function `randomForest` was used to fitting the random forest.

4.1.2 Step 2: Estimate a Regression or Classification Tree

Now, a regression or classification tree can be used to find a small number of covariates that strongly correlate with Z . A threshold c is set for Z_i , which we can use to define A .

Virtual Twins regression tree was estimated using Z as the response variable and X as covariates. Intuitively, if predicted Z_i for subject i is greater than threshold c , subjects i should be grouped into \hat{A} . Thus, \hat{A} can be defined by the paths along with tree to the end node such that predicted Z_i is greater than c . If no predicted Z_i is greater than c , \hat{A} is empty. Figure 4.1.2 illustrates the Virtual Twins regression tree in Simulation Study Scenario 3 setting. If We set c equals to 1.158, Then subgroup defined by $x_1 < 0.5$ and $x_3 < 0.5$ are

the preferred subgroup.

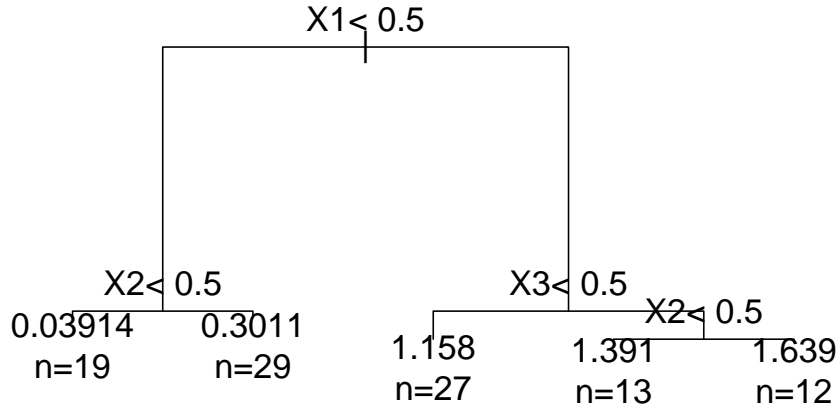


Figure 4.1: Virtual Twins Regression Tree in Simulation Study Scenario 3

To apply Virtual Twins classification tree, we design a new binary outcome measure Z^* defined as $Z_i^* = 1$ if $Z_i > c$ and $Z_i^* = 0$ if $Z_i < c$. Z^* is used as outcome in fitting the classification tree. Different from regression tree, all nodes of covariates X s in the classification tree defined set \hat{A} . If the classification tree has no splits, \hat{A} is empty.

4.2 Evaluating subgroup regions

In order to get a quantitative measurement on the enhanced treatment effect on a particular covariates region A with the average treatment effect on

the overall population, define

$$Q(A) = \frac{\sum_{x \in A} (Y_i^1 - Y_i^0)}{N_A} - \frac{\sum (Y_i^1 - Y_i^0)}{N}$$

where N_A is the number of subjects in region A and N is the total number of subjects. $Q(A)$ is defined as 0 if A is a empty set. $Q(\hat{A})$ quantifies the effectiveness of the recommended region \hat{A} in defining a region of enhanced treatment effects. The larger $Q(\hat{A})$ is, the more useful the region \hat{A} is. In simulation study, $Q(A)$ can be used to compare the efficiency of different subgroups.

Chapter 5

Bayesian Model Selection

Sivaganesan, Laud and Muller developed a Bayesian model selection approach to identify and report subgroup effects. The overall null hypothesis is that the treatment is not effective among all the study population. The overall alternative hypothesis is that the treatment is effective. The subgroups of interest are defined by covariates such as gender, age and biomarkers. Each subgroup effect model is only concerned about one single covariate in this setting. The overall hypothesis was tested at first. The overall effect model is compared with all the subgroup effect models. The overall effect model is selected when a strong overall treatment effect is observed than any other subgroup effect. If the overall effect model is not selected at this stage, a Bayesian model selection procedure will be conducted to search for prevailing subgroups continued. The posterior probability of subgroup effects models are computed for each single covariate. Fixed thresholds are set before the procedure to determine whether to select a subgroup model. The threshold is set to match the desired operating characteristics. Bayesian models with subjective prior were used in defining the overall and subgroup models. The authors use a zero-enriched Polya Urn scheme to define the prior probability distribution for model space. We will use a simplified prior distribution for

model space for the purpose of illustrating the concepts of Bayesian model selection.

5.1 Bayesian Models Construction for Overall and Subgroup Effects

First overall effects model is defined in usual regulatory setup to determine whether treatment effects is effective among overall study population. Assume M_{00} is the overall null hypothesis that there is no treatment effect, whereas M_{01} is the overall alternative hypothesis that there is a overall treatment effect.

$$M_{00} : \delta = \mu_1 - \mu_0 = 0 \quad M_{01} : \delta = \mu_1 - \mu_0 \neq 0$$

We define the null model space as

$$\mathbb{M}_0 = \{M_{00}, M_{01}\}$$

The null hypothesis can be set to quantity other than 0, for example, an interval or an point null with reasonable effect sizes. The method will remain equally valid. We will use the point null hypothesis for simplicity purpose.

Then we discuss the construction and indexing of subgroup models here. We will follow the problem setup in simulation study of chapter 2 for the rest of discussion. We have total $K = 3$ candidate binary covariates. Consider a single covariate x with two possible value s with $s \in \{0, 1\}$. Let μ_{0s} , μ_{1s} denote the mean outcome under control and treatment respectively. Define

the treatment effect in subgroup s as

$$\delta_s = \mu_{1s} - \mu_{0s} \quad \text{for } s = 1, 0$$

The goal is to identify the subgroups with no treatment effects and the subgroups with enhanced treatment effects. Moreover, we want to evaluate the size of enhanced treatment effects among different subgroups. Thus we are interested in testing whether δ_s is equal to 0. If δ_s is not equal to 0 meaning there is a subgroup treatment effect, we want test if there is a different treatment effect in δ_s for $s = 1, 0$. Let \mathbb{M}_x denote the all subgroup models for the covariate x . We use the all the possible combination of δ_s to define the subgroup model in \mathbb{M}_x . It's clear that there are total 5 subgroup models in our problem setting, which are $M_0 : \delta_1 = \delta_0 = 0$, $M_1 : \delta_1 \neq 0, \delta_0 = 0$, $M_2 : \delta_1 = 0, \delta_0 \neq 0$, $M_3 : \delta_1 \neq \delta_0 \neq 0$, and $M_4 : \delta_1 = \delta_0 \neq 0$. In order to keep consistency with simulation study setup, we match 5 subgroup models above into linear likelihood probability models with normality assumption imposed. Linear models setup make sure that the derivation of the posterior distribution can be done analytically. The detail of this mapping listed in table 5.1. Thus the sampling distributions of subgroup models can be summarized into independent normal distributions. we assign non-informative normal conjugate prior distribution for θ and σ^2 , which assign a inverse gamma prior for σ^2 and multivariate normal prior for $\theta|\sigma^2$

$$y_i \sim N(X_i\theta, \sigma^2)$$

$$\sigma^2 \sim Inv - Gamma(a_0, b_0) \quad \theta|\sigma^2 \sim N(\mu_0, \Sigma)$$

Table 5.1: Subgroup models in Model space \mathbb{M}_x

Model	$P(y_i \theta)$
$M_0 : \delta_1 = \delta_0 = 0$	$y_i = \beta_0 + \epsilon_i$
$M_1 : \delta_1 \neq 0, \delta_0 = 0$	$y_i = \beta_0 + x_i\beta + \epsilon_i$
$M_2 : \delta_1 = 0, \delta_0 \neq 0$	$y_i = \beta_0 + \gamma t_i x_i + \epsilon_i$
$M_3 : \delta_1 \neq \delta_0 \neq 0$	$y_i = \beta_0 + x_i\beta_1 + \gamma t_i x_i + \epsilon_i$
$M_4 : \delta_1 = \delta_0 \neq 0$	$y_i = t_i\alpha + x_i\beta + \gamma t_i x_i + \epsilon_i$

As a result, the posterior distribution of the parameters θ and σ^2 given a specific subgroup model M can be derived analytically. The posterior distribution for θ follows a multivariate normal distribution and the posterior distribution for σ^2 follows a inverse-gamma distribution.

$$P(\theta, \sigma^2|y, M) = \frac{P(y|\theta, \sigma^2, M)P(\theta, \sigma^2|M)}{\int_{\theta} \int_{\sigma^2} P(y|\theta, \sigma^2, M)P(\theta, \sigma^2|M)d\theta d\sigma^2}$$

5.2 Probability Distribution for Models

Given a specific covariate x , we have 5 subgroup models associated with it as we discussed in last section. Each subgroup models determined whether a subgroup effects is zero or not. The authors assign probabilities for models using the Polya Urn methods. Here we design a simplified version to be consist with our simulation study set-up. We assign a probability to each model in model space \mathbb{M}_x listed in table 5.2.

To be noted that M_0 in model space \mathbb{M}_x corresponding to the overall null model and M_1 in model space \mathbb{M}_x corresponding to the overall alternative model. Overall models was assigned a total p probability and the rest of

Table 5.2: Probabilities for Models in Model space \mathbb{M}_x

Model	$P(M_i p)$
$M_0 : \delta_1 = \delta_0 = 0$	$\frac{p}{2}$
$M_1 : \delta_1 \neq 0, \delta_0 = 0$	$\frac{1-p}{3}$
$M_2 : \delta_1 = 0, \delta_0 \neq 0$	$\frac{1-p}{3}$
$M_3 : \delta_1 \neq \delta_0 \neq 0$	$\frac{1-p}{3}$
$M_4 : \delta_1 = \delta_0 \neq 0$	$\frac{p}{2}$

probability assigned evenly to each subgroup models in model space \mathbb{M}_x .

All the probabilities for models in model space \mathbb{M}_x have to add up to 1. Thus parameter p is restricted in between 0 and 1. Thus we use Beta prior for the parameter p .

$$p \sim \text{Beta}(p; a, b)$$

Consider the combined sample data y , $P(y|M, p) = P(y|M)$ enforced by conditional independence. Then we have the posterior probability of a model M_i in model space \mathbb{M}_x as

$$P(M|y) = \frac{P(y|M)P(M)}{\sum_{M' \in \mathbb{M}_x} P(y|M')P(M')}$$

In above, we have

$$P(M) = \int P(M|p)P(p)dp$$

$$P(y|M) = \int \int P(y|M, \theta, \sigma^2)P(\theta, \sigma^2|M)d\theta d\sigma^2$$

Consider the cases where there are more than 1 covariate of interest. First, we have the overall null model M_{00} and overall alternative model M_{01} , which constitute the null space \mathbb{M}_0 . Suppose we have I covariates of interest x_1, \dots, x_i . We define a model space for each covariate, for example, $\mathbb{M}_{x_i}\mathbb{M}_i$ for covariate x_i . In our problem set-up, we have 3 covariates of interest and there are $H = 5$ models in each of the covariate space for the corresponding 3 covariates. We index the model in a covariate spaces based on the order in table 5.2.

Suppose we are interested in calculating the posterior distribution for model $M \in \mathbb{M}_i$. Let $P_i(M) \equiv P(M|\mathbb{M}_i)$ denote the prior probabilities of M . Let $P_i(M|y) \equiv P(M|y, \mathbb{M}_i)$ denote the prior probabilities of M conditional on model space \mathbb{M}_i . For a specific model $M^* \in \mathbb{M}_i$

$$\bar{P}(M^*|y) = \frac{P(y|M^*)P_i(M^*)}{\sum_{M' \in \mathbb{M}_i} P(y|M')P(M')}$$

5.3 Bayesian Model Selection Procedure

We get the posterior distributions for models in multiple model spaces. Here we will introduce the Bayesian model selection procedure to search for the target models. The goal of this procedure is to compare overall effect model and to select the subgroup models if there exist a strong subgroup enhanced treatment effects. The algorithm for the stepwise procedure are as following:

- Step 0: Choose M_{01} and stop, if

$$\bar{P}_0(M_{01})/\bar{P}_0(M_{00}) > c_0 \quad \text{and,}$$

$$\bar{P}_i(M_{ih})/\bar{P}_i(M_{iH}) < c_1 \quad \text{for } 1 < i < I \quad \text{and } 0 < h < H,$$

Else continue to next step.

- Step i : for $i = 1, \dots, I$ Choose M_{ih} for $0 < h < H$, if

$$\bar{P}_i(M_{ih}) = \max_{0 < h' < H} \bar{P}_i(M_{ih'}) \quad \text{and}$$

$$\bar{P}_i(M_{ih})/\bar{P}_i(M_{i0}) > c_0 \quad \text{and}$$

$$\bar{P}_i(M_{ih})/\bar{P}_i(M_{iH}) > c_1$$

Continue to next step.

- Step $I + 1$: Choose M_{00} if no subgroup effect model is selected in the previous steps.

The first step of the procedure compare the overall null model and overall effect model. If the overall effect model is preferred, then overall effect model continue comparing with subgroup effect models. If there is no subgroup model is preferred over the overall effect model, the procedure stops and the overall effect model is selected. Otherwise, the procedure continues to search through the subgroup models in the model space \mathbb{M}_i for $i = 1, \dots, I$. One or more subgroup may be selected if they surpass the overall null model and the overall effect model to some extent. Specifically, we use two thresholds c_0 and c_1 for posterior odds. c_0 represents the threshold for the posterior odds of the

overall or a subgroup effect model against the null model, and c_1 represents the threshold for the posterior odds of a subgroup effect model against the overall effect model. If no subgroup effect model is selected during step $1, \dots, I$, the overall null effect model is selected at last. Thus, the final outcome of the procedure has three possibilities: the overall effect model M_{01} selected in step 0, one or more subgroup effects model M_{ih} selected in step $1, \dots, I$ or the overall null model M_{00} selected in step $I + 1$.

Chapter 6

Discussion

We present three approaches for doing subgroup analysis, a traditional Bayesian methods with treatment-covariate interaction, the Virtual Twins method and a Bayesian model selection approach. The traditional Bayesian with treatment-covariate interaction has the important feature that the estimates of the subgroup treatment effects is weight between the overall treatment effects and within-subgroup differences. Scientists can also specify a prior belief on the strength of the interaction. This method is generally easy to understand and implement. Many developments of this method extend the idea to a variety of models such as generalized linear models or proportional hazards models. The Virtual Twins method is developed for identifying the subgroups with possible enhanced treatment effects. It predicts the response outcome for treatment and control for each subject. The classification tree or regression tree is applied to the response outcome to determine the potential related subgroups. The Virtual Twins methods appears to be reasonably good in identify the magnitude of the enhanced treatment effect. However, it is not adept at identifying the correct covariates. It tends to identify more subgroups than it should. Many variations of this method may be studied in the future such as using other non-parametric regression methods instead

of the random forest in the first step of the procedure. The Bayesian model selection approach represents the treatment-subgroup interaction by subgroup models and calculated the posterior probability for overall models and the subgroup models. The determination of overall treatment effects or the subgroup treatment effect is made by comparing the ratios of the posterior probability of the models with some pre-determined thresholds. These threshold can be set to achieve a specific overall Type-I error rate. Appropriate thresholds can help to deal with multiplicity issues. The Bayesian model selection approach presented here can only account for subgroups that are defined by one covariate. If multiple covariates are needed to define the subgroups, some additional covariates can be created using the combination of the existing covariates. A similar approach can be extended to those cases. In general, identifying and estimating subgroups effects is hard and usually require large sample size or strong subgroup treatment effects.

Bibliography

- [1] James O Berger, Xiaojing Wang, and Lei Shen. A bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics*, 24(1):110–129, 2014.
- [2] Donald A Berry. Subgroup analyses. pages 1227–1230, 1990.
- [3] Wei Chen, Debashis Ghosh, Trivellore E Raghunathan, Maxim Norkin, Daniel J Sargent, and Gerold Bepler. On bayesian methods of exploring qualitative interactions for targeted treatment. *Statistics in Medicine*, 31(28):3693–3707, 2012.
- [4] David I Cook, Val J GebSKI, and Anthony C Keech. Subgroup analysis in clinical trials. *Medical Journal of Australia*, 180(6):289–292, 2004.
- [5] Dennis O Dixon and Richard Simon. Bayesian subset analysis. *Biometrics*, pages 871–881, 1991.
- [6] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, 2011.
- [7] Stuart J Pocock, Susan E Assmann, Laura E Enos, and Linda E Kasten. Subgroup analysis, covariate adjustment and baseline comparisons

- in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21(19):2917–2930, 2002.
- [8] Peter M Rothwell. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365(9454):176–186, 2005.
- [9] Richard Simon. Bayesian subset analysis: application to studying treatment-by-gender interactions. *Statistics in Medicine*, 21(19):2909–2916, 2002.
- [10] Siva Sivaganesan, Purushottam W Laud, and Peter Müller. A bayesian subgroup analysis with a zero-enriched polya urn scheme. *Statistics in Medicine*, 30(4):312–323, 2011.