Meta-analysis for continuous outcomes with a baseline and follow-up measurement

Chen, Z.

Meta-analysis for continuous outcomes with a baseline and follow-up measurement

Zixuan Chen

March 31, 2022

Abstract

Using individual participant data (IPD) has many advantages over using aggregate data (AD) in clinical meta-analysis. However, access to the IPD is often limited, yet the aggregate data is available from most clinical trials. Papadimitropoulou's et al. [4] propose a method for studies with continuous outcomes at baseline and follow-up measurement to generate pseudo-IPD from the aggregate data, which can be analyzed as IPD, using analysis of covariance (ANCOVA) models and linear mixed models. The pseudo-IPD is generated based on the mean, standard deviation at baseline and follow-up, and the correlation between baseline and follow-up, which are sufficient statistics of the linear mixed model. This thesis exemplified the pseudo-IPD models, standard meta-analysis models, and a Trowman meta-regression model on Obstructive Sleep Apnea Data with 2 treatment groups. We further explored the performance of the models under different conditions by a simulation study. The estimates of the Trowman meta-regression suffered from significant variance, and the standard AD models provided bias estimation when baseline imbalance exists. The ANCOVA models for pseudo-IPD and AD offered more accurate and stable results. The pseudo-IPD ANCOVA model is the most preferred since it can account for baseline difference and interaction between treatment and baseline, and different residual structures can be used.

KEYWORDS:

Meta-analysis, pseudo individual participant data, aggregate data, analysis of covariance, simulation study,

Contents

Acknowledgement

My appreciation, first and foremost, goes to my supervisor Prof. Dr. Saskia le Cessie. She is the most patient teacher I have ever met. When I have any questions, she is always there. Saskia supervised me from basic questions to global research directions to ensure I was working on the right things. During the weekly meeting, Saskia supports me for knowledge consultation and spiritual encouragement. Without her guide on the background which I was unfamiliar with, I would fail many research goals.

Thanks to my parents for their unlimited and invaluable support and love in the 25 years. We are the most important parts of each other's life. Besides that, I want to thank my brother for his advice on my life choice. I also appreciate my grandparents. I hope they can always keep joy and health.

Finally, I want to thank every professor I met in the Netherlands. They help me to finish my study. And thank you to all the people I met in the Netherlands. The two-year study in this country is my life treasure.

Chapter 1

Introduction

The effect of a treatment or a drug can be studied by an experiment with multiple measurements, which is common in randomized controlled trials. Researchers can compare the measurements among the time points and the treatment groups to find their difference. For most treatments or drugs, many studies are performed in different places. And the results may be different in studies for many reasons, for instance, ethnic differences or some local restriction of the participants. Hence, researchers prefer synthesizing or comparing findings from multiple studies to acquire better results. The technic to combine results of multiple studies is called **Meta-Analysis**. In this thesis, we will discuss Meta-Analysis methods for randomized clinical trials, with two treatment groups, continuous outcomes, and two measurements of the outcomes at baseline and follow-up. We illustrate the details of the theories of the methods and compare their performance under different conditions with a simulation study.

1.1 Individual Participant Data and Aggregate Data

In the clinic trials, researchers record the information, e.g., the ages, symptom severity, sex of the patients, treatment received, and outcomes which are termed as **Individual Participant Data**. For instance, in a study of an Alzheimer's Disease treatment, the individual participant data is composed of the pre-treatment, and post-treatment measurements using the Alzheimer's Disease Assessment Scale(ADAS)[2], and the essential characteristics such as the treatments, ages, IQ, and sex of the patients.

The Aggregate data (AD) is composed of the summary statistics for each arm in each study, which can be obtained from publications or requested from the authors[5]. In the same Alzheimer's disease example, the AD could include the mean and standard deviation of the ADAS at baseline and follow-up, the mean change from baseline, the proportion of males, the mean and standard deviation of ages, and IQ in each treatment arm of each study. All the information can be derived from the IPD. Hence, the AD is the 'summary data' of the IPD.

Aggregate data suffers from problems such as a different representation in different studies(e.g., risk ratio verse odd ratio) and missing data. More importantly, the analysis of AD is difficult to detect how the individual covariates can modify the effect of our interest. Lacking information and an incorrect summary can cause poor estimation in the analysis of AD. The IPD is more reliable since the individual participant data can be analyzed across all studies.

1.2 Meta-analysis of Clinic Trials

Meta-analysis is the method to synthesize the results of several studies or publications, which is widely used in clinical research. Meta-analysis offers a summary result to describe the effect of the treatment or drugs used in the studies. When a narrative review of studies is conducted, many problems may emerge since the narrative review takes no consistency of the studies into account. For instance, the p-value of 0.0001 can either imply a significant effect or a small effect for a large sample size. That is, without considering the size of the study, the p-value can 'cheat' us. In the meta-analysis, researchers analyze the effect size directly instead of p-values. The primary interest of the meta-analysts is whether the studies are consistent in the estimation of the effect size.

In randomized clinical studies, the core is to compare outcomes between the treatment groups. For a continuous outcome, the patients are often measured at both baseline (pre-treatment) and follow-up (post-treatment), especially in the study of chronic conditions. Many statistics comparisons can be applied to assess the effects of the treatments in this situation. For example, the follow-up scores can reflect the condition of the patients in each group after the treatment. Therefore, the difference in the mean follow-up scores between the treatment group and the control group can measure the treatment effect. Additionally, the 'change score' calculated by subtracting the follow-up score and baseline score can represent the outcome change during the study time. Hence, the change score can compare the treatment effects between the groups. An alternative method is to perform an analysis of covariance to estimate the treatment effect, which can adjust the baseline imbalance and other patients' characters.

If the IPD is available, researchers can perform the meta-analysis using the complete data of all studies. The data can be analyzed using the family of linear mixed models, which has a lot of modeling flexibility in statistics software. For instance, Various stratified models can be applied to perform the analysis. There are two main approaches for the meta-analysis on the IPD, the one-stage approach, and the two-stage approach. All the participant-level data from all studies are analyzed simultaneously for the one-stage method. In contrast, the two-stage method conducts the process in two steps. First, estimate each study separately with linear regression. The second step is to apply a suitable standard meta-analysis method to synthesize the results acquired from the first step. The two-stage approach is considered a solid method since the standard meta-analysis in the second step is stable. However, the one-stage method has the advantage that it can model the impact of within-study variation. Most of the time, although they use different methods to estimate the summary effect, the estimation results are similar.[1]

When meta-analysts plan a meta-analysis to study the impact of drugs or treatments, they prefer IPD. However, often the access of IPD is not available. Instead, access to the aggregate data (AD) of a study is easier.

To be able to use the linear mixed model framework, when only aggregate data are available, Papadimitropoulou's et al. [4] propose a new method by generating a pseudo-IPD from the sufficient statistics of the linear mixed model. The AD should include all the sufficient statistics (The means, standard deviations, Correlations, and sample sizes). In this way, we can apply the IPD meta-analysis methods by using only the AD following two steps. First generating a pseudo-IPD and then applying IPD meta-analysis methods on the pseudo data. The relationship of the data and the methods is summarised in figure 1.1. However, the estimation process of the methods are various, especially the estimation of the random effects. Since various approaches recovered different information of the data, the conditions of which methods can produce a better estimation need to be explored further. The aim of this paper is to compare these methods on different conditions, and assess the performance of the pseudo-IPD approaches and standard meta-analysis approaches, and conclude which methods can be used in each condition.

1.3 Outline Structure of the Thesis

In Chapter 2, we will discuss the various meta-analysis models for studies with continuous outcomes measured at two time points and the background knowledge, such as the assumption of the models, the definitions and formulas of the parameters and statistics. In Chapter 3, an application of the models on a dataset, the Obstructive Sleep Apnea data is demonstrated. In Chapter 4, a simulation study of the models is produced. We assess the performance of the models under the conditions of various random structures, and parameter settings. In the Chapter 5, the results of the simulation study will be concluded. And also the findings and the connection of the models will be explained.

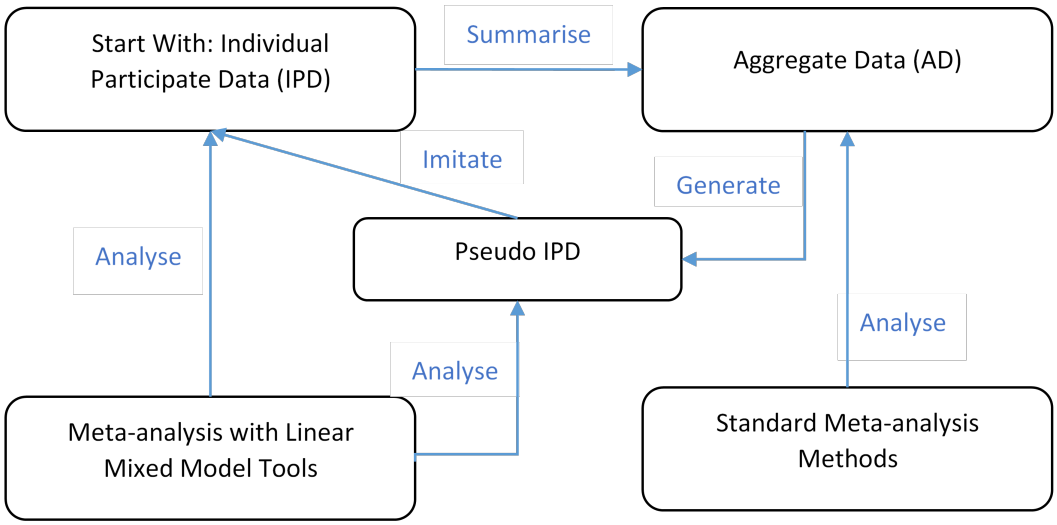

Figure 1.1: The relationship of various methods and data

Chapter 2

Models Explanation and Methodology

2.1 Definitions and Notations

We assume that there are n studies. In each study, two treatments are compared. We denote them with treatment 1 and 2. One of them is the treatment group which received the treatment of our interest. The other group received the standard or no treatment, with other conditions identical with the treatment group. We use the term 'control group' for this group The sample size two groups are n_1 and n_2, respectively. We assume that Y is the continuous outcome. It is measured at two time points, the baseline and Follow-up. We use the following notation Y_{Bij}, for the j-th person, in study i at baseline, and Y_{Fij} for the j-th person in study i, at follow up.

We assume that the following aggregate data is available: $\bar{Y}_{FTi}$, S_{FTi} are the mean, and standard deviation at follow up, in the treatment group in study i. $\bar{Y}_{FCi}$, S_{FCi} are the mean, and standard deviation at follow up, in the control group in study i. $\bar{Y}_{BTi}$, S_{BTi} are the mean, and standard deviation at baseline, in the treatment group in study i. $\bar{Y}_{BCi}$, S_{BCi} are the mean, and standard deviation at baseline, in control group in study i, the r_{iT} and r_{iC} are the correlation between the baseline and follow-up in study i in treatment group and control group respectively.

For convenience, we define the indicator variable X_{ij} that:

$$\begin{cases} X_{ij} = 1, & \text{if the patient } j \text{ in study } i \text{ is in the treatment group} \\ X_{ij} = 0, & \text{if the patient } j \text{ in study } i \text{ is in the control group} \end{cases} \tag{2.1}$$

Then the mean of the outcome can be calculated as:

$$\bar{Y}_{FTi} = \frac{\sum_{j,X_{ij}=1} Y_{Fij}}{n_1}$$

and

$$\bar{Y}_{FCi} = \frac{\sum_{j,X_{ij}=0} Y_{Fij}}{n_2}$$

where Y_{Fij} denotes the outcome of patient j at the follow-up measure in study i.

In the clinical trials, we expect to compare the treatment or drug effect in different groups. This is often done by using an overall summary measure of effect, the **effect size**. In the terminology, the effect size is a metric quantifying the relationship between the outcome and treatment groups. The relationship can be expressed by a mean difference, a ratio, a log-ratio, or for a binary outcome an odd-ratio among the groups. The effect size we choose should be comparable, computable, interpretable, and with good statistics properties.

For a continuous outcome, the mean outcomes difference between groups is often used. Denote the true treatment effect in study i is θ_i, and denote the estimate of the effect in study i by $\hat{\theta}_i$ and its standard error by σ_i.

2.2 Standard Meta-Analysis

2.2.1 Fixed Effect Models

When the standard meta-analysis is performed, various model assumptions are made depending on whether there is any evidence of heterogeneity in treatment effect among studies. If there is no heterogeneity, we can assume the treatment effects are identical among all studies in the fixed-effect model. That is,

$$\hat{\theta}_i = \theta + \epsilon_i \tag{2.2}$$

where $\hat{\theta}_i$ denotes the estimated treatment effect in study i. The true treatment effects are 'all exactly the same'. We denote the true effect as θ. The ϵ_i is the residual of the estimation in each study which has $\epsilon_i \sim N\left(0, {\sigma_i}^2\right)$. For this model, the estimation of θ is of primary interest.

2.2.2 Random Effect Model

If the heterogeneity exists, we assume that the true treatment effect is various in each study. Denote the true treatment effect in each study as θ_i, where $i \in 1, 2, ...n$. We call the model with the 'non-identical' assumption as the random effect model. The equation expression of the random effect model is:

$$\begin{aligned} \theta_i &= \theta + u_i \\ \hat{\theta}_i &= \theta_i + \epsilon_i \end{aligned} \tag{2.3}$$

The θ is the fixed value which denotes the average of the true treatment effects among all studies. The random effect model assumes that the treatment effect can vary between studies, with $Var(\theta_i) = \tau^2$. We can choose between model 2.2 and model 2.3 based on the specific circumstance. (i.e., The assumption of the treatment effect) If the treatment effect can assume to be identical among all studies, the fixed-effect model can be used. In contrast, if the experiments used various patient groups or some other evidence which implies heterogeneity among studies, it is more appropriate to use the random effect model to account for the difference among studies.

2.2.3 Raw Mean Difference

For continuous outcome variables, the means and standard deviations are recorded in the AD. We can use the difference of the mean score between the groups as the effect size to measure the effect of the treatment. We use the term **Raw Mean difference** to refer to the effect size. The raw mean difference is an intuitive way to interpret the treatment effect. To calculate the raw mean difference of two independent groups, we can use the formula:

$$D = \bar{Y_1} - \bar{Y_2} \tag{2.4}$$

where $\bar{Y_1}$ and $\bar{Y_2}$ are the sample mean of the two groups. D is the estimated effect size. Suppose the standard deviations of the two groups are identical. That is, $\sigma_1 = \sigma_2 = \sigma$. Denote S_1 and S_2 for the calculated standard deviation and n_1 and n_2 denote the sample size of each group. The variance of the effect size can be calculated as:

$$V_D = \frac{n_1 + n_2}{n_1 n_2} \times \frac{(n_1 - 1){S_1}^2 + (n_2 - 1){S_2}^2}{n_1 + n_2 - 2} \tag{2.5}$$

The standard error of D is

$$SE_D = \sqrt{V_D} \tag{2.6}$$

Without the assumption of $\sigma_1 = \sigma_2 = \sigma$. The variance can be calculated as:

$$V_D = \frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2} \tag{2.7}$$

Final Score Model

The final score method ignores the information from the baseline. It calculates the mean difference at follow up between the two groups. Then, we can use the model of the standard meta-analysis models. We can construct the **Final Score** treatment effect by using:

$$\hat{\theta}_i^{FS} = \bar{Y}_{FTi} - \bar{Y}_{FCi} \tag{2.8}$$

where the $\hat{\theta}_i^{FS}$ denotes the estimated treatment effect in study i using final score. The standard error of $\hat{\theta}_i^{FS}$ can be calculated using formula 2.5 and 2.6, or 2.7 and 2.6

Change Score Model

In the AD, the pre-treatment severity of the patients is recorded by the baseline score. Generally, the patients who are more severe ill before treatment are likely to be more severe ill after the treatment. That is, the values of the outcome before the experiment have impact on the treatment effect. For this reason, we can adjust the impact of baseline variable. An intuitive method is to summarise the 'Change Score' which is calculated by subtracting the baseline score from the follow-up score. Denote $\bar{Y}_{CTi}$, $\bar{Y}_{CCi}$ as the change score of the treatment group and control group in study i respectively. The formulas are:

$$\bar{Y}_{CTi} = \bar{Y}_{FTi} - \bar{Y}_{BTi} \tag{2.9}$$

and

$$\bar{Y}_{CCi} = \bar{Y}_{FCi} - \bar{Y}_{BCi} \tag{2.10}$$

With the standard deviation of the Y in the treatment group and control group, and the correlation between the baseline score and follow-up score, r_{Ti}, the standard deviation of the change score in the treatment group can be calculated as:

$$S_{CTi} = \sqrt{{S_{BTi}}^2 + {S_{FTi}}^2 - 2 \times r_{Ti} \times S_{BTi} \times S_{FTi}} \tag{2.11}$$

And in the same way, the standard deviation in the control group, S_{CCi} can be calculated as:

$$S_{CCi} = \sqrt{{S_{BCi}}^2 + {S_{FCi}}^2 - 2 \times r_{Ti} \times S_{BCi} \times S_{FCi}} \tag{2.12}$$

Follow the same rules as in the Final Score Model, the **Change Score** treatment effect is:

$$\hat{\theta_i^{CS}} = \bar{Y}_{CTi} - \bar{Y}_{CCi} \tag{2.13}$$

The notation $\bar{Y}_{CTi}$ and $\bar{Y}_{CCi}$ are as defined above. The $\hat{\theta_i^{CS}}$ denotes the estimated treatment effect of the Change Score Model. The standard error of $\hat{\theta}_i^{CS}$ can be calculated using formula 2.5, 2.6, and 2.7. The only change is to replace the follow-up score's standard deviations with the standard deviations of the change score which is calculated in formula 2.11 and 2.12.

2.3 ANCOVA

In the clinic research trials, researchers try to explain the impact of the treatment or drugs. However, often the dependent variable is depended on one or more other variables. It is possible to control for these variables to analysis the variation. This procedure is ANCOVA, which is the combination of regression analysis and the ANOVA.

The ANOVA method can analyze the variation source and divide the variation and degree of freedom to the within-group part and between-group part. Additionally, the regression analysis can estimate the impact of the independent variables. Hence, we can employ the regression analysis to adjust the impact of the variables in addition to our interest. We use the term 'covariates' to refer to these variables. We can remove the effect of the covariates by subtracting the variance from them. When only the variation from the primarily interested variables left, the ANOVA of the modified variation provides more reasonable results.

Recovered ANCOVA Model

Based on the ANCOVA theories, we introduce a standard meta-analysis model with adjustment of the baseline scores. We terms this model as **Recovered ANCOVA model**. The difference between the Change Score model and the Recovered ANCOVA Model is that in the recovered ANCOVA model we use a linear regression to adjust the baseline effect. The linear regression for study i is formulated as:

$$Y_{Fij} = \beta_{0i} + \theta_i^{ANCOVA} X_{ij} + \beta_i Y_{Bij} + \epsilon_{ij} \tag{2.14}$$

We assume that the variance of ϵ_{ij}, is the same in each study and each group. Denote S_{Fi} and S_{Bi} as the pooled standard deviations of the follow-up and baseline measurements in study i. Denote the correlation between Y_{Fij} and Y_{Bij} as r_i, assumed to be the same in the treatment and control group. $\hat{\beta}_i$ represents the regression coefficient for the baseline measurement, which can be calculated as:

$$\hat{\beta}_i = r_i \frac{S_{Fi}}{S_{Bi}} \tag{2.15}$$

The treatment effect in study i can be estimated by:

$$\hat{\theta}_i^{ANCOVA} = \left(\bar{Y}_{FTi} - \bar{Y}_{BTi}\right) - \hat{\beta}_i \left(\bar{Y}_{FCi} - \bar{Y}_{BCi}\right) \tag{2.16}$$

All the statistics can be calculated from the aggregate data. We can use standard meta-analysis methods to estimate the overall treatment effect using the models in Chapter 2.2

2.4 Trowman Model

An alternative method to adjust for the baseline measurement is to perform meta-regression. Trowman[6] performs a meta-regression with the mean follow-up score per treatment group as the outcome, the mean baseline score, and the treatment groups as the independent variables. That is, each study provides two observations. In addition, the interaction of the baseline and treatment group can be used for extension. We use the term **Trowman Model** to refer to the model. Suppose there are n studies and 2 treatment groups. The formula of the model without interaction is:

$$\bar{Y}_{Fik} = \beta_0 + \beta_1 X_{ik} + \beta_2 \bar{Y}_{Bik} + \epsilon_{ik} \tag{2.17}$$

where $\bar{Y}_{Fik}$ denotes the mean follow-up score of study $i \in 1, 2, ..., n$ and treatment $k \in 1, 2$. $\bar{Y}_{Bik}$ denotes the mean baseline score of study $i \in 1, 2, ..., n$ and treatment $k \in 1, 2$.

X_{ik} denotes the indicator variable with $X_{i1} = 1$ and $X_{i2} = 0$. ϵ_{ik} denotes the error term. The β_0, β_2, are the corresponding parameters of intercept and slopes, β_1 is the treatment effect.

2.5 Methods When IPD is Available

The IPD can be modeled on the participant level. There are two popular approaches, the one-stage method, and the two-stage method. The logic of the methods is to apply linear mixed model to the IPD.

2.5.1 One-stage Analysis IPD with LMM

The simplest one-stage model is to fit the linear mixed model with only the treatment effect. The **Base Model** can be written as:

$$Y_{Fij} = \beta_{0i} + (\beta_1 + b_{1i}) X_{ij} + \epsilon_{ij} \tag{2.18}$$

where b_{1i} denotes the random effect of the treatment effect on the study level. We assume b_{1i} follows a normal distribution with mean 0 and standard deviation τ. β_{0i}, is the study-specific intercept, β_1 is the overall treatment effect. We can remove the random treatment effect by constraining the standard deviation $\tau = 0$.

The ϵ_{ij} denotes the within-study individual variation which follows a normal distribution with mean 0 and standard deviation σ_{ik}. We can specify that the within-study variance depends on the

study by assuming $\sigma_{ik} = \sigma_{i.}$. Similarly, the variance can depend on the treatment group by $\sigma_{ik} = \sigma_{.k}$. Also, we can assume the simplest structure that all within-study variance is equal. That is, $\sigma_{ik} = \sigma$.

The baseline imbalance can be adjusted by introducing the baseline effect into the model. Hence, we can fit a ANCOVA model with stratified intercept and slope, which can be written as:

$$Y_{Fij} = \beta_{0i} + (\beta_1 + b_{1i})X_{ij} + \beta_{2i}(Y_{Bij} - \bar{Y}_{Bi}) + \epsilon_{ij} \tag{2.19}$$

where $\bar{Y}_{Bi}$ denotes the mean of the baseline score in study i. Hence, $(Y_{Bij} - \bar{Y}_{Bi})$ denotes the centered baseline score.

Alternatively, the intercept β_{0i} and baseline effect β_{2i} can be assumed to be random variables. Then the model 2.19 will become:

$$Y_{Fij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\, X_{ij} + (\beta_2 + b_{2i})\left(Y_{Bij} - \bar{Y}_{Bi}\right) + \epsilon_{ij} \tag{2.20}$$

where the assumption of the random effects are:

$$\begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_0^2 & \tau_{01} & \tau_{02} \\ \tau_{01} & \tau_1^2 & \tau_{12} \\ \tau_{02} & \tau_{12} & \tau_2^2 \end{bmatrix}\right) \tag{2.21}$$

In linear regression, the covariates can affect the variable of primary interest. For instance, the treatment may perform better on the more severe patients. To explore the relationship between the treatment effect and the baseline score in this case, we take the **interaction** into account. Hence, we can extend equation 2.20 to:

$$Y_{Fij} = \beta_{0i} + (\beta_1 + b_{1i})\, X_{ij} + \beta_{2i}\left(Y_{Bij} - \bar{Y}_{Bi}\right) + \beta_{3i}\left[\left(Y_{Bij} - \bar{Y}_{Bi}\right) X_{ij}\right] + \epsilon_{ij} \tag{2.22}$$

where, the β_3 denotes the increase of treatment effect for a one-unit increase of the baseline score in the treatment group. And b_{3i} is the random effect of the treatment-baseline interaction. $\left(\bar{Y}_{Bi}X_{ij}\right)$ is the interaction of the study-specific mean baseline values and the treatment group.

Similar with the equation 2.20, we can introduce random intercept and random slope to extend the model 2.19 as:

$$\begin{aligned} Y_{Fij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\, X_{ij} + (\beta_2 + b_{2i})\left(Y_{Bij} - \bar{Y}_{Bi}\right) \\ + (\beta_3 + b_{3i})\left[\left(Y_{Bij} - \bar{Y}_{Bi}\right) X_{ij}\right] + \beta_{4i}\left(\bar{Y}_{Bi}X_{ij}\right) + \epsilon_{ij} \end{aligned} \tag{2.23}$$

where β_{4i} denotes the increase of the treatment effect for a one-unit increase in the study-specific mean baseline score in the treatment group. There are many model specifications available. We can apply various combinations of covariates, fixed and random effects for the intercept and slopes, and whether the interaction is specified. In this thesis, we mainly consider model 2.18 and 2.19.

2.5.2 Two-stage Meta-Analysis With LMM

Another popular approach is the two-stage meta-analysis of the IPD. In the first-step, the ANCOVA model is fitted with respect to each study from 1 to n. Inherit the notation of the one-stage model, the model for study $i \in 1, 2, ..., n$ can be written as:

$$Y_{Fij} = \beta_{0i} + \beta_{1i}X_{ij} + \beta_{2i}Y_{Bij} + \epsilon_{ij} \tag{2.24}$$

The β_{0i} is the intercept of study i, β_{1i} and β_{2i} are the treatment effect and baseline effect in study i, respectively. The ϵ_{ij} is the residual.

The second-step is to synthesis the results of each study with the standard meta-analysis methods. The fixed effect model 2.2 and random effect model 2.3 can be used to estimate the summary treatment effect.

The one-stage and two-stage methods will acquire similar results if the model's assumptions are identical. For instance, if $\sigma_{ij} = \sigma_{i.}$ and no interaction is involved, equation 2.19 and the two-stage model 2.24 apply the same treatment variance structure and exclude interaction variables. Hence, in this case, we expect the estimated results from these models to be similar.

2.6 Pseudo-IPD Methods

The meta-analysis of IPD is flexible since we can apply LMM using various residual structures. In addition, Modeling options are much fewer in a standard meta-analysis of the AD. Therefore, Papadimitropoulou et al. [4] propose the **Pseudo-IPD** method to overcome these challenges. The idea of the method is based on the properties of **sufficient statistics**. In the likelihood linear mixed model, the mean, standard deviation, and correlation of the baseline and follow-up variables are sufficient statistics since they are sufficient to calculate all the estimated results. More specifically, if the original IPD is available, we can summarize the identical aggregate data from the pseudo-IPD as the original IPD. Hence, we can improve the flexibility of the analysis when there is only access to the AD. Based on the sufficient statistics property, the estimation of the pseudo-IPD should be identical with the estimation of the original IPD. We can use a simple algorithm to implement the IPD-generation process:

1. Generate two random samples from the standard normal distribution with the size equal to n, the size for each measure in each study. Denote them Y_i^*1 $(i = 1, ..., n)$ and Y_i^*2 $(i = 1, ..., n)$

2. Standardize the samples and calculate the correlation r^* of the two samples.

3. Fit the linear regression model by setting the Y_{i2}^* as the dependent variable and Y_{i1}^* as the independent variable. Denote the coefficients and residual as $\hat{\beta}$ and $\hat{\epsilon}$ respectively.

4. Create a new variable $Y_{i3}^* = Y_{i1}^* r + \hat{\epsilon}_i \sqrt{1 - r^2} \left[\sqrt{1 - r^{*2}}\right]^{-1}$, where r denotes the true correlation in the studies.

5. Generate $Y_{Bi} = Y_{i1}^* Sd_B + \bar{Y}_B$ and $Y_{Fi} = Y_{i1}^* Sd_F + \bar{Y}_F$. Y_{Bi} and Y_{Fi} are the pseudo-IPD we need.

The pseudo-IPD comprises the study, treatment group, baseline score, and follow-up score for each subject. We can apply the linear mixed model tools to estimate the treatment effect and specify various variance structures.

Chapter 3

Example: Meta-analysis of obstructive sleep apnea data

In this chapter, we will apply the standard meta-analysis methods and the pseudo-IPD methods to the aggregate data of obstructive sleep apnea. Obstructive Sleep Apnea is characterized by recurring episodes of cessation (apnea) or reduction (hypopnea) in airflow during sleep caused by obstruction of the upper airway [3]. In Canada, the moderate and severe prevalence of disease ranges from 3% to 50% depending on sex and age. The typical symptoms of OSA are daytime sleepiness, unrefreshing sleep or fatigue, poor concentration, etc. Researchers use the frequency of apnea-hypeapnea events per hour in the total sleep time (AHI) to measure the severity. A common standard is:

1. Mild OSA: AHI ≥ 5 and < 15 events per hour
2. Moderate OSA: AHI ≥ 15 and < 30 events per hour
3. Severe OSA: AHI ≥ 30 events per hour

A positive treatment for symptomatic patients is the continuous positive airway pressure (CPAP) [3] which can release sleepiness and reduce AHI. The main task of the meta-analysis is to analyze the treatment effect of a CPAP treatment device. In each study, one of the patient groups receives an active CPAP, and the control group patients are treated with a sham CPAP device. The methods model from chapter 2 will be applied to the data.

3.1 The Aggregate Obstructive Sleep Apnea Data

We perform the meta-analysis methods on the aggregate Obstructive Sleep Apnea data. The outcome measure is the apnea-hypopnea index(AHI) which indicates the ratio of the number of apnea or hypopnea events per hour in the total hours of sleep. In all studies, the AHI score was measured at the baseline (pre-treatment) and follow-up (post-treatment) time point for each patient. The mean, standard deviation, correlation, and sample size of the active CPAP group (Treatment group) and sham CPAP group (Control group) are provided in the aggregate data. The aim is to estimate the treatment effect of the CPAP device. The full dataset is comprised in table 3.1.

We first perform the standard meta-analysis models 2.8, 2.13, 2.16. Next, we apply the Trowman method, the formula 2.17 to the aggregate data. Then we generate pseudo-IPD data and fit model 2.18 and model 2.19 to the pseudo-IPD data. Finally, we apply the two-stage model (formula 2.24) to estimate the treatment effect. In this way, we reproduce analysis of the pseudo-IPD experiments in the paper of Papadimitropoulou's[4].

We apply the model fitting functions in the 'metafor' package in R for the standard meta-analysis. The random effect meta-analysis model on the final scores and change scores are fitted with the estimation method 'REML'. We use the forest plot to show the estimates of the treatment effect in each trial and use the 'summary' function to acquire the essential information of the estimation, and we extract the estimated treatment effect, estimated standard error, and confidence interval.

The idea of the Trowman method is straightforward. We apply linear regression on the aggregate data with the follow-up score as the dependent variable and the treatment group and baseline score

Table 3.1: The final Dataset used in Papadimitropoulou's meta-analysis

ID	Study	MB	sdB	MPB	sdPB	NFB	Cor	group
1	Egea	43.1	22.9	10.8	11.4	27	0.4979	1
2	Haensel	65.9	28.6	3.5	3.4	25	0.4981	1
3	Loredo99	56.4	24.1	3.3	3.8	23	0.4442	1
4	Mills	65	34	2.56	2.4	17	0.4969	1
5	Loredo06	65.9	28.6	3.0	4.7	22	0.5704	1
6	Norman	66.1	29.1	3.4	3.0	18	0.4967	1
7	Becker	62.5	17.8	3.4	3.1	16	0.5025	1
8	Spicuzza	55.3	11.9	2.1	0.3	15	0.5052	1
1	Egea	35.3	16.7	28.0	24.8	29	0.4979	0
2	Haensel	57.5	32.1	53.4	32.9	25	0.4981	0
3	Loredo99	44.2	25.3	28.3	22.7	18	0.4442	0
4	Mills	61.2	41.0	57.3	41.0	16	0.4969	0
5	Loredo06	57.5	32.1	52.5	37.5	19	0.5704	0
6	Norman	53.9	29.8	50.1	32.1	15	0.4967	0
7	Becker	65.0	26.7	33.4	29.2	16	0.5025	0
8	Spicuzza	59.2	17.3	57.0	8.6	10	0.5052	0

Abbreviations: NFB denotes the sample size of each group in each study;
MB and sdB denote the mean and standard deviation of Apnea scores among patients at the baseline;
The MPB and sdPB denote the mean and standard deviation of Follow-up measurement;
The number represents the apnea/hypopnea events happened per hour;
Cor denotes the correlation between the baseline and follow-up. Group 1 and group 0 denote the treatment group and control group respectively.

as independent variables, which is written as formula 2.17.

The final method is based on the pseudo Individual participant data(IPD). We generate the pseudo-IPD from the aggregate data such that the mean, standard deviation, correlation, and sample size are identical with the aggregate data. That is, we can summarise the same aggregate data as table 3.1 from the pseudo-IPD. Then we apply meta-analysis on the pseudo-IPD directly by fitting the linear mixed model with various variance structures with function 'lme' in package 'nlme' in R.

3.2 Results

3.2.1 Results for each methods

All the forest plots are provided in the Appendix. The estimation values of the Final Score Model are shown in figure A.1

The lines with black points at the center are the description of the estimated effect size with the 95% confidence interval. The size of the black points denotes the weight of the studies. The line at the bottom is the scale of all the measures in the plot. In the 'Study' column, the Studies' names are listed respectively.

On the right side, the first column is the exact weight of each study. The second column is the estimated treatment effect for each study. And the last column is composed of the confidence interval of the estimates. The last row is the summary effect of primary interest.

Similarly, the forest plot of the change score model and recovered ANCOVA model are shown in figure A.2 and A.3.

We report the estimated results of all the meta-analysis models in table 3.2.

Based on the standard methods, the pseudo-IPD methods, and the Trowman method, the estimations of the treatment effect are ranging from -40.43 to -45.52 events per hour, which are relatively stable. Therefore, the pseudo-IPD method can provide a reasonable estimation of the effect size. The range of the standard deviations of the methods is from 4.67 to 5.46. And the random effect variance is various from 152.6 to 190.6.

From the standard meta-analysis methods, the random effect model of the change scores acquires a smaller standard error and random effect than the final score model. Since we take the Baseline score

Table 3.2: The summary results of various meta-analysis models

Models	Estimate	SE	CI	Random Effect
Change Score Model (Formula 2.13)	-45.52	5.29	[-55.89, -35.16]	152.65
Final Score Model (Formula 2.8)	-40.43	5.46	[-51.13,-29.73]	190.57
Recovered ANCOVA Model (Formula 2.16)	-42.41	5.22	[-52.65, -32.18]	181.79
Trowman Model (Formula 2.17)	-41.74	4.76	[-51.06, -32.42]	NA
Simplest Pseudo-IPD Model (Formula 2.18)	-40.64	5.20	[-50.87,-30.40]	170.78
Full Pseudo-IPD Model (Formula 2.19)	-42.41	5.23	[-52.70,-32.12]	180.36
Two-stage Model (Formula 2.24)	-42.41	5.23	[-52.66,-32.16]	180.44

The negative estimated values mean that the treatment effect can **reduce** the times of breathing difficulty.

Table 3.3: The summary results of various residual structure

Residual Structure	Estimate	SE	CI	Random Effect
All-equal	-42.61	5.18	[-52.81, -32.40]	171.50
Study-specific	-42.41	5.23	[-52.70, 32.12]	180.36
Group-specific	-41.18	5.16	[-51.34, -31.02]	162.36
Study and Arm-specific	-41.07	5.25	[-51.40, -30.74]	176.15

into the change score model analysis and acquire a better performance, we conclude that the Baseline value has an impact on the treatment effect estimation.

The Trowman method is based on the linear model without random effect. And the Trowman model acquires the smallest standard error (4.76) among all methods.

The Recovered ANCOVA model, the Full Pseudo-IPD Model, and the two-stage pseudo-IPD model produce very similar estimates and performance. The relationship of the methods will be explored in the simulation study.

3.2.2 Results for different Residual structure

For the Full Pseudo-IPD Model, we explore further on the various residual structures. With fixed predictors, we vary the residuals as all-equal ($\sigma_{ik} = \sigma$ where i is the study and k denotes the groups), study-specific ($\sigma_{ik} = \sigma_{i.}$), group-specific ($\sigma_{ik} = \sigma_{.k}$), and arm-specific ($\sigma_{ik} = \sigma_{ik}$). The estimation results are shown in table 3.3.

For different residual structures, the estimates are range in $[41.07, 42.61]$. However, the estimated random effects vary significantly, ranging from 162.36 to 180.36. Compared with the results in table 3.2, the study-specific model acquired almost identical results with the recovered ANCOVA model and the two-stage model and the estimated results of the other residual structures are slightly different. We will use the simulation study to further assess the performance of residual structures.

3.2.3 Results for within-study interaction

An extension of the full pseudo-IPD model is to include interaction terms in the model. We fitted the model with a within-study interaction term (the model 2.22). The results are shown in table 3.4

The baseline score in the pseudo-IPD we used is centered to 0. We can interpret the results as:

Table 3.4: The results of the model with interaction

	Estimation	Standard Error	DF	t-value	p-value
Treatment	-42.63	5.14	293	-8.29	0.00
Within-study interaction	-0.40	0.07	293	-5.41	0.00
Baseline	0.53	0.15	293	3.75	0.00

1. The treatment effect for the patient with a centered baseline score 0 has an average treatment effect -42.63. That is, the active CPAP can reduce breathing difficulty for approximately average 42 times per hour. The active CPAP treatment is significantly better than the sham CPAP.

2. The interaction term is significant that the baseline effects in the treatment group and control group are different. For a one-unit increase in the baseline score, the mean increase of the treatment effect of the active CPAP will be 0.40. E.g., a patient with a centered baseline score 10 received the active CPAP treatment, his or her treatment effect can reduce approximately $-42.63 - 0.40 \times 10 = -46.63$ times of breathing difficulty per hour.

Chapter 4

Simulation Process

4.1 Notation

The notations used in this chapter are listed in table 4.1

4.2 Aim

To assess the performance of an estimation method, we mainly consider two measurement, the bias and variance. The bias is a systematic tendency to cause a difference between the true value of a parameter and the mean estimated value. An unbiased estimation satisfies that

$$\lim_{N \to +\infty} \frac{\sum_{i=1}^{N} \hat{\theta}_i}{N} - \theta = 0 \tag{4.1}$$

In addition, the variance of the estimation should be taken into account. Imagine an unbiased estimation where the result of each estimation varies widely. Such an estimation will be a disaster for researcher. Therefore, a biased estimation with much smaller variance is not entirely undesirable.

When we assess the performance of the methods of chapter 2 in chapter 3. We cannot measure the bias and error since the true treatment effect is unknown. Therefore, we desire a experiment with known treatment effect to assess the estimation models. In general, the methods with small MSE is preferred. The MSE is equal to the sum of bias and variance which can be regarded as a measure of the difference between the estimated values and true values. We can calculate the MSE by formula 4.2

$$MSE = \frac{\sum_{i=1}^{n_{sim}} (\theta - \hat{\theta}_i)^2}{n_{sim}} \tag{4.2}$$

Additionally, in the obstructive sleep apnea example, the baseline imbalance exists in each study which can lead to bias and Type 1 error if no reasonable adjustment is specified[7]. The baseline imbalance between treatment groups is common in small trials and can confound the inference of the treatment effect. In general, the ANCOVA model with baseline as a covariate accounts better for the baseline imbalance. However, which method in chapter 2 provides the best estimation under different conditions is unknown. To assess the bias and error of each method, we use a simulation study for further exploration.

4.3 Data-generation machine

The simulation study generated individual participant data for a meta-analysis with n_{study} studies, with an equal number of subjects within each arm of the study. Baseline values were drawn from a normal distribution with study and arm-specific mean, $\hat{\mu}_{Bik}$ for study i with $k = 1$ for the treatment group and $k = 2$ for the control group, and standard deviation of the baseline, σ_{YB}, was constant in all studies and all groups.

Table 4.1: Frequently used Notations

Symbol	Description
Change	The Change score model.
Final	The final score model
RA	The recovered ANCOVA model.
Trowman	The Trowman model
PB	The pseudo base model with the only predictor, the treatment group
PF	The pseudo full model is the one-stage method with the treatment effect, study-specific baseline effects and intercept.
PTS	The pseudo Two-stage model with baseline, treatment effect as predictors.
Treatment effect, β_1	The true treatment effect we set in the simulation.
Baseline effect, β_2	The true baseline treatment effect we set in the simulation.
Random effect, τ	The true standard deviation of the random treatment effect.
σ_{YB}	The true standard deviation of the values at baseline.
Residual, σ_{ik}	The true standard deviation of the experiment error in study i.
n_{study}	The total number of studies in the simulated dataset.
n_1 and n_2	The number of patients for the treatment group and control group.
imbalance	The quantity of imbalance between the mean baseline of the treatment group and control group.
$\hat{\beta}_1$	The mean of the estimated treatment effect value in all the simulation experiments.
standard error	The mean of the standard error of the estimated treatment effects in the output of the simulation experiments.
tau	The mean of the estimated random effect standard deviation in all the simulation experiments.
MSE	The mean square of error among all the simulation experiments.
Observed standard deviation of $\hat{\beta}_1$	The standard deviation of the estimated value in all simulation.
n_{sim}	The number of experiments in the simulation study. We generate $n_{sim} = 200$ data sets and apply all the models n_{sim} times and summary the results.

We kept the baseline outcome balance in the generated data ($\hat{\mu}_{Bi2} = \hat{\mu}_{Bi1}$) for all studies in most of the experiments. In addition, we also generated the data in one study where the baseline outcome was not balanced ($\hat{\mu}_{Bi2} = \hat{\mu}_{Bi1} + imbalance$) to check the performance.

After the baseline score was determined, we generated the follow-up values based on the baseline values using the stratified model of formula 4.3

$$Y_{Fij} = \beta_{0i} + (\beta_1 + b_{1i})X_{ij} + \beta_2 Y_{Bij} + \epsilon_{ij} \tag{4.3}$$

with $\beta_{0i} = 40$ in all studies and simulations. b_{1i} denotes the random effect following a normal distribution with mean 0 and standard deviation τ. ϵ_{ij} follows a normal distribution with mean 0 and standard deviation σ_{ik} (i and k represent the study and group respectively). We used the fixed mean treatment effect $\beta_1 = 50$ in all experiments. We did not vary the treatment effect since the performance of the models (MSE, standard error, and bias) would not react to the change of the treatment effect.

The parameters and designs were varied as follows:

1. Number of Studies $n_{study} = 4, 8, 16$
2. Number of patients in each arm $n_1 = n_2 = 10, 20, 30$
3. $\sigma_{ik} = 4, 8, 32$ for all i, k.
4. The study-specific, group-specific, all-equal, and arm-specific residual structures in the pseudo full model.
5. $\sigma_{YB} = 10, 20, 30$. We use the same values for all studies and all groups.
6. Imbalance $= 0, 5$. We compare the conditions with and without baseline imbalance.
7. $\beta_2 = 0.2, 0.5, 0.8$. We compare the estimation under different baseline effect.
8. $\tau = 6, 13, 20$. The MSE and bias under different random effect will be assessed.

For each parameter, we performed $n_{sim} = 200$ simulations. We regarded $imbalance = 0$, $N = 8$, $n_1 = n_2 = 20$, $\sigma_{ik} = 4$, $\sigma_{YB} = 20$, $\beta_1 = 50$, $\beta_2 = 0.5$, $\tau = 13$ as the standard case. In each simulation experiment we varied one of the parameters, and compare as standard values for the other parameters.

For each simulated dataset, we aggregated the data by study and arm and then performed the meta-analysis methods 2.8, 2.13, 2.16, 2.17, 2.18, 2.19, and 2.24, like what we did in the obstructive sleep apnea case. We referred them to the terms as following:

- **Final Score Model**: the model 2.8, the standard meta-analysis method with follow-up score as the outcome
- **Change Score Model**: the model 2.13, the standard meta-analysis method with change score as the outcome
- **Recovered ANCOVA Model**: the model 2.16, standard meta-analysis method with the adjusted outcome
- **Trowman Method**: the model 2.17, meta-regression with mean baseline effect, and treatment as predictors
- **Pseudo Base Model**: pseudo-IPD is generated and the model 2.18, linear mixed model with treatment group as the predictor is used
- **Pseudo Full Model**: pseudo-IPD is generated and the model 2.19, a linear mixed model with study-specific baseline effects and intercept, treatment group as the predictors is used. In most simulations, we used a study-specific residual error structure, except for the simulation experiment of Table 4.5.
- **Pseudo Two-stage Model**: Using the pseudo-IPD, a linear model was fitted in each study, using baseline values and treatment as covariates, and a synthesis standard meta-analysis was performed.

In every simulation, we recorded the estimated treatment value, standard error, and estimated random effect τ, and we calculated the bias $(\hat{\beta}_1 - \beta_1)$, where β_1 denotes the true treatment effect, and $\hat{\theta}$ denotes the mean estimation of all repetition.

Finally, We calculated over all simulation:

- Mean bias : $(\hat{\beta}_1 - \beta_1)$

- Mean estimated standard error : $(se(\beta_1))$
- Mean estimated random effect : $(\hat{\tau})$
- The MSE

The results of the simulation are summarised in Tables and shown by boxplots. The whole experiment was implemented by R (version 4.0.3). Linear model were fitted using the package nlme. We applied the control of maxiter = 1000 times to the linear mixed model to prevent convergence issues.

Standard meta-analysis were performed using the package 'metafor'. The flow graph of the simulation is given in figure 4.1

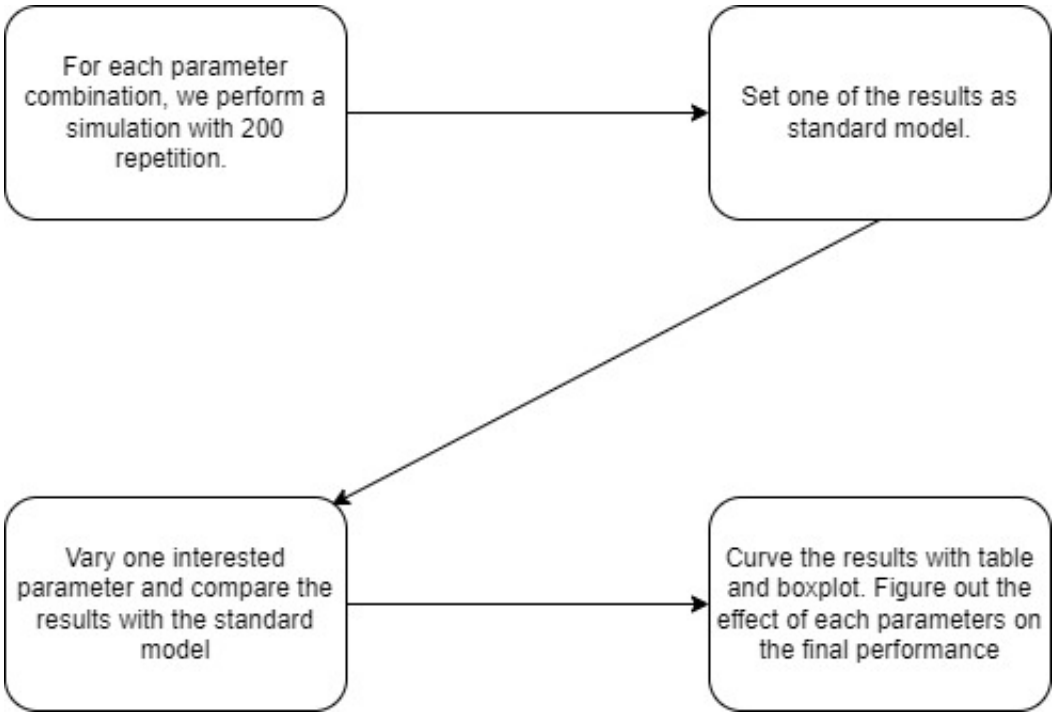

Figure 4.1: The logic of the simulation experiment

4.4 Results

In this section, we provide the results of the simulation study. We compared the performance for different parameters. First of all, we show the results of the standard case in table 4.2.

Regarding the bias, all the methods provide an unbiased estimation. However, a much higher MSE was detected in the Trowman method. Moreover, the estimated τ of the pseudo base model is smaller than the true value. In addition the pseudo full model, recovered ANCOVA model, and pseudo Two-stage model provided almost identical results, and acquired the lowest MSE. The Trowman model acquired the largest MSE. We found that the estimated random effects for all models on average were somewhat lower than the true value. The mean standard errors reported in the output were smaller than the observed standard deviation of the estimated treatment effect.

Figure B.1 is the boxplot with the distribution of the estimated treatment effect. The top line and bottom line of each box are the $Q1 = 25\%$ and $Q3 = 75\%$ quantiles, respectively. The black line in the middle is the median $Q2 = 50\%$. The range from 25% to 75% is termed as **Interquartile Range**, and the abbreviation is $IQR = Q3 - Q1$. The top short line and bottom short line are equal to $Q3 + 1.5 * IQR$ and $Q1 - 1.5 * IQR$.

To assess the performance of the methods under various conditions, we varied one parameter each time to compare with the standard case as we previously mentioned. We measured the stability and accuracy of the method by the bias and MSE. First, we checked the performance for various σ_{ik}.

Table 4.2: The results of standard case

Model	Estimation	Standard error	tau	MSE	Observed standard deviation of β_1	Bias
Change	49.79	4.54	12.34	23.32	4.84	0.21
Final	49.77	4.47	12.12	22.54	4.75	0.23
RA	49.79	4.38	12.33	21.77	4.67	0.21
Trowman	49.73	4.53	NA	24.42	4.95	0.27
PB	49.74	4.33	11.55	22.58	4.76	0.26
PF	49.79	4.38	12.32	21.77	4.67	0.21
PTS	49.79	4.38	12.32	21.77	4.67	0.21

The standard parameters are 1. $\beta_1 = 50$
2. $\beta_2 = 0.5$
3. $\sigma_{ik} = 4$
4. $\sigma_{YB} = 20$
5. $n_{study} = 8$
6. $n_1 = n_2 = 20$
7. $\tau = 13$
8. Without imbalance

4.4.1 Performance of different values for σ_{ik}

Table 4.3 and Figure B.2, Figure B.3 show the results of $\sigma_{ik} = 8$, and $\sigma_{ik} = 32$ with the same values in all studies, respectively. All methods provide unbiased estimates in both experiments. The standard error, observed standard deviation of β_1, and the MSE of the standard case (Table 4.2) are slightly lower than the case when $\sigma_{ik} = 8$. However, the error parts are much larger when $\sigma_{ik} = 32$.

For standard case ($\sigma_{ik} = 4$) and the $\sigma_{ik} = 8$ case, the Recovered ANCOVA model, pseudo full model, and pseudo two-stage model provide smaller MSE compared to other methods. Similar with the standard case, the Trowman model acquires the largest MSE.

For the case when $\sigma_{ik} = 32$, we observed similar pattern. In the real clinic trials, the residual should not reach such a large number. Here, the large value is used to assess the impact of the residual on the performance. Figure B.2 and B.3 show the overall distribution of the estimated values for $\sigma_{ik} = 8$, and $\sigma_{ik} = 32$, respectively. The box of $\sigma_{ik} = 32$ is larger than the standard case and $\sigma_{ik} = 8$, which indicates a more unstable estimate.

In clinical trials, we apply different treatments to different groups. It is common that the residual variance in different treatment groups is different. Table 4.4 and figure B.4 show the results of an experiment where the two groups have different residual standard deviations with $\sigma_{i1} = 8$ and $\sigma_{i2} = 4$. We observed that the MSEs in table 4.4 are slightly larger than the MSEs in Table 4.2, where $\sigma_{i1} = \sigma_{i2} = 4$, but lower than in Table 4.3 where $\sigma_{i1} = \sigma_{i2} = 8$. This was what we expected.

Similar to the Apnea example in chapter 3, we fitted pseudo IPD models with different residual structures. In this experiment, we used $\sigma_{i1} = 16$, and $\sigma_{i2} = 4$. Table 4.5 shows the results of different models. The results of both the estimations and the performance are similar for the different residual structures. This was different from what we observed in table 3.3 in the Apnea example, where we found some difference among the different models.

In addition, for each simulation, we regarded the dataset which is generated from formula 4.3 as the 'original data'. We applied the same linear mixed model to both the 'original data' and the pseudo-IPD. For all simulation, all models with different residual structures did provided identical results for both datasets. That is, the pseudo-IPD can recover all the information of the original data.

4.4.2 Performance of different values for n_{study}

Table 4.6, figure B.6, and figure B.5 show the results of $n_{study} = 4$ and $n_{study} = 16$. The $n_{study} = 16$ case provides a much lower reported standard error, observed standard deviation of β_1, and MSE than the standard case in table 4.2. And the $n_{study} = 4$ experiment provides exact opposite results. Similar to the standard case, the recovered ANCOVA model, pseudo full model, and the pseudo two-stage

Table 4.3: The results of different residual

Model	Estimation		Standard error		tau	
	$\sigma_{ik} = 8$	$\sigma_{ik} = 32$	$\sigma_{ik} = 8$	$\sigma_{ik} = 32$	$\sigma_{ik} = 8$	$\sigma_{ik} = 32$
Change	49.76	49.62	4.59	5.63	12.26	11.19
Final	49.73	49.56	4.52	5.62	12.04	11.07
RA	49.75	49.59	4.44	5.57	12.28	11.60
Trowman	49.70	49.53	4.58	5.71	NA	NA
PB	49.70	49.55	4.38	5.52	11.47	10.45
PF	49.75	49.60	4.44	5.54	12.27	11.21
PTS	49.75	49.60	4.44	5.52	12.27	11.35
	MSE		Observed standard deviation of β_1		Bias	
	$\sigma_{ik} = 8$	$\sigma_{ik} = 32$	$\sigma_{ik} = 8$	$\sigma_{ik} = 32$	$\sigma_{ik} = 8$	$\sigma_{ik} = 32$
Change	24.36	41.93	4.94	6.48	0.24	0.38
Final	23.31	38.50	4.83	6.20	0.27	0.44
RA	22.64	39.06	4.76	6.25	0.25	0.41
Trowman	25.55	43.70	5.06	6.61	0.30	0.47
PB	23.41	39.01	4.84	6.25	0.30	0.45
PF	22.63	39.63	4.76	6.30	0.25	0.40
PTS	22.63	39.40	4.76	6.28	0.25	0.40

The parameters are as follows 1. $\beta_1 = 50$
2. $\beta_2 = 0.5$
3. $\sigma_{ik} = 8$ and $\sigma_{ik} = 32$
4. $\sigma_{YB} = 20$
5. $n_{study} = 8$
6. $n_1 = n_2 = 20$
7. $\tau = 13$
8. Without imbalance

Table 4.4: The results of the study and arm-specific residual

Model	Estimation	Standard error	tau	MSE	Observed standard deviation	Bias
Change	49.80	4.56	12.28	23.79	4.89	0.20
Final	49.78	4.49	12.08	22.91	4.79	0.22
RA	49.79	4.41	12.29	22.16	4.71	0.21
Trowman	49.75	4.55	NA	24.93	5.00	0.25
PB	49.75	4.35	11.50	23.00	4.80	0.25
PF	49.80	4.41	12.28	22.16	4.72	0.20
PTS	49.80	4.41	12.28	22.16	4.72	0.20

In this case, we generate the data with group-specific residual setting. The parameters are as follows:
1. $\beta_1 = 50$
2. $\beta_2 = 0.5$
3. $\sigma_{i1} = 8$ and $\sigma_{i2} = 4$
4. $\sigma_{YB} = 20$
5. $n_{study} = 8$
6. $n_1 = n_2 = 20$
7. $\tau = 13$
8. Without imbalance

Table 4.5: The results of different residual structures

Residual Structure	Estimation	Standard Error	tau	MSE	Observed Standard Deviation	Bias
All-equal	49.81	4.52	12.17	23.53	4.86	0.19
Study-specific	49.82	4.51	12.16	23.65	4.87	0.18
Group-specific	49.82	4.50	12.13	23.60	4.87	0.18
Study and Arm-specific	49.83	4.50	12.13	23.74	4.88	0.17

In this experiment, we generate the data with $\sigma_{i1} = 16$, $\sigma_{i2} = 4$. And we fit the pseudo full model with 4 different residual structures like we did to the Apnea data in Chapter 3. The parameters are:
1. $\beta_1 = 50$
2. $\beta_2 = 0.5$
3. $\sigma_{i1} = 16$ and $\sigma_{i2} = 4$
4. $\sigma_{YB} = 20$
5. $n_{study} = 8$
6. $n_1 = n_2 = 20$
7. $\tau = 13$
8. Without imbalance

Table 4.6: The results of different n_{study}

Model	Estimation		Standard error		tau	
	$n_{study} = 4$	$n_{study} = 16$	$n_{study} = 4$	$n_{study} = 16$	$n_{study} = 4$	$n_{study} = 16$
Change	50.02	49.85	6.50	3.23	12.40	12.43
Final	50.14	49.86	6.35	3.21	12.06	12.35
RA	50.10	49.86	6.25	3.12	12.42	12.43
Trowman	49.85	49.84	6.74	3.17	NA	NA
PB	50.15	49.86	6.23	3.09	11.70	11.72
PF	50.10	49.86	6.25	3.12	12.42	12.43
PTS	50.10	49.86	6.25	3.12	12.42	12.43
	MSE		Observed standard deviation of β_1		Bias	
	$n_{study} = 4$	$n_{study} = 16$	$n_{study} = 4$	$n_{study} = 16$	$n_{study} = 4$	$n_{study} = 16$
Change	47.93	11.61	6.94	3.41	0.02	0.15
Final	50.71	11.61	7.14	3.41	0.14	0.14
RA	46.89	10.90	6.86	3.31	0.10	0.14
Trowman	56.34	10.84	7.52	3.30	0.15	0.16
PB	50.97	11.54	7.16	3.40	0.15	0.14
PF	46.89	10.90	6.86	3.31	0.10	0.14
PTS	46.89	10.90	6.86	3.31	0.10	0.14

The parameters are: 1. $\beta_1 = 50$
2. $\beta_2 = 0.5$
3. $\sigma_{i2} = 4$
4. $\sigma_{YB} = 20$
5. $n_{study} = 4, 16$
6. $n_1 = n_2 = 20$
7. $\tau = 13$
8. Without imbalance

model acquire almost identical results. In addition, their performance is more solid than the others regarding the MSE.

When only a small number of studies is available in the meta-analysis, the estimation will be unstable. Although all the model provide unbiased estimation, the standard deviation of the estimated β_1 is large. Hence, we prefer as much available studies as possible to guarantee the accuracy and stability of the meta-analysis.

4.4.3 Performance of different values for σ_{YB}

Table 4.7, figure B.7, and figure B.8 show the results of the cases when $\sigma_{YB} = 10$ and $\sigma_{YB} = 30$. For the Trowman model, recovered ANCOVA model, pseudo full model, and the pseudo two-stage model, the results are identical among $\sigma_{YB} = 10$, the standard case, and $\sigma_{YB} = 30$. That is, the impact of different σ_{YB} is adjusted perfectly in these models. In clinical trials, the treatment effect is of our primary interest. The standard deviation of the baseline score, and the imbalance between the baseline score are the noise that should be removed. These methods can successfully achieve the goal. For the other methods, the estimated results are slightly different for different σ_{YB}'s, which means the change score model, final score model, and the pseudo base model are unstable for various standard deviations of the baseline.

Table 4.7: The results of different σ_{YB}

Model	Estimation		Standard Error		tau	
	$\sigma_{YB} = 10$	$\sigma_{YB} = 30$	$\sigma_{YB} = 10$	$\sigma_{YB} = 30$	$\sigma_{YB} = 10$	$\sigma_{YB} = 30$
Change	49.79	49.78	4.43	4.71	12.35	12.27
Final	49.78	49.76	4.39	4.60	12.24	11.91
RA	49.79	49.79	4.38	4.38	12.33	12.33
Trowman	49.73	49.73	4.53	4.53	NA	NA
PB	49.75	49.73	4.24	4.47	11.67	11.33
PF	49.79	49.79	4.38	4.38	12.32	12.32
PTS	49.79	49.79	4.38	4.38	12.32	12.32
	MSE		Observed standard deviation of β_1		Bias	
	$\sigma_{YB} = 10$	$\sigma_{YB} = 30$	$\sigma_{YB} = 10$	$\sigma_{YB} = 30$	$\sigma_{YB} = 10$	$\sigma_{YB} = 30$
Change	22.29	24.89	4.73	5.00	0.21	0.21
Final	21.89	23.73	4.69	4.88	0.22	0.22
RA	21.77	21.77	4.67	4.67	0.21	0.21
Trowman	24.42	24.42	4.95	4.95	0.27	0.27
PB	21.93	23.79	4.69	4.88	0.25	0.27
PF	21.77	21.77	4.67	4.67	0.21	0.21
PTS	21.77	21.77	4.67	4.67	0.21	0.21

The parameters are: 1. $\beta_1 = 50$
2. $\beta_2 = 0.5$
3. $\sigma_{i2} = 4$
4. $\sigma_{YB} = 10, 30$
5. $n_{study} = 8$
6. $n_1 = n_2 = 20$
7. $\tau = 13$
8. Without imbalance

Table 4.8: The results of different n_{group}

Model	Estimation		Standard error		tau	
	$n_{group} = 10$	$n_{group} = 30$	$n_{group} = 10$	$n_{group} = 30$	$n_{group} = 10$	$n_{group} = 30$
Change	50.24	49.95	4.63	4.62	12.12	12.74
Final	50.38	50.04	4.69	4.63	12.30	12.78
RA	50.30	49.99	4.40	4.54	12.32	12.79
Trowman	50.23	49.89	4.56	4.72	NA	NA
PB	50.36	50.05	4.39	4.54	11.01	12.41
PF	50.30	49.99	4.40	4.54	12.31	12.79
PTS	50.30	49.99	4.40	4.54	12.31	12.79
	MSE		Observed standard deviation of β_1		Bias	
	$n_{group} = 10$	$n_{group} = 30$	$n_{group} = 10$	$n_{group} = 30$	$n_{group} = 10$	$n_{group} = 30$
Change	24.01	19.02	4.91	4.37	0.24	0.05
Final	23.34	20.00	4.83	4.48	0.38	0.04
RA	20.91	18.66	4.57	4.33	0.30	0.01
Trowman	21.19	22.91	4.61	4.80	0.23	0.11
PB	23.89	19.89	4.89	4.47	0.36	0.05
PF	20.91	18.66	4.57	4.33	0.30	0.01
PTS	20.91	18.66	4.57	4.33	0.30	0.01

The parameters are: 1. $\beta_1 = 50$
2. $\beta_2 = 0.5$
3. $\sigma_{i2} = 4$
4. $\sigma_{YB} = 20$
5. $n_{study} = 8$
6. $n_1 = n_2 = 10, 30$
7. $\tau = 13$
8. Without imbalance

4.4.4 Performance of different values for n_{group}

Table 4.8, figure B.9, and figure B.10 show the results for $n_{group} = 10$ and $n_{group} = 30$. Compare the results of table 4.8 and table 4.2, we can conclude that the larger group size, the smaller the bias is. We also see that the reported standard errors are very similar, but that the standard deviation of β_1 is decreasing somewhat smaller. However, no significant relationship between the standard error and the n_{group} values. In addition, the estimated variance of the random effect approach the true value if n_{group} increases, and the bias of the estimated treatment effect decreases when the n_{group} increases.

4.4.5 Performance of different values for β_2

Table 4.9 and figure B.11, B.13 show the results of the cases where the baseline effect, $\beta_2 = 0.2$ and $\beta_2 = 0.8$, respectively. If $\beta_2 = 0$, model 4.3 becomes $Y_{Fij} = \beta_{0i} + (\beta_1 + b_{1i})X_{ij} + \epsilon_{ij}$. This is a final score model , with $\theta_i = \beta_1 + b_{1i}$. In the same way if $\beta_2 = 1$, model 4.3 becomes $(Y_{Fij} - Y_{Bij}) = \beta_{0i} + (\beta_1 + b_{1i})X_{ij} + \epsilon_{ij}$. This is a change score model, with $\theta_i = \beta_1 + b_{1i}$. Hence, for small baseline effect, e.g., $\beta = 0.2$, the final score model performs better. Conversely, change score model is better for the large baseline effect, e.g., $\beta = 0.8$. That is, we can choose the standard meta-analysis method based on the developmental features of the disease. If the disease severity grows more rapidly for the

Table 4.9: The results of different β_2

Model	Estimation		Standard error		tau	
	$\beta_2 = 0.2$	$\beta_2 = 0.8$	$\beta_2 = 0.2$	$\beta_2 = 0.8$	$\beta_2 = 0.2$	$\beta_2 = 0.8$
Change	49.93	50.14	4.94	4.41	12.84	12.32
Final	49.99	50.17	4.63	4.64	12.95	11.88
RA	49.98	50.15	4.61	4.38	12.98	12.31
Trowman	49.92	50.13	4.85	4.52	NA	NA
PB	50.02	50.18	4.45	4.50	12.30	11.29
PF	49.98	50.15	4.61	4.38	12.97	12.31
PTS	49.98	50.15	4.61	4.38	12.97	12.31
	MSE		Observed standard deviation of β_1		Bias	
	$\beta_2 = 0.2$	$\beta_2 = 0.8$	$\beta_2 = 0.2$	$\beta_2 = 0.8$	$\beta_2 = 0.2$	$\beta_2 = 0.8$
Change	28.22	24.44	5.33	4.95	0.07	0.14
Final	24.17	29.98	4.93	5.49	0.01	0.17
RA	23.91	24.87	4.90	5.00	0.02	0.15
Trowman	26.18	27.13	5.13	5.22	0.08	0.13
PB	24.31	29.82	4.94	5.47	0.02	0.18
PF	23.91	24.87	4.90	5.00	0.02	0.15
PTS	23.91	24.87	4.90	5.00	0.02	0.15

The parameters are: 1. $\beta_1 = 50$
2. $\beta_2 = 0.2, 0.8$
3. $\sigma_{i2} = 4$
4. $\sigma_{YB} = 20$
5. $n_{study} = 8$
6. $n_1 = n_2 = 20$
7. $\tau = 13$
8. Without imbalance

patients with worse base, we can choose the change score model to estimate the treatment effect. For the disease with approximate equal effect for the patients with different baseline values, the final score model can provide a solid estimation.

For any value of β_2, all the models provide an unbiased estimation since the assumption that there is no baseline imbalance between groups is always satisfied.

4.4.6 Performance when baseline imbalance exists

Table 4.10 and figure B.14, B.15, B.16 show the performance of different baseline effect β_2 when baseline imbalance exists. Unlike the experiments before, the change score model, final score model, and the pseudo base model provide biased estimation, which means that the imbalance between the baseline is the source of the bias. The recovered ANCOVA method, pseudo full model, and pseudo two-stage model succeed in adjusting for the imbalance and recovering the true treatment effect.

In addition, the bias of the change score model decreases for the increasing of the β_2 ($\beta_2 = 0.2$, bias = 4.14; $\beta_2 = 0.8$, bias = 1.45). For the final score model and the pseudo base model, the bias grows for the increasing β_2. ($\beta_2 = 0.2$, bias = 1.07; $\beta_2 = 0.8$, bias = 3.61)

For the Trowman method, the standard error and the MSE are much larger than the experiment without imbalance, which indicates that the Trowman method performs unstable for the case with

baseline imbalance.

One thing to note is that the difference between the estimated effect from the change score model and the estimated effect from the final score model is approximately equal to the baseline imbalance. For example, when we set the baseline imbalance to 5, the mean difference between the estimated values of the change score model and the final score model will be approximately 5.

4.4.7 Performance of different values for random effect τ

Table 4.11 and figure B.17, B.18 show the results of the simulation experiment with $\tau = 6$ and $\tau = 20$. A large random effect τ indicates that the treatment effect has a large difference among studies. Hence, it is natural to acquire results with large variations in the simulation study. For the $\tau = 20$ case, we only use 8 studies, and the minimum MSE reaches 51.12 (The MSE of all the adjusted methods). For the $\tau = 6$ case, the maximum MSE is 6.25, which is much smaller than the standard case and the $\tau = 20$ case. For the estimated random effect variance, the estimated τ was smaller than the real values. In addition, the bias was larger for the larger random effect.

Table 4.10: The results when baseline imbalance exists

Model	Estimation			Standard error		
	$\beta_2 = 0.2$	$\beta_2 = 0.5$	$\beta_2 = 0.8$	$\beta_2 = 0.2$	$\beta_2 = 0.5$	$\beta_2 = 0.8$
Change	54.14	52.37	51.45	4.64	4.46	4.56
Final	48.93	47.26	46.39	4.42	4.47	4.77
RA	49.97	49.83	50.44	4.41	4.39	4.55
Trowman	50.17	49.53	50.25	5.72	5.70	5.80
PB	49.03	47.20	46.18	4.45	4.32	4.50
PF	49.97	49.83	50.44	4.41	4.39	4.55
PTS	49.97	49.83	50.44	4.41	4.39	4.55
	tau			MSE		
	$\beta_2 = 0.2$	$\beta_2 = 0.5$	$\beta_2 = 0.8$	$\beta_2 = 0.2$	$\beta_2 = 0.5$	$\beta_2 = 0.8$
Change	12.32	12.28	12.82	36.41	24.92	25.12
Final	12.41	12.31	12.75	20.27	29.28	37.33
RA	12.42	12.36	12.84	18.62	19.96	23.01
Trowman	NA	NA	NA	38.28	34.83	43.80
PB	12.30	11.54	11.29	25.26	31.70	44.40
PF	12.42	12.36	12.84	18.62	19.96	23.01
PTS	12.42	12.36	12.84	18.62	19.96	23.01
	Observed standard deviation of β_1			Bias		
	$\beta_2 = 0.2$	$\beta_2 = 0.5$	$\beta_2 = 0.8$	$\beta_2 = 0.2$	$\beta_2 = 0.5$	$\beta_2 = 0.8$
Change	4.40	4.40	4.81	4.14	2.37	1.45
Final	4.38	4.68	4.94	1.07	2.74	3.61
RA	4.33	4.48	4.79	0.03	0.17	0.44
Trowman	6.20	5.90	6.63	0.17	0.47	0.25
PB	4.94	4.90	5.47	0.97	2.80	3.82
PF	4.33	4.47	4.79	0.03	0.17	0.44
PTS	4.33	4.47	4.79	0.03	0.17	0.44

The parameters are: 1. $\beta_1 = 50$
2. $\beta_2 = 0.2, 0.5, 0.8$
3. $\sigma_{i2} = 4$
4. $\sigma_{YB} = 20$
5. $n_{study} = 8$
6. $n_1 = n_2 = 20$
7. $\tau = 13$
8. $imbalance = 5$

Table 4.11: The results of different random effects

Model	Estimation		Standard error		tau	
	$\tau = 6$	$\tau = 20$	$\tau = 6$	$\tau = 20$	$\tau = 6$	$\tau = 20$
Change	49.88	49.69	2.33	6.85	5.52	19.04
Final	49.86	49.68	2.29	6.77	5.30	18.81
RA	49.88	49.69	2.05	6.73	5.66	18.99
Trowman	49.86	49.60	2.12	6.95	NA	NA
PB	49.85	49.63	2.34	6.54	5.03	17.92
PF	49.88	49.69	2.05	6.73	5.65	18.98
PTS	49.88	49.69	2.05	6.73	5.65	18.98
	MSE		Observed standard deviation of β_1		Bias	
	$\tau = 6$	$\tau = 20$	$\tau = 6$	$\tau = 20$	$\tau = 6$	$\tau = 20$
Change	6.25	52.82	2.50	7.28	0.12	0.31
Final	5.74	51.74	2.40	7.20	0.14	0.32
RA	4.86	51.12	2.21	7.16	0.12	0.31
Trowman	5.49	57.17	2.35	7.57	0.14	0.40
PB	5.75	51.83	2.40	7.21	0.15	0.37
PF	4.86	51.12	2.21	7.16	0.12	0.31
PTS	4.86	51.12	2.21	7.16	0.12	0.31

The parameters are: 1. $\beta_1 = 50$
2. $\beta_2 = 0.5$
3. $\sigma_{i2} = 4$
4. $\sigma_{YB} = 20$
5. $n_{study} = 8$
6. $n_1 = n_2 = 20$
7. $\tau = 6, 13, 20$
8. Without imbalance

Chapter 5

Conclusion and Discussion

5.1 Discussion of the findings

In the thesis, we considered the standard meta-analysis methods, the Trowman method, and the pseudo-IPD methods for meta-analysis with continuous outcomes measured at two time points. We first fitted the different models on the apnea data. The results from the methods were somewhat different in estimated treatment effect and estimated random effects yet similar in the standard errors. The methods which use only the final score, such as the Final score model (Formula 2.8) and the Simplest pseudo-IPD model (Formula 2.18) which has the treatment group as the only predictor are similar. The adjusted models, like Recovered ANCOVA model, Full Pseudo-IPD Model, and the Two-stage model also yielded similar results. The estimated treatment effect of these models were in between the results of the Change Score model and the Final Score model results.

In the simulation studies, we designed various scenarios to test bias and MSE. Considering each experiment independently, the MSE of the Trowman method was larger than the others, which indicates the instability of this meta-regression method. Apart from that, the estimated variance of the random effects were always smaller than the true values, especially in the pseudo base model. Having more studies, more samples in each arm, and smaller variations (residual variances or tau) did generally increase the performance of all the methods according to the bias of the estimated values and estimated random effects, and the standard errors.

Increasing the standard deviation of the baseline measurement did not change any of the results. The recovered ANCOVA model, the pseudo full model, pseudo two-stage model, acquired **identical** results for all standard deviations of the baseline measurement. That is, these methods completely adjust the difference of the baseline variations. However, the change score model, final score model, and the pseudo base model performed worse as the baseline standard deviation increased.

Additionally, We found that the performance of all the methods are sensitive to the baseline effect (β_2) and the baseline imbalance. For the experiment without baseline imbalance, all the methods provide the unbiased estimation. When baseline imbalance exists, the bias merges only in the change score model, final score model, and the pseudo base model. And the difference between the estimated values of the final score model and the change score model is approximately equal to the mean baseline imbalance. We also found that the performance of the final score model will be better if the baseline score has a small impact on the follow-up score. Similarly, the change score model will be closer to the true conditions if the baseline effect approaches to 1 (s,one follow-up score increase associated with one-unit increase of the baseline score). In addition, the Trowman method acquired an unbiased estimation yet with a much larger standard error in the experiment with imbalance.

Comparing the results of various baseline effect with and without imbalance experiments, the performance of the adjusted methods had no significant difference for all baseline effects. That is, the recovered ANCOVA model, pseudo full model, and pseudo two-stage model all can eliminate the impact of the baseline to the estimation.

5.2 Advantages of the way we performed the simulation

The Simulation process in this thesis has advantages in data-generation machine, performance measure, and model fitting. We will discuss as following.

1. **Data-generation Machine.** In this thesis, we assess the performance of the pseudo-IPD models. The pseudo-IPD is generated from the aggregate data. And the aggregate data is the summary of the original IPD which is not available for most of clinical trials. Based on the relationship between the datasets. There are two ways to perform the simulation. The first approach is to generate the aggregate data directly and then fitting all the models as we did in Chapter 3. The second method is to generate the original IPD and aggregate the IPD to acquire the AD. Then fitting the models as in the first approach. Comparing the two potential methods, we select the second method since that

   - We can compare the results of the pseudo-IPD and the original IPD to check whether the linear mixed model can acquire identical results on these two datasets.
   - The means and standard deviations of the change score can be calculated directly from the original IPD, which is more convenient than calculating from aggregate data.

   In addition, the data-generation formula 4.3 provides flexibility to vary each parameters separately, which allows us to analyze how each parameter affects the performance of the models.

2. **Performance Measure.** We use the MSE and bias to assess the performance of each method. Moreover, we calculated the standard deviation of the estimated values and compared it with the standard error in the output. Combining the results with the boxplots can provide a clear view of the performance.

3. **Model Fitting** In the model fitting process, we applied identical fitting control to the standard meta-analysis models and linear regression models to avoid convergence error. The control of the iteration times and convergence tolerance can reduce the time-cost of the simulation, which is a crucial challenge for many simulation studies.

5.3 Limitations of the Simulation

Although the simulation can generate suitable dataset to check the performance of the models, the flexibility and reliability have some limitations.

- In the generated data, there are many negative values. In some clinical trials (e.g. the Obstructive Sleep Apnea Data we used in Chapter 3), negative values will not appear. The negative values will not affect the methodology study. However, if we want to generate data which is very similar to a specific trials where all the values are positive, we have to add restriction to the simulation process.

- In the real trials, studies with larger scores also had larger standard deviations. In our simulation, the relationship between the mean scores and standard deviations is not taken into account.

- We only tried limited number of studies and number of patients in each arm. That is, we did not check the asymptotic properties of the methods. For instance, we did not use a extremely large number of patients in each arm (E.g. 1000) to the simulation. Hence, we do not know whether the negative bias of the estimated random effects exists when $n_{group} \to +\infty$

- In the generation process, we only tried the all-equal residual and the group-specific residual. The study-specific and arm-specific structures worth further exploration.

- We average the results of each model based on 200 simulated dataset. However, 200 times repetition may not enough. We can set the repetition time to e.g. 500 times to acquire a more reliable assessment.

For improvement, we can increase the number of repetition and assess the performance of the models on more circumstance. E.g. We can adjust the data-generation machine so that a larger mean corresponds to a larger standard deviation.

5.4 Comments of the methods

From the results of the simulation, the **pseudo full model** which has the study-specific baseline score and intercept, treatment group as the predictors can provide unbiased, solid estimation. For any parameters, the pseudo full model can estimate the treatment effect without bias and with a relatively small standard error.

The **pseudo two-stage model** with the baseline score, treatment group as predictors can acquire similar results as the pseudo full model. Pseudo two-stage model uses both linear model and standard meta-analysis, which is flexible to satisfy different assumptions.

For the standard meta-analysis models, the **Recovered ANCOVA model** can acquire stable and unbiased estimation in any conditions. The estimation result is similar with the pseudo full model and the pseudo two-stage model.

The **Trowman model** is a meta-regression model. It uses the aggregate data to fit a linear model directly. In the aggregate data, there are 2 treatment groups in each study and totally $2 \times n_{study}$ observations in the data. Since the number of observations is small, the standard error of the estimation is large. For the data with baseline imbalance, the performance of Trowman method is much worse than the other methods which adjust for the baseline imbalance.

The **Final Score Model** and the **Change Score Model** are the commonly used standard meta-analysis models since statistician can program without complex adjustment to the aggregate data. However, the performance of the methods are highly depends on the baseline effect. And most of time, the baseline effect is unknown before the clinic trials. Moreover, the baseline imbalance between treatment groups can cause bias to the methods.

5.5 Conclusion

In this thesis, we studied the performance of the standard meta-analysis methods, Trowman method, and pseudo-IPD methods. We checked the performance in scenarios with various residual variances, number of study, sample sizes in each group, standard deviations of the baseline scores, baseline effects, random effects, and the existence of baseline imbalance. Following findings are concluded from the simulation study.

- The **Pseudo Full Model**, **Pseudo Two-stage Model**, and the **Recovered ANCOVA Model** are stable unbiased methods. For all conditions, they provide almost equal results and the best performance among all models.
- The estimated variance of the random effects have negative bias compared to the true values. The larger sample sizes in each arm can reduce the bias.
- Without baseline imbalance, all the methods provide unbiased estimation. That is, the baseline imbalance is the only source of bias.
- All pseudo-IPD models can acquire identical results on the original data and the pseudo-IPD. Hence, the pseudo-IPD recovered all the information of the original IPD.

Based on the findings in the simulation, we recommend the improvements as following.

- For the future meta-analysis with continuous outcomes and a baseline and follow-up measurement, we recommend to use the **Pseudo Full Model**, **Pseudo Two-stage Model**, and the **Recovered ANCOVA Model** based on the aggregate data.
- Since the standard deviation of the adjusted score in the Recover ANCOVA model and change score in the Change Score Model are calculated from the correlation value of the baseline and follow-up, researchers can improve the quality of aggregate data, by reporting the correlation between baseline and follow up measurement in each group, or report the results of an adjusted ANCOVA analysis in the aggregate data.

5.6 Future Work

In this thesis, we assessed the pseudo base model with only treatment groups as predictor, pseudo full model with study-specific baseline effect and intercepts, treatment groups as predictors, and pseudo two-stage model. However, as we mentioned in Chapter 2, there are many other options for the pseudo full model. They should be studies and extended further. For the standard meta-analysis model, we only assessed the performance of the meta-analysis with raw mean difference measure. Some other effect sizes for the continuous outcome (e.g. the standard mean difference) worth exploration as well.

For both the Obstructive Apnea data and the simulated data in our study, only the baseline scores and treatment groups variables are included. For more complicated clinical trials, more variables like the sex, ages, smoking frequency of the patients may be provided. If the mean score, standard deviation of other continuous variables, and their correlation with the baseline and follow-up are provided in the aggregate data, how to adjust the effect of the other covariates is a potential topic for future study.

Bibliography

[1] Danielle L Burke, Joie Ensor, and Richard D Riley. "Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ". eng. In: *Statistics in medicine* 36.5 (2017), pp. 855–875. ISSN: 0277-6715.

[2] Julian P. T Higgins et al. "Meta-analysis of continuous outcome data from individual patients". eng. In: *Statistics in medicine* 20.15 (2001), pp. 2219–2241. ISSN: 0277-6715.

[3] Cheryl R Laratta et al. "Diagnosis and treatment of obstructive sleep apnea in adults". eng. In: *Canadian Medical Association journal (CMAJ)* 189.48 (2017), E1481–E1488. ISSN: 0820-3946.

[4] Katerina Papadimitropoulou et al. "Meta-analysis of continuous outcomes: Using pseudo IPD created from aggregate data to adjust for baseline imbalance and assess treatment-by-baseline modification". In: *Research Synthesis Methods* 11.6 (2020), pp. 780–794. DOI: https://doi.org/10.1002/jrsm.1434. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1434. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1434.

[5] Richard D Riley, Paul C Lambert, and Ghada Abo-Zaid. "Meta-analysis of individual participant data: rationale, conduct, and reporting". eng. In: *BMJ* 340.7745 (2010), pp. 907–525. ISSN: 0959-8138.

[6] Rebecca Trowman et al. "The impact of trial baseline imbalances should be considered in systematic reviews: a methodological case study". eng. In: *Journal of clinical epidemiology* 60.12 (2007), pp. 1229–1233. ISSN: 0895-4356.

[7] Lynn Wei and Ji Zhang. "Analysis of Data with Imbalance in the Baseline Outcome Variable for Randomized Clinical Trials". eng. In: *Therapeutic innovation regulatory science* 35.4 (2001), pp. 1201–1214. ISSN: 2168-4790.

Appendix A

Results of Obstructive Sleep Apnea Case

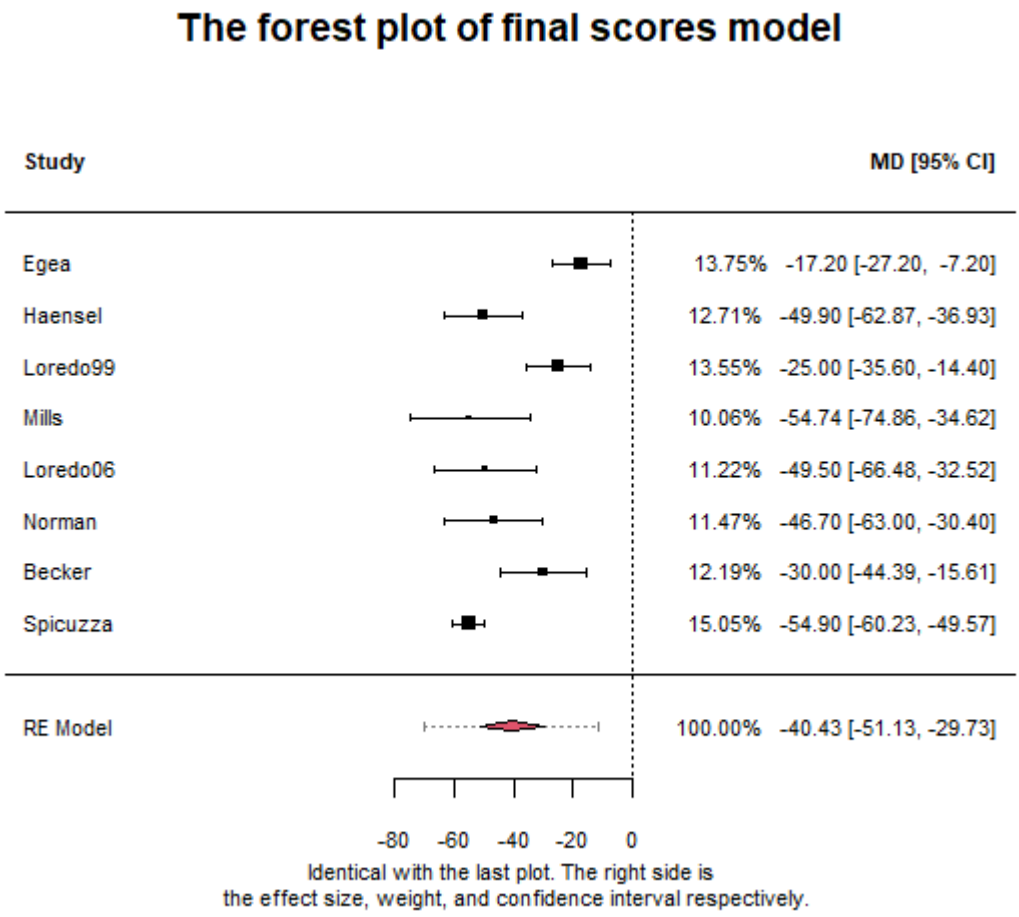

Figure A.1: Final Score Model Curve

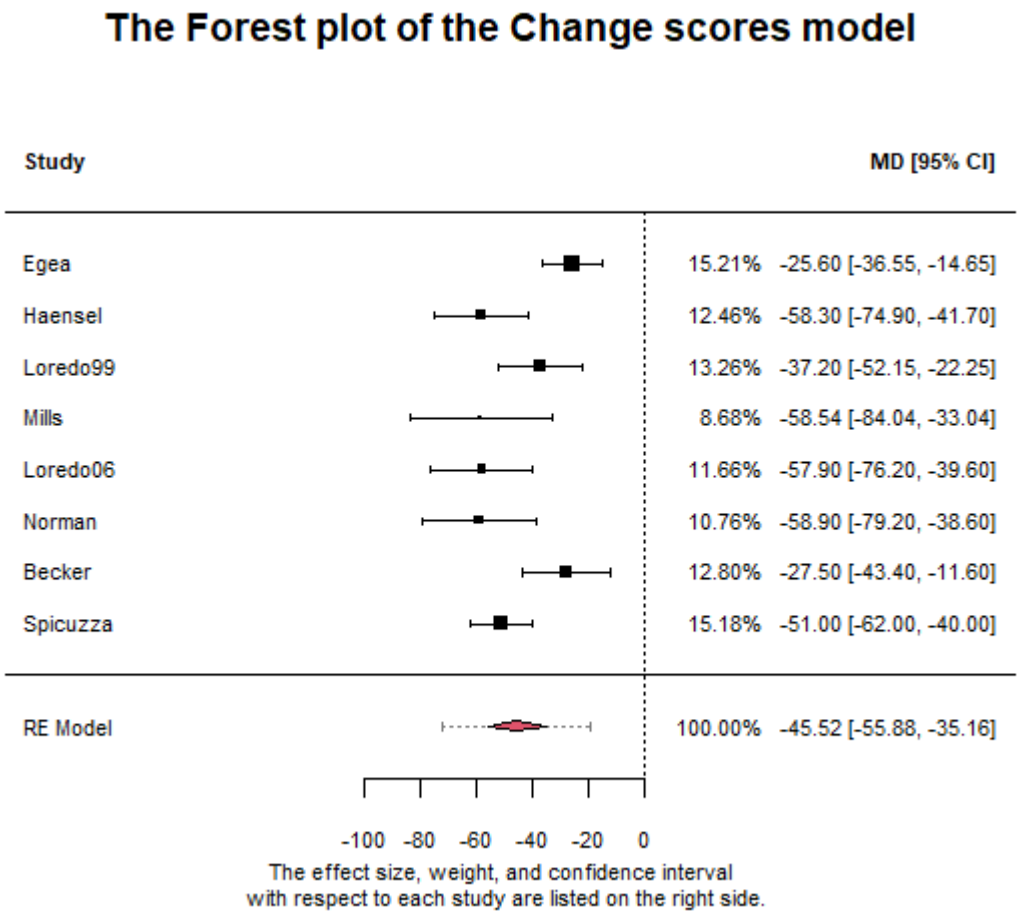

Figure A.2: Change Score Model Curve

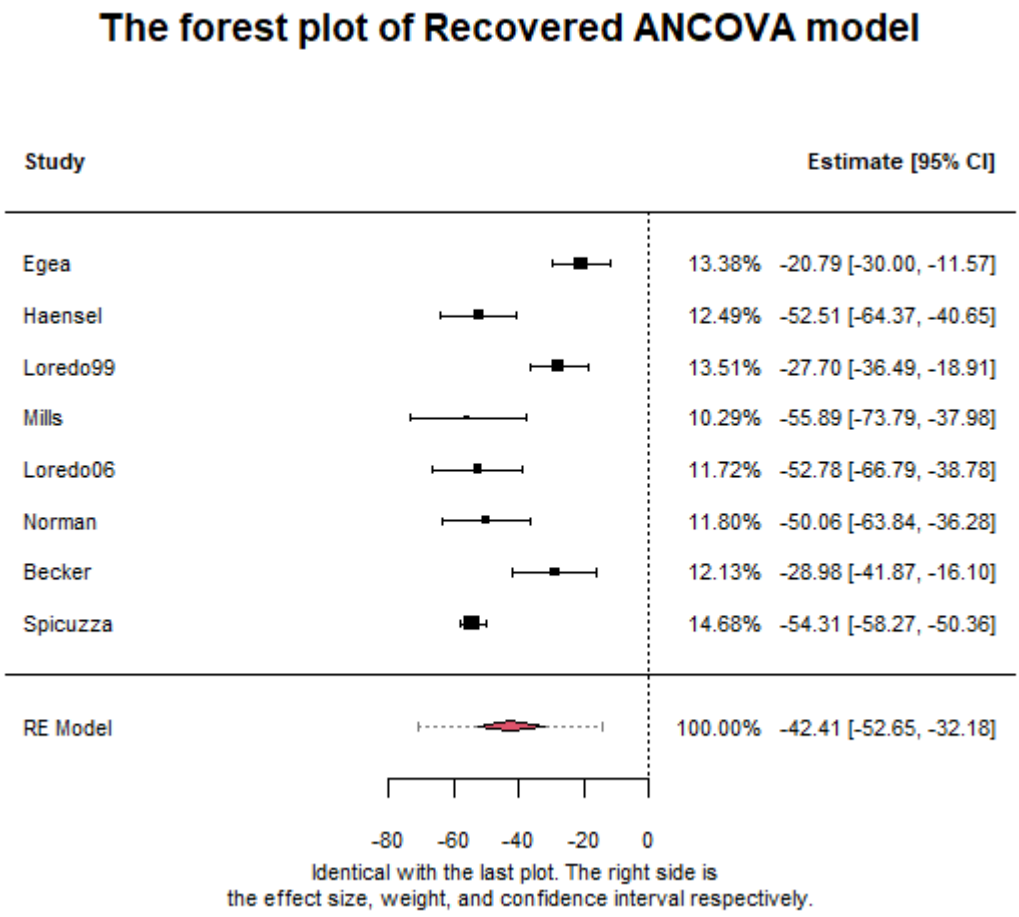

Figure A.3: Recovered ANCOVA Model Curve

Appendix B

Results of Simulation Study

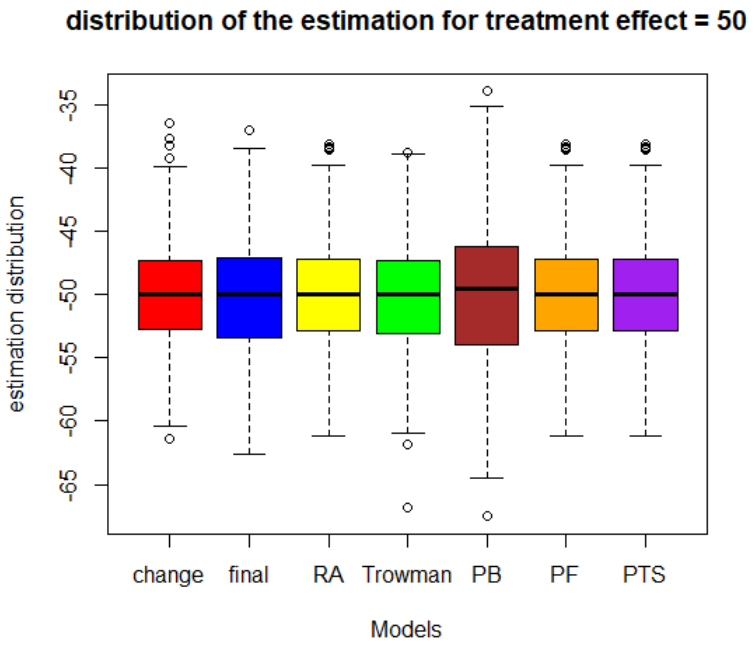

Figure B.1: The result of the standard case

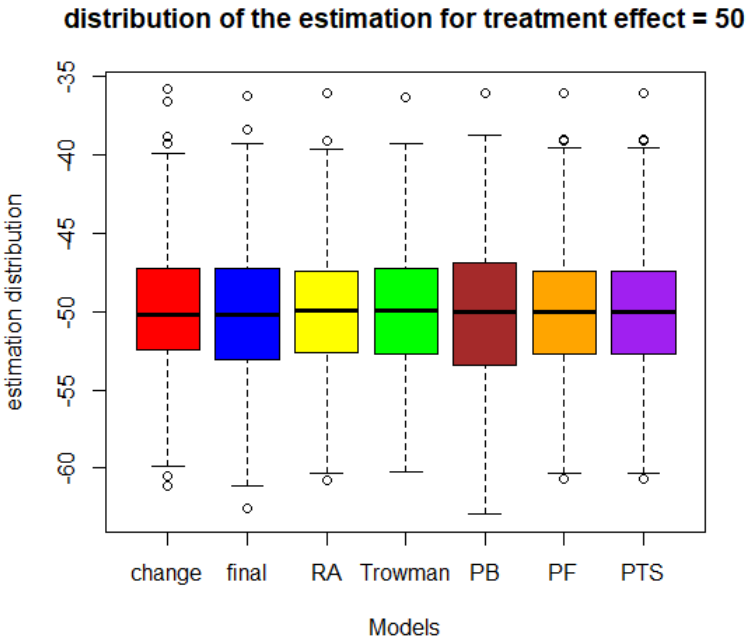

Figure B.2: The result of the case when $\sigma_{ik} = 8$

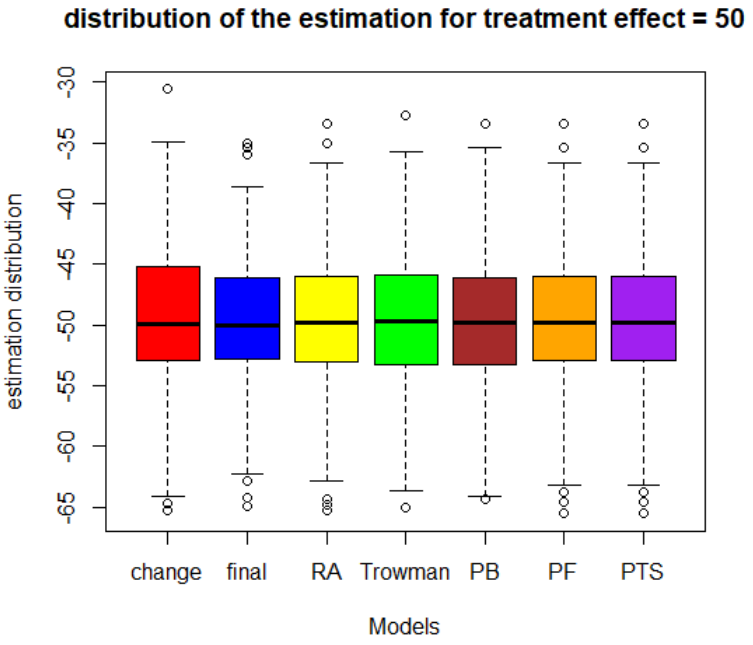

Figure B.3: The result of the case when $\sigma_{ik} = 32$

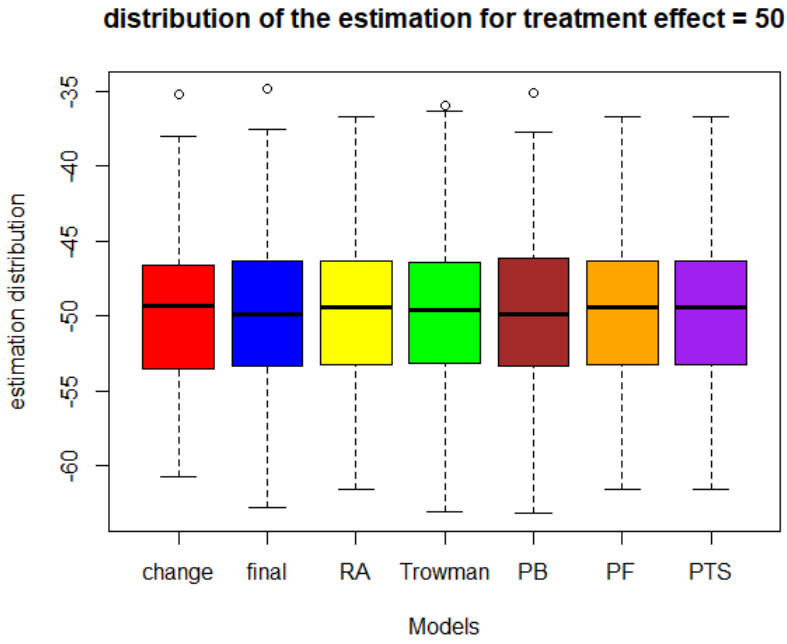

Figure B.4: The result of the arm-specific residual case

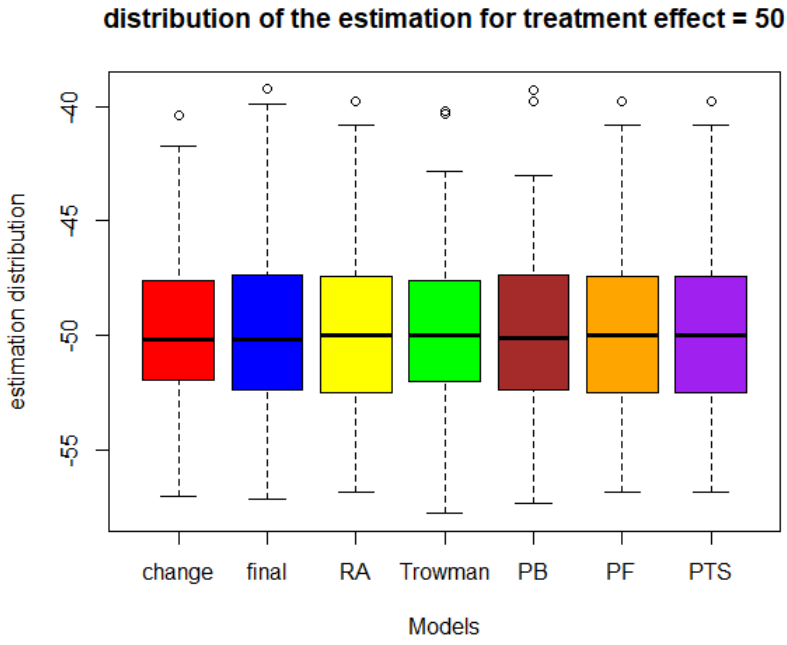

Figure B.5: The result for $n_{study} = 16$

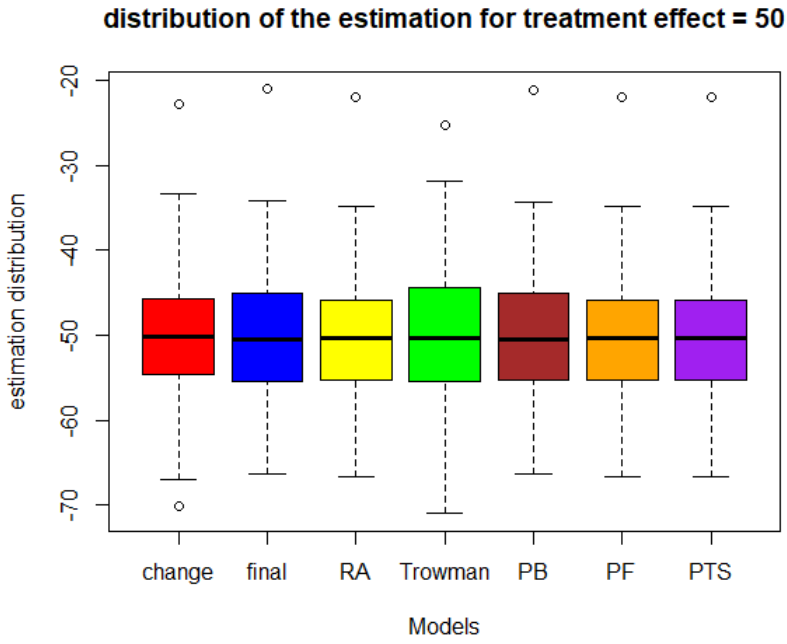

Figure B.6: The result for $n_{study} = 4$

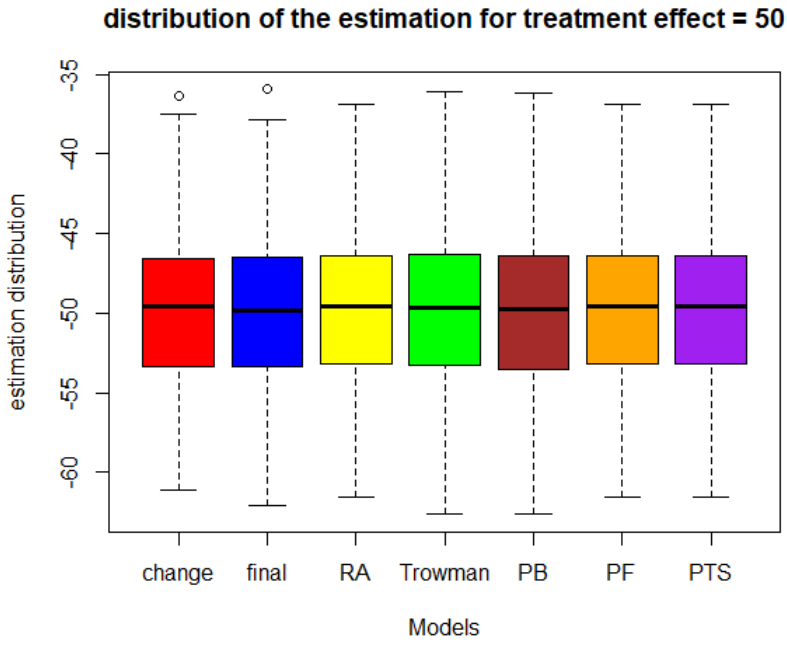

Figure B.7: The result for $\sigma_{YB} = 10$

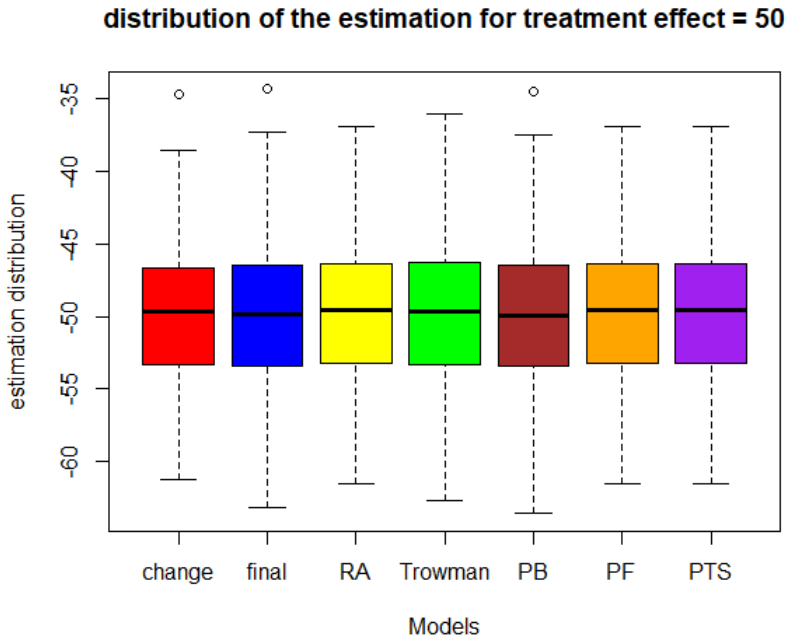

Figure B.8: The result for $\sigma_{YB} = 30$

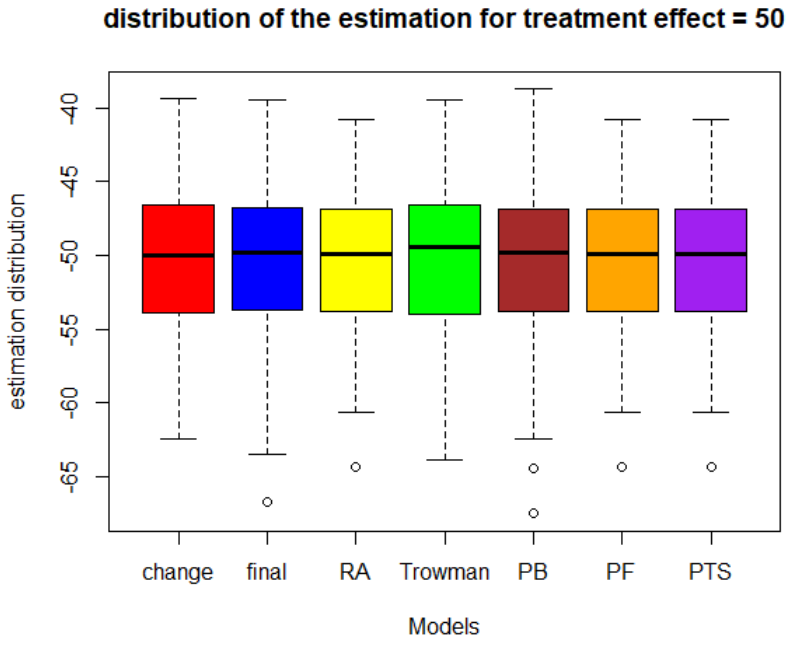

Figure B.9: The result for $n_{group} = 10$

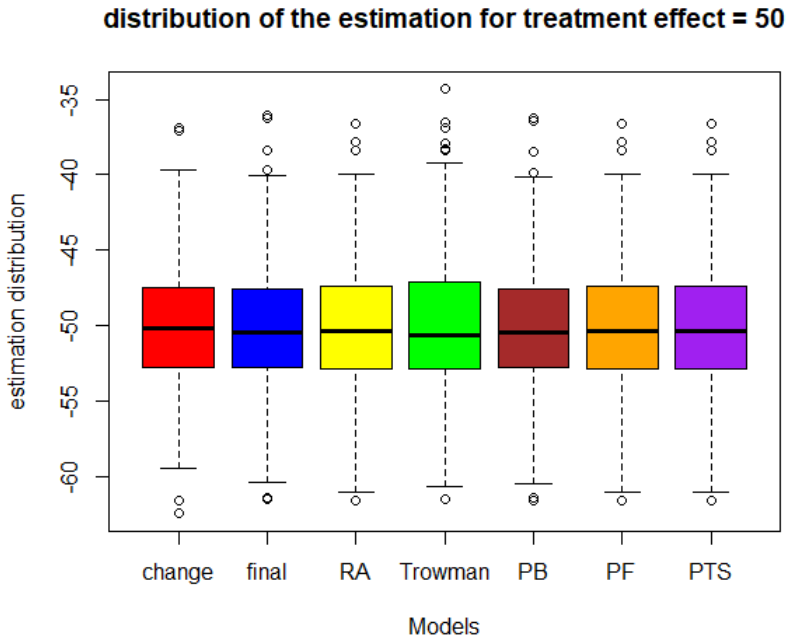

Figure B.10: The result for $n_{group} = 30$

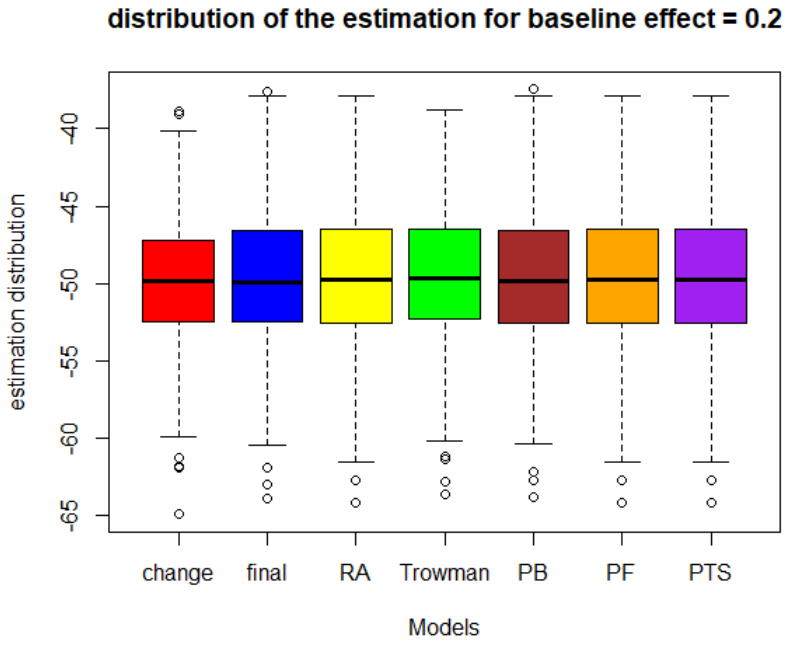

Figure B.11: The result for $\beta_2 = 0.2$ without imbalance

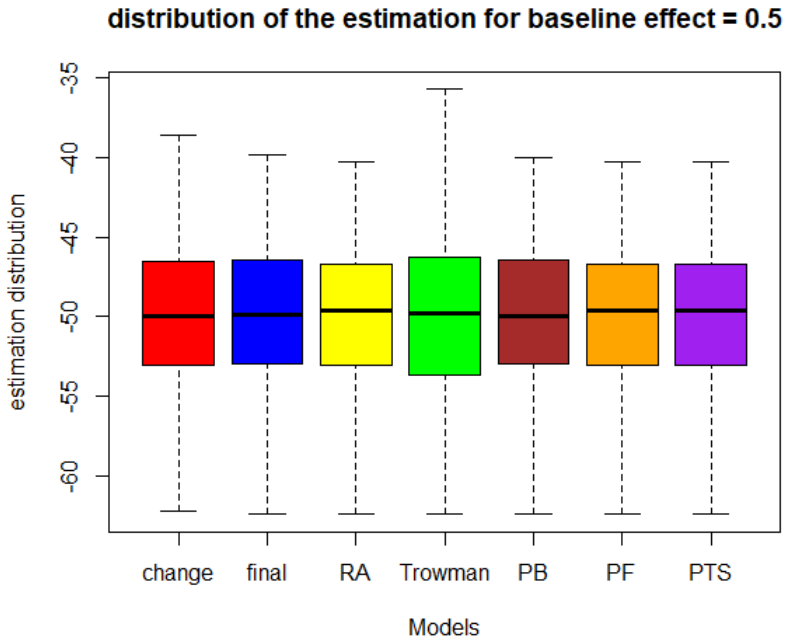

Figure B.12: The result for $\beta_2 = 0.5$ without imbalance

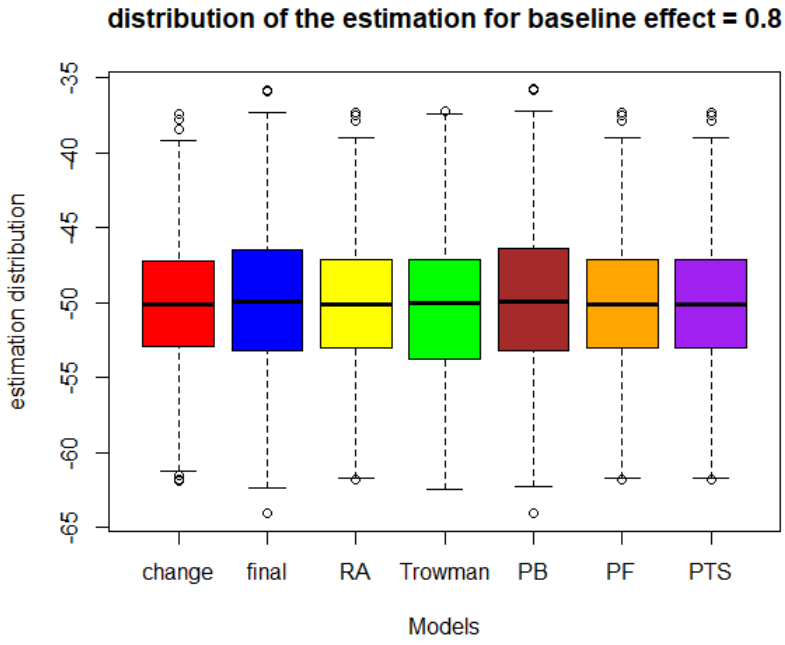

Figure B.13: The result for $\beta_2 = 0.8$ without imbalance

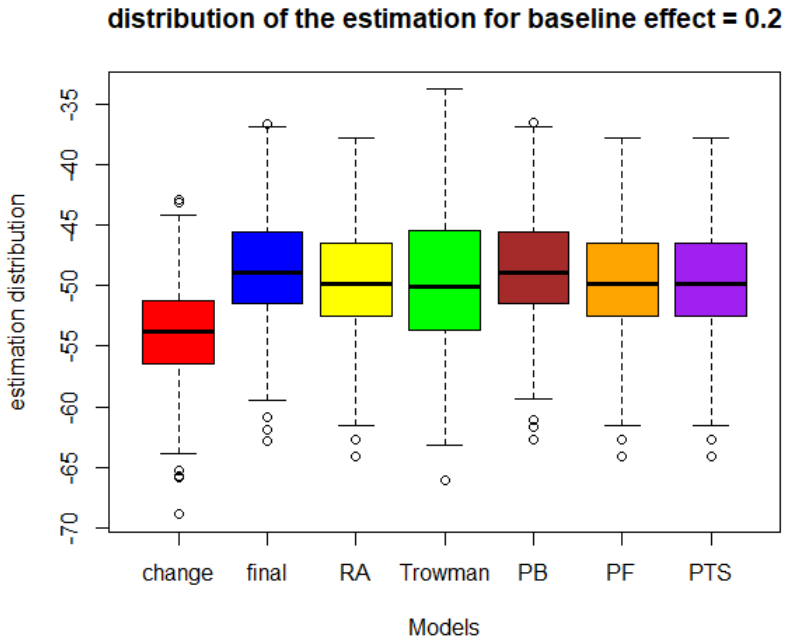

Figure B.14: The result for $\beta_2 = 0.2$ with imbalance

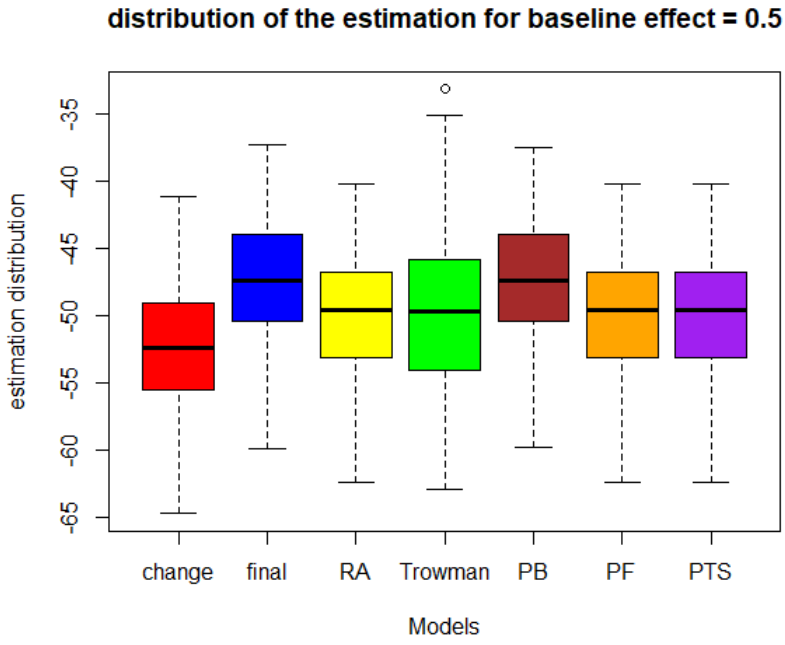

Figure B.15: The result for $\beta_2 = 0.5$ with imbalance

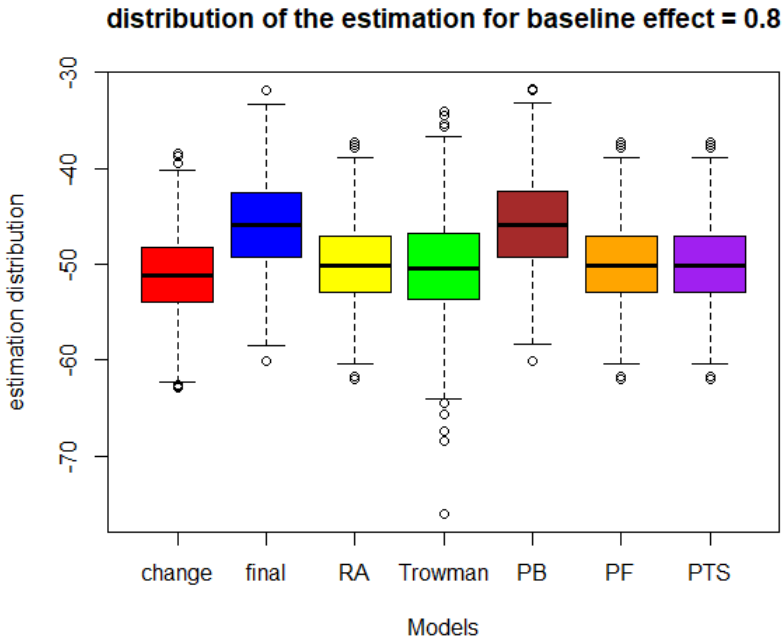

Figure B.16: The result for $\beta_2 = 0.8$ with imbalance

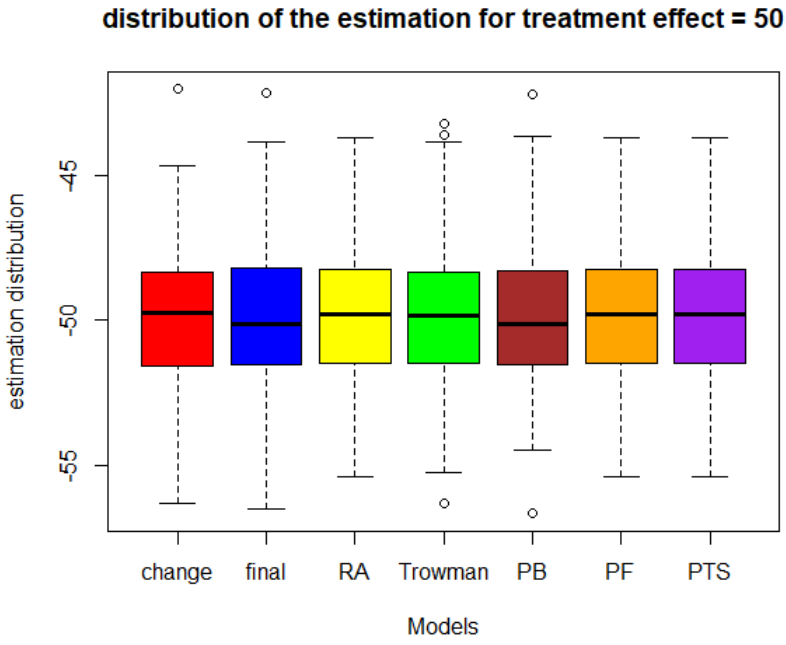

Figure B.17: The result for $\tau = 6$

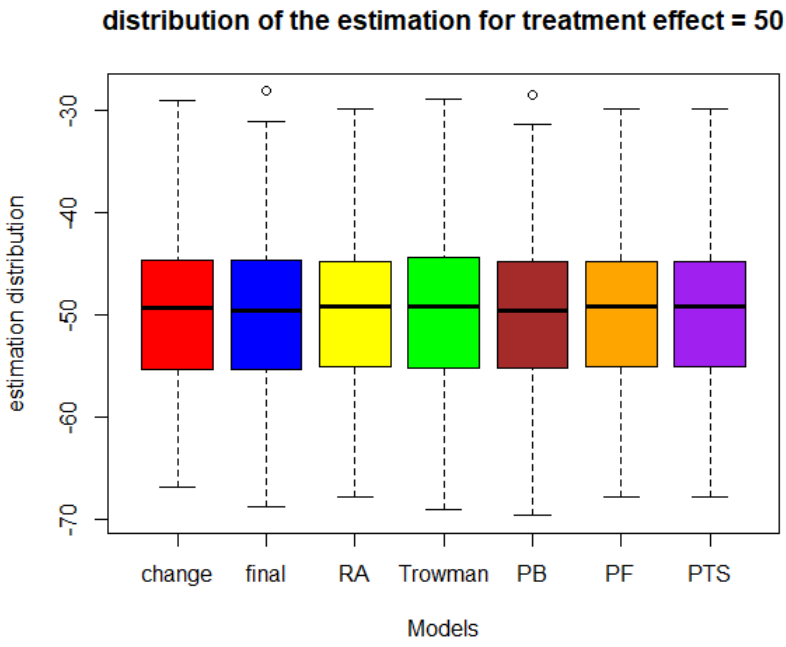

Figure B.18: The result for $\tau = 20$

Appendix C

R Code for this thesis

Researcher can access the R code of chapter 3 and chapter 4 for reproducing the results or further research need.