Choosing appropriate statistical models for multicenter randomized controlled trials with continuous and binary endpoints

by

Hui Wu

B.S., China Agricultural University, 2009
M.S., China Agricultural University, 2011

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Abstract

Many randomized controlled trials (RCTs) recruit individuals to multiple clinical centers rather than a single center. This strategy may provide a larger sample size and, therefore, have greater power to detect potential differences among experimental treatments. After balanced randomization, analyses of multicenter studies still need to appropriately adjust for the center effects and possibly other covariates to ensure valid inference. However, for real-world studies, regardless of whether the outcome is continuous or binary, it is often challenging to properly account for them or decide whether to remove the interaction effects, especially when the treatment arm ratio is unequal or the total sample size is small.

This work considered two projects. In the first project, (1) continuous outcomes were explored under a blocked randomization process, which can induce correlation among treatment groups and violate the statistical assumption that all individuals were independent. (2) The correct test statistics for the null hypothesis of no treatment effect under the mixed-effects models were derived under homogeneous and heterogeneous scenarios. (3) The method-of-moments based on Type III sum of squares and restricted maximum likelihood (REML) procedures with Kenward-Roger (KR*)* adjustment were examined for the impact on inference for center, treatment, and center-by-treatment interaction effects under mis-specified models in two simulations studies. One study considered only center effects, and the other considered center and block nested within the center (e.g., litter). In the second project, (1) the performance characteristics of Cochran-Mantel-Haenszel (CMH), generalized linear mixed model (GLMM), and generalized estimating equations (GEEs) for binary data were compared under both balanced and unbalanced designs via simulation under homogeneous scenarios. (2) Furthermore, these three primary methods were explored in various situations with and without accounting for the center or center-by-treatment interaction effects.

In summary, for both continuous and binary outcomes, it is important to assess whether or not treatment effects are heterogeneous across centers. In particular for a continuous outcome when the number of centers is small, whether or not to exclude the center-by-treatment interaction term should be determined by *p*-values from the Type III sum of the squares analysis instead of REML. The simulation results showed that REML was too conservative even with KR adjustment. When center-by-treatment was statistically significant, the center-by-treatment interaction term should be included in the model as random rather than fixed, otherwise Type I error rates for testing

treatment effects were inflated. However, when models were mis-specified (i.e., missing non-zero random effects), increasing the number of centers (n_c= 25) and considering the nested variable (e.g., litter) as fixed could improve model performance in terms of Type I error rate and power for inference on treatment effects. For a binary outcome, CMH is always recommended as the most preferable approach in homogeneous scenarios. When center-by-treatment effects were present in the data, GLMM incorporating center and center-by-treatment interaction as random could be used to provide precise and valid inference for treatment effect. When the number of centers was large (e.g., 25), GEE with robust standard errors (center as the subject for exchangeable working correlation structure) was shown to be the best option for both homogeneous and heterogeneous cases.

Choosing appropriate statistical models for multicenter randomized controlled trials with continuous and binary endpoints

by

Hui Wu

B.S., China Agricultural University, 2009
M.S., China Agricultural University, 2011

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Approved by:

Major Professor
Dr. Christopher I.Vahl

Copyright

Abstract

Many randomized controlled trials (RCTs) recruit individuals to multiple clinical centers rather than a single center. This strategy may provide a larger sample size and, therefore, have greater power to detect potential differences among experimental treatments. After balanced randomization, analyses of multicenter studies still need to appropriately adjust for the center effects and possibly other covariates to ensure valid inference. However, for real-world studies, regardless of whether the outcome is continuous or binary, it is often challenging to properly account for them or decide whether to remove the interaction effects, especially when the treatment arm ratio is unequal or the total sample size is small.

This work considered two projects. In the first project, (1) continuous outcomes were explored under a blocked randomization process, which can induce correlation among treatment groups and violate the statistical assumption that all individuals were independent. (2) The correct test statistics for the null hypothesis of no treatment effect under the mixed-effects models were derived under homogeneous and heterogeneous scenarios. (3) The method-of-moments based on Type III sum of squares and restricted maximum likelihood (REML) procedures with Kenward-Roger (KR*)* adjustment were examined for the impact on inference for center, treatment, and center-by-treatment interaction effects under mis-specified models in two simulations studies. One study considered only center effects, and the other considered center and block nested within the center (e.g., litter). In the second project, (1) the performance characteristics of Cochran-Mantel-Haenszel (CMH), generalized linear mixed model (GLMM), and generalized estimating equations (GEEs) for binary data were compared under both balanced and unbalanced designs via simulation under homogeneous scenarios. (2) Furthermore, these three primary methods were explored in various situations with and without accounting for the center or center-by-treatment interaction effects.

In summary, for both continuous and binary outcomes, it is important to assess whether or not treatment effects are heterogeneous across centers. In particular for a continuous outcome when the number of centers is small, whether or not to exclude the center-by-treatment interaction term should be determined by *p*-values from the Type III sum of the squares analysis instead of REML. The simulation results showed that REML was too conservative even with KR adjustment. When center-by-treatment was statistically significant, the center-by-treatment interaction term should be included in the model as random rather than fixed, otherwise Type I error rates for testing

treatment effects were inflated. However, when models were mis-specified (i.e., missing non-zero random effects), increasing the number of centers (n_c= 25) and considering the nested variable (e.g., litter) as fixed could improve model performance in terms of Type I error rate and power for inference on treatment effects. For a binary outcome, CMH is always recommended as the most preferable approach in homogeneous scenarios. When center-by-treatment effects were present in the data, GLMM incorporating center and center-by-treatment interaction as random could be used to provide precise and valid inference for treatment effect. When the number of centers was large (e.g., 25), GEE with robust standard errors (center as the subject for exchangeable working correlation structure) was shown to be the best option for both homogeneous and heterogeneous cases.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my gratitude to my advisor, Dr. Christopher I. Vahl for his guidance, support, and encouragement throughout this journey. His expertise, patience and mentorship have been invaluable to me. I am also profoundly grateful to my dissertation committee members, Dr. Qing Kang, Dr. Juan Du, and Dr. Majid Jaberi-Douraki, for their insightful feedback, constructive criticism, and scholarly guidance. Their expertise has enriched this work and contributed significantly to its refinement. I would like to extend my thanks to the faculty and staff of the department of Statistics for all the instances in which their assistance helped me along the way. At last, I would like to thank my family for the support they provided me through my study, especially my husband and two daughters. Their patience, unwavering love, and belief in me have sustained me during the challenging moments of this endeavor.

Chapter 1 - Introduction

Multicenter randomized clinical trials (RCTs) have been frequently employed in clinical research (Agresti and Hartzel, 2000; Babor, 2004; Chu et al., 2011; Pickering and Weatherall, 2007). Compared to single-center RCTs, one of the main reasons for conducting multicenter RCTs is that it could provide a feasible way to generate a larger sample size and then achieve greater power to detect the differences among experimental treatments (Kraemer, 2000; Localio et al., 2001). Additionally, multicenter trials allow treatments to be tested in various settings, providing a better basis for generalizing study results than results from one single-center trial (Lewis, 1999). However, the major disadvantage is the potentially increasing variability and heterogeneity across different centers. Variability in factors such as patient populations, healthcare practices, and adherence to protocols among multiple centers can introduce additional complexity and may compromise the study's internal validity. Managing and controlling for these variations becomes more challenging, making it harder to draw definitive conclusions about the intervention's effectiveness. Additionally, coordinating and standardizing procedures across diverse centers may increase logistical challenges, impacting the overall efficiency and consistency of the trial.

1.1 Center Effects

As expected, individuals within a single center may be more similar to each other than individuals from different centers, which is known as a center effect. The center effect can be caused by various factors, such as the characteristics of the patient population (*e.g.,* age and race), the specific clinical procedures, and the quality of staff at each center. In statistical terms, observations within a center tend to be correlated; those in different centers should be independent.

The center effect can lead to bias in multicenter RCTs if not adequately accounted for in the trials. Simply pooling data as if they arose from one population (ignoring the center effect)

could provide biased results (Chu et al., 2011). Two primary statistical methods are accounting for the center effect. One is to include the center as a random effect in the statistical analysis, allowing for the variation between centers to be considered and separated from the variation within centers. Another way is to stratify the sample by center so that the results from each center can be analyzed separately. This statistical method could identify specific patterns or trends unique to a particular center, which helps understand its characteristics and potentially make improvements or adjustments to meet the needs of the individuals it serves.

The regulatory guideline suggests that the heterogeneity of treatment effects across centers should be explored first. Still, it should be noted that this model with such statistical significance tests generally has little power to detect the main effects of treatment (ICH, 1998).

1.2 Other Covariates

Besides center effects, it is also necessary to address the adjustment problem for other covariates at the center level. There are many different types of covariates at the center level, such as demographic factors (e.g., age, gender, weight, and race), disease characteristics (e.g., duration or severity), and management teams (Lachin, 2009). Researchers should identify those covariates that are expected to have a significant impact on the primary variables. Furthermore, it is crucial to consider how to account for these factors in the randomization procedure and the analysis. When more than two factors are used to stratify the design procedure, adjusting all these factors in the primary analysis is challenging.

Moreover, since the treatment effect may vary across subgroups (heterogeneity), a statistical model including interactions should also be explored in a planned confirmatory analysis first (ICH, 1998). Although the researchers expected the level of imbalance in these trials to be zero after all considerations, no single trial is likely to achieve complete balance on all critical

prognostic variables. Much literature illustrates the effect of the imbalance of prognostic variables on statistical inference of treatment effect (Ciolino et al., 2015; Pocock et al., 2002).

1.3 Randomization

Randomization is the most essential feature of RCTs, as it inherently balances known and unknown baseline prognostic factors between treatment groups. Moreover, it plays a crucial role in minimizing selection bias on average. In mainly speaking, randomization refers to randomly arranging participants into treatment and control groups, assuming that each participant has an equal chance of being assigned to any group (Fleiss et al., 2013). It could reduce bias and control for confounding variables in clinical studies.

1.3.1 Categories

There are several different techniques of randomization, including treatment balance (simple, block, and urn randomization) and covariate balance (stratification and minimization randomization) within treatments (Altman and Bland, 1999; Callegaro et al., 2021; Ma et al., 2020; Taves, 1974). Several methods might be combined to achieve both treatment balance and covariate balance across treatments (Hedden et al., 2006).

1.3.1.1 Treatment Balance

a) Simple Randomization

Simple randomization is the most straightforward approach to the randomization procedure. For instance, when assigning subjects to groups A and B, the assignments are made entirely random for each allocation. Typically, simple randomization could be reliable in generating equal numbers of participants among groups when the total number of samples exceeds 100 (Kim & Shin, 2014). When the total number of samples is small, it could lead to an uneven number of participants among groups across some important prognostic factors. Kang et al. (2008)

demonstrated an example of this imbalance that employing a coin toss with a small sample size (e.g., n = 10) might lead to assigning 7 participants to the control group and 3 to the treatment group (shown in Figure 1.1).

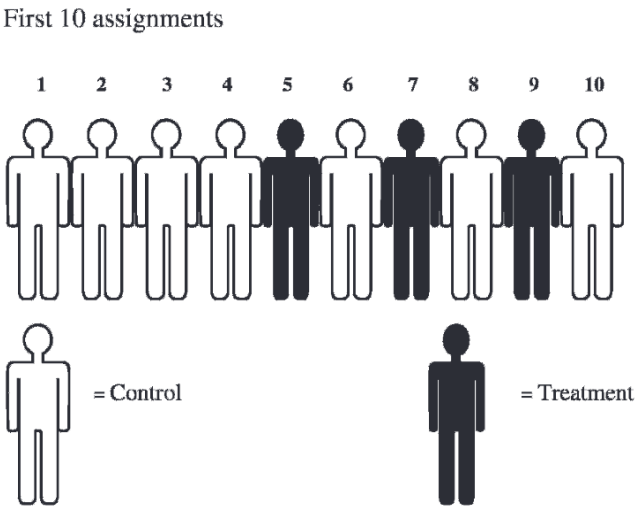


Figure 1.1. An example of the imbalance of sample size between treatment arms due to simple randomization (coin toss) in a small trial (n = 10) from Kang et al. (2008).

b) Block Randomization

The second method is the block randomization, designed to enhance the balance of treatment assignments over time. In this approach, participants are grouped into small blocks, and within each block, the allocation to different treatment groups is random. The primary advantage of block randomization is that it ensures more even numbers of participants among treatment groups at various stages of the study, thereby minimizing the risk of imbalances in important prognostic factors. Altman and Bland (1999) suggested that block size should be several times that of treatment arms (*i.e.,* with two treatment groups, block size of either 4 or 6), which the researcher should determine.

While block randomization offers advantages in maintaining balance in treatment group allocations, it also comes with disadvantages. One notable drawback of this method is groups may

be generated that are rarely comparable concerning specific covariates (Hedden et al., 2006). Such an imbalance could introduce bias and reduce the power of the study, mainly when dealing with small sample sizes (Schulz & Grimes, 2002).

1.3.1.2 Covariate Balance

a) Stratified Randomization

This approach addresses the necessity of controlling and balancing the impact of covariates across different treatment groups. Thus, researchers' identification of relevant stratification covariates that potentially influence the dependent variable, such as age, gender, or disease severity, is crucial in this process (Kernan et al., 1999). By considering these factors, the randomization process is stratified, leading to more homogeneous treatment groups with respect to the specified characteristics. Participants are first categorized into different strata based on several specific characteristics, and randomization is then independently assigned to each corresponding covariate stratum. Subsequently, one of the randomization techniques (*i.e.,* simple, block, or urn randomization) for treatment balance could be applied within each stratum to allocate participants to one group randomly.

Commonly, stratified randomization is a straightforward and valuable technique, especially for smaller clinical studies. Still, it becomes complicated to implement if too many covariates need to be controlled (Weir & Lees, 2003). Another limitation is that it depends on having all participants identified before group assignment since covariates must first categorize participants. This method is rarely applicable in practice, as clinical trial participants are typically enrolled continuously, one at a time. The difficulty arises when the baseline characteristics of all participants are not accessible before the assignment, which makes implementing stratified randomization challenging (Lachin et al., 1988).

b) Minimization

When many important prognostic factors need to be considered, a randomization procedure called covariate-adaptive can be used to balance selected covariate factors (Lin et al., 2015). Minimization is the oldest and most popular covariate-adaptive randomization method, first developed by Taves (1974) and then expanded by Pocock and Simon (1975). After it considers relevant prognostic factors for the study outcomes, minimization achieves an excellent balanced distribution of these factors across treatment groups, enhancing the comparability of the groups.

Despite increased RCTs on covariate adaptive randomization, some arguments still arise. Due to its potential impact on subsequent analysis, regulatory agencies have questioned minimization. Furthermore, the U.S. Food and Drug Administration (FDA) could commonly require an additional complimentary analysis of data, typically as sensitivity analysis using a re-randomization test (Callegaro et al., 2021).

1.3.2 Concerns

Usually, which randomization method will be applied based on the specific study design and the type of collected data. To minimize the impact of center and other covariates on treatment effect, stratified randomization and minimization randomization have been more often commended for RCTs. Scott et al. (2002) showed that in 2001, 45% of randomized trials in the Lancet or the New England Journal of Medicine used either stratification or minimization. More recently, Pond et al. (2010) manually reviewed 476 clinical papers published in 13 major oncology journals from 1995-2005 and found that 84.7% of trials stratified on at least one baseline variable in their randomization.

Nevertheless, randomization does not guarantee the complete balance of participant characteristics, especially when the sample size is small (Chu et al., 2011). Meanwhile,

stratification and minimization of randomized controlled trials will arise another concern. Because randomization inevitably uses the covariate information to balance treatment groups, the validity of classical statistical methods after randomization is often unclear. In other words, it could introduce a correlation among treatment groups (Kahan & Morris, 2012; Kang et al., 2008), which violates the statistical assumption that all patients are independent. This "cluster" effects could be quantified by the intra-class correlation coefficient (ICC, ρ), which is defined as the ratio of its variance between clusters to its total variance (sums of between and within clusters) (Rabe-Hesketh & Skrondal, 2008). The range of ρ is between zero and one, and the larger ρ values indicate a higher level of correlation between individuals in the same "cluster", which means individuals within "cluster" contain less unique information. Also, this implies that the independence assumption for many statistical models is violated.

An improper analysis could potentially result in a beneficial treatment being denied to patients or a treatment not beneficial being adopted. Therefore, when centers or other covariates have been involved in randomization, it is necessary to adjust appropriately for these variables in the analysis to obtain valid results. Scott et al. (2002) suggested that adjustments should always be made for minimization factors when analyzing trials where minimization is the allocation method used. Kernan et al. (1999) suggested that investigators should choose factors carefully and account for them in the analysis once stratified randomization is made in the trial. Kahan and Morris (2012) state that if the correlation is induced when using stratification and performing unadjusted analysis, it will give invalid inference regarding low type I error rates and a power reduction.

Ideally, the covariate factors used in randomization should also be integrated into the subsequent analysis according to regulatory guidelines (ICH E9, 1998). However, suppose more than two or three important prognostic factors are used as stratification factors. In that case, it is

less successful at achieving balance or accounting for these stratification factors in subsequent analysis. Thus, unadjusted analyses are still commonly used in practice. Kahan and Morris (2012) reviewed the medical literature published in four major medical journals (the Lancet, the British Medical Journal, the Journal of the American Medical Association, and the New England Journal of Medicine) in 2010. Only 14 of 41 eligible studies reported accounting for their primary analysis for stratification or minimization variables. Also, Pocock et al. (2002) summarized several vital statistical issues that exist in the adjusted analysis, such as the overuse and overinterpretation of subgroup analyses, contradictions in the use of covariate adjustment, the lack of clear guidelines on covariate selection, the inappropriate application of statistical significance tests in assessing the comparability of baseline covariates, and the importance need to have a predefined statistical analysis plan encompassing all uses of baseline data. Hence, many working models may encounter the challenge of invalid statistical inference.

1.4 Intraclass Correlation Coefficient

After the randomization process, participants in the same 'block' tend to have correlated outcomes, meaning they are more similar to other individuals in the same 'block' than individuals from other 'blocks'. This similarity can be measured by the intraclass correlation coefficient (ICC, ρ), which can be fundamentally understood as the proportion of total variability explained by the variance between blocks. (Fisher, 1928). The fundamental definition of ICC is derived from the expression for the correlation between responses from two different individuals within the same 'block'. For instance, the correlation between the *j*th and *l*th responses in the *i*th 'block' can be expressed as:

$$\rho = cov(y_{ij}, y_{il}) / \sqrt{var(y_{ij})\, var(y_{il})} \quad \text{for } j \neq l. \tag{1.1}$$

The algebraic form for ICC depends on the outcome type and model researchers apply. There are two main modeling approaches, including (i) directly modeling the covariance structure in the data, and (ii) specifying the variation at the cluster and individual levels separately in a hierarchical model, then indirectly inducing the correlation structure (Eldridge et al., 2009). This study only focuses on the expression of ρ for continuous and binary outcomes under a hierarchical model. Therefore, hierarchical models usually need to specify a distribution for each level of data. For instance, a model with two levels is defined by separate distribution for the 'blocks' and for individuals within 'block'.

For continuous outcomes, McGraw and Wong (1996) derived ten forms of ICC based on five different cases (Selected results are shown in Table 1.1). When applied to the same data set, these various forms of ICC could result in different conclusions. Mixed effect models are usually used to estimate ICC since they can obtain an estimated random effect value directly using maximum likelihood estimation.

Table 1.1. Definitions of ICC for selected models with continuous outcomes by McGraw and Wong (1996).

Case	Name	Model[1]	Note	Forms for ICC[2]
1	One-way random effects	$y_{ij} = \mu + r_i + e_{ij}$	r_i is random	$\frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2}$
2A	Two-way random effects, with interaction	$y_{ij} = \mu + r_i + c_j + rc_{ij} + e_{ij}$	r_i, c_j and rc_{ij} are random	$\frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2}$
2B	Two-way random effects, without interaction	$y_{ij} = \mu + r_i + c_j + e_{ij}$	r_i, and c_j are random	$\frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_e^2}$
3A	Two-way mixed effects, with interaction	$y_{ij} = \mu + r_i + c_j + rc_{ij} + e_{ij}$	c_j is fixed, r_i and rc_{ij} are random	$\frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + \sigma_{rc}^2 + \sigma_e^2}$
3B	Two-way mixed effects, without interaction	$y_{ij} = \mu + r_i + c_j + e_{ij}$	c_j is fixed, and r_j are random	$\frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2}$

[1]where i = 1, …, n and j = 1, …, k.
[2]ICC: intraclass correlation coefficient.

For dichotomous outcomes, Gulliford et al. (2005) found a relationship between ICC and the overall prevalence, and Evans et al. (2001) also discovered that the definition of ICC may depend on the type of model adopted to analyze the data. Eldridge et al. (2009) presented two expressions for ICC under hierarchical models: modeling the cluster-level proportions or the cluster-level log-odds. The latter approach used to be under the logistic-normal model as follows,

$$y_{ij} \sim Bernoulli(p_i)$$

$$logit\ (p_i) = log\frac{p_i}{1-p_i} = \mu + u_i \tag{1.2}$$

where y_{ij} is a binary outcome generated from a Bernoulli distribution (p_i). Meanwhile, u_i is the random effect of the i^{th} stratum (*i.e.,* Center), which is generally assumed to follow a normal distribution with mean 0 and variance σ_c^2. The parameter σ_c^2 which is expressed on the log-odd scale, summarizes the variation among clusters, whereas the overall outcome variance $p(1-p)$ is as the proportion scale. Since there is no close form for expressing the relationship between the log-odd scale σ_c^2 and the proportion scale $p(1-p)$ (Goldstein et al., 2002; Turner et al., 2001), Eldridge et al. (2009) introduced an alternative ICC using the log-odds scale without considering covariates:

$$\rho_c = \frac{\sigma_c^2}{\sigma_c^2 + \frac{\pi^2}{3}} \tag{1.3}$$

With the assumption that the model-based error term follows a logistic distribution. The variance of the standard logistic distribution is ($\pi^2/3$), where π is the mathematical quantity (~3.14159).

1.6 Objective

To date, limited studies have considered the issue of interaction (heterogeneity) in their simulations. Although regulatory guidelines recommended that the statistical analysis for multicenter/multi-region RCTs evaluate the consistency of treatment effect across subgroups, there is no clear guideline for the appropriate statistically significant test for heterogeneous treatment across subgroups, especially for data with binary outcomes.

The main goal of this research is to (1) emphasize the need for careful consideration of methods with inference from heterogeneous data; (2) provide practical recommendations for a statistically significant test for heterogeneous treatment across subgroups; (3) develop an appropriate strategy when and how to adjust for the center and other covariates in the analysis for the continuous and binary outcomes under homogeneous and heterogeneous scenarios.

Chapter 2 - Multicenter Randomized Clinical Trials with Continuous Outcomes

Multicenter randomized clinical trials are often carried out in practice, and significant challenges are faced in analyzing trials. When balanced randomization is implemented to ensure that treatment allocation is balanced within the 'block' (*i.e.,* center), then when and how to adjust for these factors is crucial for the analysis, especially under homogeneous or heterogeneous cases. Many researchers performed simulation studies to investigate the appropriate model under different ICC values for RCTs with continuous outcomes (Chu et al., 2011; Kahan & Morris, 2013a; Localio et al., 2001; Pickering & Weatherall, 2007). Generally, most of these studies assume that the treatment effects were homogeneous among 'block' groups, that is no block-by-treatment interaction exists (σ_{rc}^2 is zero). However, in real-world data, this assumption may not be held (heterogeneous scenarios may happen). As discussed before, the center effect can also be caused by various factors, especially the characteristics of the population (*e.g.,* age, gender and race), which means that the treatment effect could be heterogeneous for a particular subpopulation. An ascertain analysis for heterogeneous treatment across subgroups is essential before the primary analysis.

Furthermore, as McGraw and Wong (1996) stated, the forms of ICC for heterogeneous cases differ from those for homogeneous cases. The recommendations on adjusting center and other covariate factors for various ICC values, derived from studies in homogeneous scenarios, may not be suitable for heterogeneous cases. Recall the two-way mixed model in Chapter 1.4.

$$y_{ij} = \mu + \tau_i + b_j + \tau b_{ij} + e_{ij}, \text{ with } i = 1, \ldots., t, j = 1, \ldots, c \tag{2.1}$$

where y_{ij} represents a continuous outcome measured for the subject in the *j*th 'block' factor, which is also one factor or combination group used in the randomization process receiving the *i*th treatment. μ represents an intercept term (the overall response mean). The treatment effects τ_i's are fixed, with $\sum_i \tau_i = 0$. The term b_j denotes the random effect of the *j*th 'block', and τb_{ij} denotes the random interaction effect on the continuous outcomes, and e_{ij} denotes the random error term. The additional assumptions regarding the random variables within this model include as

$$b_j \sim i.i.d.\, N\,(0, \sigma_b^2)$$

$$\tau b_{ij} \sim i.i.d.\, N.\,(0, \sigma_{\tau b}^2)$$

$$e_{ij} \sim i.i.d.\, N\,(0, \sigma_e^2)$$

When $\sigma_{\tau b}^2$ = 0, meaning that the treatment effects were homogeneous among 'block' groups, then ICC_b is defined as

$$ICC_b = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2} \tag{2.2}$$

When $\sigma_{\tau b}^2 \neq 0$, which means the treatment effects were heterogeneous among 'block' groups, then ICC_b is defined as

$$ICC_b = \frac{\sigma_b^2 - \sigma_{\tau b}^2/(k-1)}{\sigma_b^2 + \sigma_{\tau b}^2 + \sigma_e^2} \tag{2.3}$$

Mathematically, when $\sigma_{\tau b}^2 \neq 0$, if accidentally using Eq. 2.2 to estimate ICC_b, the true ICC_b could be overestimated. Furthermore, this can be problematic because the final models must be determined by researchers considering suggestions from the literature related to ICC_b, which may not be applicable.

Commonly, linear fixed/mixed-effect models are tools for analyzing multicenter datasets with continuous response. The fixed-effect model makes strong assumptions regarding the independence of observations and the distribution of residuals, while the mixed-effect model makes more complex assumptions, particularly about the distribution of random effects and residuals. Also, it allows researchers to consider heterogeneous situations in models. Violations of these assumptions often happen in the real world, yet it is unclear how much these violations impact accurate and unbiased estimation of interest or hypothesis testing.

For this chapter, we investigated (1) the appropriate method to test the existence or lack of heterogeneous treatment effects across 'block'; (2) the consequences of missing random effect components on model estimates of interests.

2.1 Center Effect without Other Covariates

Many researchers have investigated when and how to account for center effects when the center is applied as a stratification variable in the randomization process for multicenter trials with continuous outcomes (Chu et al., 2011; Kahan and Morris, 2013; Pickering and Weatherall, 2007). When using individuals as a unit of analysis, there are three commonly used approaches: ignoring centers, including centers as fixed effects, and including them as random effects.

Thus, simple linear regression completely ignoring centers is inappropriate in general cases. Otherwise, one would assume that the data is homoscedastic by ignoring the centers, which means that the variance of the error is constant across centers. This strong assumption may not hold because there is a potential correlation among patients within centers (Parzen et al., 1998). Although stratified randomization has already attempted to reduce the impact of center on the treatment effect's standard error and improve the study's overall validity, it is still essential to adjust stratified randomization variables (e.g., center) in the analysis. The survey by Kahan and Morris

(2012) shows that an unadjusted analysis could cause the standard errors for the treatment effect to be biased upwards, resulting in wide 95% confidence intervals and low type I error rates. Only when center effects and size are small can ignoring the center perform better than the fixed effect model (Pickering & Weatherall, 2007; Kahan & Morris, 2013).

Therefore, accounting for center effects is crucial in randomized trials, particularly in large multicenter studies (Localio et al., 2001). The fixed-effects model is the most common and standard statistical adjustment method (Tangri et al., 2010). This method treats each center as a fixed intercept to control for possible population or environmental differences among centers. However, this model must assume that study subjects from the same center have independent outcomes (i.e., the intraclass correlation coefficient ρ statistically is fixed at zero), which may also not hold since the design has a potentially high correlation (Kahan & Morris, 2012). Meanwhile, Pickering and Weatherall (2007) have shown that fixed effect analysis was less robust, mainly when center effects were small.

The alternative method accounting for center effects is the mixed-effects model, which incorporates dependence of outcomes within a center and treats centers as random intercepts. The study from Kahan and Morris (2013) showed that random-effects models offer many advantages over fixed-effects models in certain situations. For instance, random center effects models increased power and precision when the number of patients per center was small (< 10 patients), or an imbalance between treatments within centers due to the randomization method or the distribution of patients across centers.

For the details of estimation and testing of treatment effects, regulatory guidelines (ICH E9, 1998) recommend that,

1) For the fixed-effects model, “the main treatment effect may be investigated first using a model that allows for center differences, but does not include a term for treatment-by-center interaction”. “If the treatment effect is homogeneous across centers, the routine inclusion of interaction terms in the model reduces the efficiency of the test for the main effects”.
2) Experts may also explore the heterogeneity of the treatment effect for the mixed effect model. “These models consider center and treatment-by-center effects to be random and are especially relevant when the number of sites is large”.

This statement has triggered controversy in statistical circles and caused in some challenges in both the design and analysis of multicenter RCTs (Worthington, 2004). Regardless of the model employed, the assumptions associated with it must be diagnosed by researchers first. These should include assessing the validity of the model, independence of data points, linearity in the predictor-response relationship, absence of measurement errors in the predictor, homogeneity of residuals, independence of random effects versus covariates (homogeneity), identification of data missing completely at random, and evaluating assumptions related to the distribution of residuals and random effects. Worthington (2004) provided several opinions how on the proper modeling approaches, including (*i*) analysis of the full model at first, including terms for treatment, center, and their interaction; (*ii*) secondly, working on the full model but removing the interaction depending on its level of significance; and (*iii*) finally using a reduced model, perhaps augmenting it with an interaction term if secondary analyses suggest its presence.

However, limited studies provided recommendations for how to appropriately remove the interaction term; only an early work suggested when the *p*-value for the test was 0.5, 0.25, or even as low as 0.10 in a fixed-effects model (Schwemer, 2000). The concern is that different statistical

tests for interaction terms led to inconsistent *p*-values, which may result in misinterpretation. If trials are balanced that is treatment allocated an equal number into each center, the decision from a fixed-effect model to drop the center or interaction term from the full model depends on the estimate of the residual variance and degrees of freedom of the test statistics. Typically, it is impractical that sample sizes at different clinical centers are controlled to be balanced for multicenter RCTs. Worthington (2004) found that there was controversy about the use of Type III or Type II analysis in a fixed-effect model, such as Type II provided the most powerful test for treatment effect if no interaction existed, or Type III led to an unbiased estimate for treatment effect if interaction exists. For mixed models, there are limited relative studies for the details (Verbeke & Molenberghs, 2003).

2.1.1 Simple Linear Regression

This approach estimates the effect of treatment (τ) on outcome (Y) ignoring centers via the following equation.

$$Y_{ijk} = \mu + \tau_i + e_{ijk}, i = 1,2, j = 1, \ldots, c, k = 1, \ldots, r \quad (2.4)$$

where Y_{ijk} represents a continuous outcome measured for the *k*th subject in the *j*th center receiving the *i*th treatment (control or treatment group). μ represents an intercept term (the overall response mean). The treatment effects τ_i's are fixed, with $\sum_i \tau_i = 0$. We assume that all centers are of equal size ($2r$) and that subjects are equally allocated among all treatment arms within each center. The number of subjects in each arm is rc, and the total number of subjects is $N = 2rc$. The e_{ijk} as the residual variance is assumed to be independent and identically distributed $N(0, \sigma_e^2)$.

When this model is true with valid assumption, estimation and test hypothesis of treatment effect could be straightforward. However, in a multicenter RCT study with a nonzero center effect,

a randomization procedure was done in the design stage. Then, $e_{ijk} = b_j + \varepsilon_{ijk}$, where the center effects $\{b_j\}$ are independent and identically distributed (*i.i.d.*) $N(0, \sigma_c^2)$ and $\sigma_c^2 \geq 0$, while the $\varepsilon_{ijk} \sim iid\ N(0, \sigma_\varepsilon^2)$. This could induce the ICC, ρ_c, defined as the proportion of the total variance due to the between-center variability (Eq. 2.2, form as homogeneous case). Let Y_{1jk} and $Y_{2jk'}$ be the outcomes from the *j*th center, assigned to treatments 1 and 2, respectively. Denote $var(Y_{1jk}) = var(Y_{2jk'}) = \sigma^2 = \sigma_\varepsilon^2 + \sigma_c^2$. The correlation between subjects in the *j*th center is ρ_c and the correlation between subjects with different centers is 0. It follows that $cov(Y_{1jk}, Y_{2jk'}) = \rho_c \sigma^2$. Let $\bar{Y}_1$ and $\bar{Y}_2$ be the mean outcomes in treatment group 1 and 2, respectively. The null hypothesis of no treatment difference is, $H_0: \tau_1 - \tau_2 = 0$.

The *t*-test Statistic

In the context of a two-arm trial, this model (2.1) is the same as a two-sample *t*-test

$$t = \frac{\hat{\tau}_1 - \hat{\tau}_2}{\sqrt{\widehat{var}(\tau_1 - \tau_2)}} \tag{2.5}$$

The ordinary least squares estimate of $\tau_1 - \tau_2$, naively assuming subject outcomes within the center are independent, is given by

$$\hat{\tau}_1 - \hat{\tau}_2 = \bar{Y}_1 - \bar{Y}_2 \tag{2.6}$$

We could rewrite *t* statistic as

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{var(\bar{Y}_1 - \bar{Y}_2)}} \tag{2.7}$$

where the variance of the treatment difference, $\bar{Y}_1$- $\bar{Y}_2$, can be found using the following formula:

$$var(\bar{Y}_1 - \bar{Y}_2) = var(\bar{Y}_1) + var(\bar{Y}_2) - 2cov(\bar{Y}_1, \bar{Y}_2)$$

(2.8)

If $\bar{Y}_1$ and $\bar{Y}_2$ are independent, $cov(\bar{Y}_1, \bar{Y}_2)$ would be zero. However, this does not hold because the center naturally clusters study subjects:

$$cov(\bar{Y}_1, \bar{Y}_2) = cov\left(\frac{1}{cr}\sum_{j,k} Y_{1jk}, \frac{1}{cr}\sum_{j,k} Y_{2jk'}\right)$$

$$= \frac{1}{(cr)^2}\sum_{j,k} cov\left(Y_{1jk}, Y_{2jk'}\right)$$

$$= \frac{\rho_c \sigma^2}{cr}$$

(2.9)

Note that $var(\bar{Y}_1) = var(\bar{Y}_2) = \frac{\sigma^2}{cr}$; the correlation between $\bar{Y}_1$ and $\bar{Y}_2$ is ρ_c, which is the same as Y_{1jk} and $Y_{2jk'}$.Consequently, the true variance of the treatment difference is

$$\frac{2\sigma^2}{cr} - \frac{2\rho_c \sigma^2}{cr} = \frac{2(1-\rho_c)\sigma^2}{cr}$$

(2.10)

yielding an elementary formula different from the classical one by a $(1-\rho_c)^{-\frac{1}{2}}$ factor.

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{2\sigma^2/cr}}(1-\rho_c)^{-\frac{1}{2}}$$

(2.11)

If the correlation ρ_c is positive between Y_{1jk} and $Y_{2jk'}$ in the same center, the usual variance estimate $\frac{2\sigma^2}{cr}$ is bigger than the true variance.

2.1.2 Fixed-effects Model

Let's consider a general fixed-effects model with treatment by center interaction term.

$$Y_{ijk} = \mu + \tau_i + b_j + (\tau b)_{ij} + \varepsilon_{ijk}, i = 1, 2, j = 1, \ldots, c, k = 1, \ldots, r$$

(2.12)

where Y_{ijk} represents a continuous outcome measured for the *k*th subject in the *j*th center receiving the *i*th treatment (control or treatment group). The term μ represents an intercept term (the overall response mean). The treatment effects τ_i's are fixed, with $\sum_i \tau_i = 0$. Also, the additive fixed center effects b_j's, and interaction effects $(\tau b)_{ij}$'s, are unknown parameters. The ε_{ijk} is the residual variance assumed to follow a normal distribution $N(0, \sigma_\varepsilon^2)$. Clearly, if the assumption for the fixed-effects model all holds, the analysis will be straightforward. Otherwise, we will have the same issue as the previous simple linear regression model, which ignores the correlation between subjects within centers.

2.1.3 Mixed-effects Model

Consider the general two-way mixed model with treatments being fixed effect and centers being a random effect as follows,

$$Y_{ijk} = \mu + \tau_i + b_j + (\tau b)_{ij} + \varepsilon_{ijk}, i = 1, \ldots, t, j = 1, \ldots, c, k = 1, \ldots, r_{ij} \tag{2.13}$$

where Y_{ijk} represents a continuous outcome measured for the *k*th subject in the *j*th center receiving the *i*th treatment. μ represents an intercept term (the overall response mean). The treatment effects τ_i's are fixed, with $\sum_i \tau_i = 0$. The term b_j denotes the random effect of the *j*th center, and $(\tau b)_{ij}$ denotes the random interaction effect on the continuous outcomes that is specific to the *j*th center receiving the *i*th treatment, and ε_{ijk} denotes the random error term associated with the *k*th subject in the *j*th center receiving the *i*th treatment.

Since there are two types of factors in a mixed model, the resulting model has two parts: a random effects part and a fixed effects part. The fixed effect part of this model is $\mu + \tau_i$, and the random effect part is $b_j + (tb)_{ij} + \varepsilon_{ijk}$. The additional assumptions include the following,

$$b_j \sim i.i.d. N(0, \sigma_c^2)$$

$$(\tau b)_{ij} \sim i.i.d.\, N(0, \sigma_{tc}^2)$$

$$\varepsilon_{ijk} \sim i.i.d.\, N(0, \sigma_{\varepsilon}^2)$$

(2.14)

and b_js, $(\tau b)_{ij}$s and ε_{ijk}s are all independent random variables. The terms corresponding to the design structure are not included in the above model. In matrix notation, the general linear mixed model expressed as the marginal distribution of y is given as

$$y = X\beta + e$$

where $e = Z_1 u_1 + Z_2 u_2 + \varepsilon$, with $Var\,(e) = \Sigma$, and

$$\Sigma = \sigma_c^2\, Z_1 Z_1' + \sigma_{tc}^2 Z_2 Z_2' + \sigma_{\varepsilon}^2 I_N$$

(2.15)

where y is an observed $N \times 1$ data vector, $X\beta$ is the fixed effects part of the model. u_i $(i = 1, 2)$ and ε are independent random variables.

2.1.3.1 Analysis of the Random Effects

The first step is to determine whether there exists enough statistical evidence to conclude that $\sigma_i^2 > 0$ $(i = c, tc)$ for model 2.13. An appropriate decision can be made by (*i*) testing the hypothesis H_0: $\sigma_i^2 = 0$ *vs* H_a: $\sigma_i^2 > 0$; or by (*ii*) constructing a confidence interval about σ_i^2. The common tests of hypotheses about variance components include constructing *F*-statistics (*F*-distributions), Wald *Z*-test, and a likelihood ratio test (chi-square distributions). Furthermore, the confidence interval about σ_i^2 could be constructed based on the approach used for testing hypotheses. In general, techniques such as method-of-moments (MoMs), maximum likelihood (ML), restricted, or residual maximum likelihood (REML), and minimum norm quadratic unbiased estimation (MINQUE) could be used to obtain the estimates of the three variance components (center, interaction, and residuals).

The SAS-Mixed code in Table 2.1 was used to fit the two-way mixed model and obtain the REML estimates of the variance components. The estimates from the other three methods could be obtained by specifying Method = ML, MIVQUE0, or type3 (or type1, or type2).

Table 2.1. SAS-Mixed Code to Obtain the REML Estimates of the Variance Components.

```
Proc mixed data=data method=REML cl covtest;
    Class center treatment;
    Model outcomes= treatment /DDFM=KR;
    Random center center*treatment;
Run;
```

The method of moments procedure (Eisenhart, 1947), first computes the sums of squares (Type I, Type II, or Type III), determines their expectations, and then estimates the variance components from the system of linear equations. For instance, the Type III sums of squares, mean squares, and their corresponding expected mean squares for balanced data are shown in Table 2.2. Then, based on the analysis of the variance table, we construct *F*-statistics for the variance component hypothesis test.

When data is balanced (*e.g.* $r_{ij} = r$), the *F*-statistics are distributed exactly as *F*-distributions. In particular, the statistic for the first testing hypothesis (H_0: $\sigma_{tc}^2 = 0\ vs\ H_a$: $\sigma_{tc}^2 > 0$), constructed by using the expected mean squares in Table 2.2, is

$$F_{tc} = \frac{MST \times C}{MSE} \tag{2.16}$$

which has a sampling distribution of *F* with $df_{MST\times C} = (t-1)(b-1)$ for numerator degrees of freedom and $df_{MSE} = bt(r-1)$ for denominator degrees of freedom. If the $T \times C$ interaction variance component were equal to zero, it would indicate that the differences observed among the treatments are similar for all centers. Similarly, the statistic for the second testing hypothesis (H_0: $\sigma_c^2 = 0\ vs\ H_a$: $\sigma_c^2 > 0$) is

$$F_c = \frac{MSCenter}{MST \times C}$$

(2.17)

which has a sampling distribution of F with $df_{MSCenter} = (b-1)$ for numerator degrees of freedom and $df_{MST\times C} = (t-1)(b-1)$ for denominator degrees of freedom.

For unbalanced cases, using the analysis of the variance table to construct *F*-statistics for some variance component hypothesis tests are not quite straightforward. Unlike balanced case, the values of k_1, k_2, k_3, and k_4 which depend on data structure, need to be determined by Hartley's synthesis method (Hartley, 1967) at first. As shown in Table 2.3, the coefficients on σ_{tc}^2 in the expected mean squares of $MSCenter$ and $MST \times C$ are not the same (*e.g.* $k_2 \neq k_3$). Thus, there is no direct *F*-statistic to be tested for the hypothesis, H_0: $\sigma_c^2 = 0$ *vs* H_a: $\sigma_c^2 > 0$. However, there is a linear combination of mean squares that has the desired expectation, that is, $E\left[\frac{k_2}{k_3} MSTC + \left(1 - \frac{k_2}{k_3}\right) MSE\right] = k_2\sigma_{tc}^2 + \sigma_\varepsilon^2$. Then, the statistic to test this hypothesis is

$$F_c = \frac{MSCenter}{Q_2}$$

(2.18)

where $Q_2 = \frac{k_2}{k_3} MSTC + \left(1 - \frac{k_2}{k_3}\right) MSE$. Also, F_c has an approximate sampling distribution of F with $df_{MSC} = b - 1$ for numerator degrees of freedom and v_2 for denominator degrees of freedom where

$$v_2 = \frac{{Q_2}^2}{\frac{[(k_2/k_3)MSTC]^2}{df_{MSTC}} + \frac{[(1-(k_2/k_3))MSE]^2}{df_{MSE}}}$$

(2.19)

as determined by approximating the distribution of $v_2 Q_2 / E(Q_2)$ by a chi-square distribution using the Satterthwaite approximation (Giesbrecht & Burns, 1985).

When ML, REML, and MINQUE0 solutions are obtained for variance components, a Wald test is also available for these variance component tests. Since the Wald test is very approximate, it is suggested to be used only when there are many levels of the random effect (Dickey, 2008).

Table 2.2. Analysis of Variance Table for the Balanced Data Using Type III Sums of Squares

Source of Variance	*df*	Sum of Squares (SS)	Mean Square (MS)	Expected Mean Square	Error Term	Error *df*
T (Treatment)	*t-1*	$rb\sum_{i=1}^{t}(\bar{y}_{i..}-\bar{y}_{...})^2$	SST/df_T	$r\sigma_{tc}^2+\sigma_\varepsilon^2+\frac{rb}{t-1}\sum_i(\bar{T}_i-\bar{T})^2$	$MST\times C$	$df_{T\times C}$
C(Center)	*b-1*	$rt\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})^2$	SSC/df_C	$tr\sigma_c^2+r\sigma_{tc}^2+\sigma_\varepsilon^2$	$MST\times C$	$df_{T\times C}$
$T\times C$	*(b-1)(t-1)*	$r\sum_{i=1}^{t}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2$	$SST\times C/df_{T\times C}$	$r\sigma_{tc}^2+\sigma_\varepsilon^2$	MSE	df_{Error}
Residual (Error)	*bt (r-1)*	$\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{r}(\bar{y}_{ijk}-\bar{y}_{ij.})^2$	SSE/df_{Error}	σ_ε^2	-	-

Hypothesis	Test Statistics
H_0: $\sigma_{tc}^2=0$ *vs* H_a: $\sigma_{tc}^2>0$	$F_{tc}=MST\times C/MSE$
H_0: $\sigma_c^2=0$ *vs* H_a: $\sigma_c^2>0$	$F_c=MSCenter/MST\times C$
H_0: $\tau_i=0$ *vs* H_a: *at least one* $\tau_i\neq 0$	$F_t=MSTreatment/MST\times C$

Table 2.3. Analysis of Variance Table for the Unbalanced Data Using Type III Sums of Squares

Source of Variance	*df*	Sum of Squares (SS)	Mean Square (MS)	Expected Mean Square	Error Term[1]	Error *df*
T (Treatment)	*t-1*	$rb\sum_{i=1}^{t}(\bar{y}_{i..}-\bar{y}_{...})^2$	SST/df_T	$k_4\sigma_{tc}^2+\sigma_\varepsilon^2+\frac{rb}{t-1}\sum_i(\bar{T}_i-\bar{T})^2$	Q_1	v_1
C (Center)	*b-1*	$rt\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})^2$	SSC/df_C	$k_1\sigma_c^2+k_2\sigma_{tc}^2+\sigma_\varepsilon^2$	Q_2	v_2
$T\times C$	*(b-1) (t-1)*	$r\sum_{i=1}^{t}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2$	$SST\times C/df_{T\times C}$	$k_3\sigma_{tc}^2+\sigma_\varepsilon^2$	MSE	df_{Error}
Residual (Error)	*N-tb*	$\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{r_{ij}}(\bar{y}_{ijk}-\bar{y}_{ij.})^2$	SSE/df_{Error}	σ_ε^2	-	-

Hypothesis	Test Statistics
$H_0: \sigma_{tc}^2=0\ vs\ H_a: \sigma_{tc}^2>0$	$F_{tc}=MST\times C/MSE$
$H_0: \sigma_c^2=0\ vs\ H_a: \sigma_c^2>0$	$F_c=MSCenter/Q_1$
$H_0: \tau_i=0\ vs\ H_a: at\ least\ one\ \tau_i\neq 0$	$F_t=MSTreatment/Q_2$

[1] $Q_1=\frac{k_4}{k_3}MSTC+(1-\frac{k_4}{k_3})MSE$, with an associated degree of freedom v_1, and $Q_2=\frac{k_2}{k_3}MSTC+(1-\frac{k_2}{k_3})MSE$, with an associated degree of freedom v_2, where v_i $(i=1,2)$ is determined by approximating the distribution of $v_iQ_i/E(Q_i)$ by a chi-square distribution using the Satterthwaite approximation.

Another alternative approach for testing hypotheses about variance components could be based on a likelihood ratio procedure (not shown here since it is not our main interest), which involves evaluating the value of the likelihood function for the complete model and evaluating the value of the likelihood function for the model under the conditions of H_0 (reduced model).

Constructing confidence intervals of variance components relies on the type of approximation. A $(1-\alpha)$ 100% confidence interval about σ_i^2 $(i = c, tc, \varepsilon)$ is given as

1) Wald Type Confidence Intervals

$$\hat{\sigma}_i^2 - Z_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2_{\hat{\sigma}_i^2}} \leq \sigma_i^2 \leq \hat{\sigma}_i^2 + Z_{\alpha/2}\sqrt{\hat{\sigma}^2_{\hat{\sigma}_i^2}} \tag{2.20}$$

which can be computed using the asymptotic normality of maximum likelihood estimates. As mentioned above, the Wald confidence intervals are appropriate only when the degrees of freedom associated with the estimated variance component are large (*e.g.,* large size of random effect levels).

2) Satterthwaite Type Confidence Intervals

$$\frac{\nu_i \hat{\sigma}_i^2}{\chi^2_{\left(\frac{\alpha}{2}\right),\nu_i}} \leq \sigma_i^2 \leq \frac{\nu_i \hat{\sigma}_i^2}{\chi^2_{\left(1-\frac{\alpha}{2}\right),\nu_i}} \tag{2.21}$$

where $\hat{\sigma}_i^2$ is the estimate of σ_i^2, ν_i is the degrees of freedom associated with $\hat{\sigma}_i^2$, and $\chi^2_{\left(\frac{\alpha}{2}\right),\nu_i}$, and $\chi^2_{\left(1-\frac{\alpha}{2}\right),\nu_i}$ denote lower and upper $\alpha/2$ percentage points from a chi-square distribution with ν_i degrees of freedom, which are determined using the Satterthwaite approximation.

$$\nu_i = \frac{2[E(\hat{\sigma}_i^2)]^2}{Var(\hat{\sigma}_i^2)} \tag{2.22}$$

If $\hat{\sigma}_i^2$ is obtained from methods of moments solution, the approximate sampling distribution of $\hat{\sigma}_i^2 = q_1 MS_1 + q_2 MS_2 + \cdots + q_k MS_k$, where MS_i denotes a mean square based on f_i degrees of freedom, the mean squares are independently distributed, and the q_i are known constants. Then, Satterthwaite approximation degrees of freedom could be written as

$$\nu_i = \frac{2(\hat{\sigma}_i^2)^2}{\sum_{i=1}^{k} \frac{(q_i MS_i)^2}{f_i}}$$

(2.23)

For instance, in Table 2.3, $\hat{\sigma}_{tc}^2 = \frac{1}{r} MST \times C - \frac{1}{r} MSE$. The number of degrees of freedom of the approximating chi-square distribution obtained through the Satterthwaite approximation is

$$\nu_{tc} = \frac{2[\hat{\sigma}_{tc}^2]^2}{\frac{\left(\frac{1}{r} MST \times C\right)^2}{(b-1)(t-1)} + \frac{\left(\frac{1}{r} MSE\right)^2}{bt(r-1)}}$$

(2.24)

If $\hat{\sigma}_i^2$ is obtained from the REML, ML, and MIVQUE solutions, then the inverse of the information matrix can be used to estimate the variances of the estimated variance components. Then, the number of degrees of freedom by Satterthwaite approximation is given by

$$\nu_i = \frac{2[E(\hat{\sigma}_i^2)]^2}{Var(\hat{\sigma}_i^2)} = 2(Z\ value)^2$$

(2.25)

In general, the method of moments technique enables the derivation of estimators without requiring the assumption of normality. Consequently, the singular established characteristic of these estimators is their lack of bias. However, a challenge arises when specific variance component estimates are negative. In such instances, the variance component estimator is adjusted to zero to maintain it within the parameter space. Unfortunately, this practice of setting the estimator to zero in response to negative solutions compromises its unbiasedness. In cases where

the design is balanced and along with all positive solutions for variance components, the method of moments, REML, and MIVQUE estimators are identical (Milliken & Johnson, 2009). In the case of unbalanced designs, method-of-moment estimates are computationally more straightforward, while the other three methods necessitate iterative algorithms. Both maximum likelihood and REML estimators demonstrate consistency and inherit the standard properties associated with large-sample-size maximum likelihood estimates. Generally, REML is the preferred method of estimating the variance components.

2.1.3.2 Analysis of the Fixed Effects

Eventually, our main interest is to determine if there is enough statistical evidence to conclude that $\beta \neq 0$ (*e.g.,* treatment effect exists) for model 2.13.

a) Estimation and Construction of Confidence Intervals

For the estimation of fixed effect, there are several methods for estimating estimable functions of β (*i.e.* $a'\beta$) in the mixed model. If the elements of the covariance matrix Σ are known, then the best linear unbiased estimator (BLUE) of an estimable function $a'\beta$ is

$$a'\hat{\beta}_{BLUE} = a'(X'\Sigma^{-1}X)^{-}X'\Sigma^{-1}y \tag{2.26}$$

For most balanced designs and some simple unbalanced designs, the estimator of β could be simplified to $\hat{\beta}_{BLUE} = \hat{\beta}_{LS} = (X'X)^{-}X'y$, which is the ordinary least squares estimator of β, and it also does not depend on the variance components (Milliken & Johnson, 2009). That is,

$$a'\hat{\beta}_{BLUE} = a'(X'X)^{-}X'y \tag{2.27}$$

When the designs are unbalanced, and the covariance matrix Σ is unknown, the $\hat{\beta}_{BLUE}$ does not exist. A weighted least squares estimator must be obtained where $\hat{\Sigma}$ is used as the weighting matrix. The estimated covariance matrix is given by

$$\hat{\Sigma} = \hat{\sigma}_c^2 \, Z_1 Z_1' + \, \hat{\sigma}_{tc}^2 Z_2 Z_2' + \hat{\sigma}_\varepsilon^2 I_N$$

(2.28)

where $\hat{\sigma}_c^2$, $\hat{\sigma}_{tc}^2$, and $\hat{\sigma}_\varepsilon^2$ are the estimators of the variance components obtained using one of the previous methods (MoMs, ML, REML, or MINQUE0). And then, the weighted least squares estimator of $a'\beta$ is as followed

$$a'\hat{\beta}_W = a'(X'\hat{\Sigma}^{-1}X)^{-}X'\hat{\Sigma}^{-1}y$$

(2.29)

with Var $(a'\hat{\beta}_W) = a'(X'\hat{\Sigma}^{-1}X)^{-}a$, where a generalized inverse (g-inverse) solution is used for inverse of $X'\hat{\Sigma}^{-1}X$. Thus, a (1- a)100% confidence interval for an estimable function of $a'\beta$ could be obtained by using the asymptotic sampling distribution of $a'\hat{\beta}_W$. That is,

$$a'\hat{\beta}_W \sim N[a'\beta, a'(X'\hat{\Sigma}^{-1}X)^{-}a]$$

However, Kackar and Harville (1984), and Kenward and Roger (1997) found that $a'(X'\hat{\Sigma}^{-1}X)^{-}a$ is too small for small sample inference from REML procedure, and then provided an adjustment to the estimated standard errors of the fixed effects using a Taylor series expansion about the unknown variance components. The *SAS-Mixed* procedure with the *DDFM = KR* option directly provides the adjusted estimated standard error for the estimates of the fixed effects parameters, along with adjusted corresponding degrees by using the generalization of the Satterthwaite approximation as

$$\hat{v} = \frac{2[a'(X'\hat{\Sigma}^{-1}X)^{-}a]^2}{\widehat{Var}[a'(X'\hat{\Sigma}^{-1}X)^{-}a]}$$

(2.30)

Therefore, an approximate (1- a)100% confidence interval of $a'\beta$ is given by

$$a'\hat{\beta}_W \pm (t_{a/2,\hat{v}})\sqrt{[a'(X'\hat{\Sigma}^{-1}X)^{-}a]}$$

(2.31)

b) Testing Hypotheses

To test $H_0: H\beta = b$ vs $H_a: H\beta \neq b$, first, we need to compute the statistic as

$$Q = \left(H\hat{\beta}_W - b\right)\left[H\left(X'\hat{\Sigma}^{-1}X\right)^{-}H'\right]^{-1}\left(H\hat{\beta}_W - b\right) \tag{2.32}$$

Under the conditions of the null hypothesis, the asymptotic sampling distribution of Q is χ_q^2 where q = Rank(H) is the number of linear independent linear combinations of β in H. A small sample test statistic is $F_c = Q/r$ where r = Rank $[H\left(X'\hat{\Sigma}^{-1}X\right)^{-}H']$, which has an approximate sampling distribution of F with $df_{\text{numerator}} = q$ for numerator degrees of freedom and $df_{\text{denominator}} = r$ for denominator degrees of freedom. The denominator degrees of freedom (DDF) could be computed using Satterthwaite approximation under the *DDFM = KR* option. Small p-values (typically less than 0.05) indicate a significant effect.

For instance, to test the hypothesis that the means are equal, $H_0: \mu_1 = \cdots = \mu_i \; vs \; H_a: not \; (H_0)$, it could construct F-statistics using the expected mean squares based on the analysis of variance table, when data is balanced shown in Table 2.2,

$$F_t = \frac{MSTreatment}{MST \times C} \tag{2.33}$$

which has a sampling distribution of F with $df_{MSTreatment} = (t-1)$ numerator degrees of freedom and $df_{MST\times C} = (t-1)(b-1)$ denominator degrees of freedom.

For unbalanced data (*e.g.* $r_{ij} \neq r$), since the coefficients on σ_{tc}^2 in the expected mean squares of $MSTreatment$ and $MST \times C$ are not the same (*e.g.* $k_4 \neq k_3$), there is no F-statistic to be tested that the means are equal. However, there is a linear combination of mean squares that has the desired expectation, that is, $E\left[\frac{k_4}{k_3}MSTC + \left(1 - \frac{k_4}{k_3}\right)MSE\right] = k_4\sigma_{tc}^2 + \sigma_\varepsilon^2$. The test statistic is

$$F_t = \frac{MSTreatment}{Q_1}$$

(2.34)

where $Q_1 = \frac{k_4}{k_3} MSTC + \left(1 - \frac{k_4}{k_3}\right) MSE$, which has an approximate sampling distribution of F with df_{MSC} numerator degrees of freedom and v denominator degrees of freedom, as determined by a Satterthwaite approximation, that is,

$$v_1 = \frac{Q_1^2}{\frac{[(k_4/k_3)MSTC]^2}{df_{MSTC}} + \frac{[(1-(k_4/k_3))MSE]^2}{df_{MSE}}}$$

(2.35)

Here is a concern, if the $\sigma_{tc}^2 \neq 0$, this should be the correct test for treatment effects (George Casella, 2006). However, if $\sigma_{tc}^2 = 0$ (the treatment effect is homogeneous across centers), the reduced model could generate greater power and improve the efficiency of the test for the main effects (shown in Table 2.4 and Figure. 2.1).

If REML, ML, and MIVQUE solutions are obtained, then F statistic construct as part b) with recomputed degrees of freedom $df = 2(Z\ value)^2$.

Table 2.4. Analysis of Variance Table for the balanced data using Type III sums of squares when $\boldsymbol{\sigma_{tc}^2}$=0.

Source of Variance	*df*	Sum of Squares	Expected Mean Square
Treatment	*t-1*	$rb\sum_{i=1}^{t}(\bar{y}_{i..} - \bar{y}_{...})^2$	$\sigma_\varepsilon^2 + \frac{rb}{t-1}\sum_i(\bar{T}_i - \bar{T})^2$
Center	*b-1*	$rt\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2$	$tr\sigma_c^2 + \sigma_\varepsilon^2$
Residual (Error)		$\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{r}(\bar{y}_{ijk} - \bar{y}_{ij.})^2$	σ_ε^2

Hypothesis	Test Statistics
$H_0: \sigma_c^2 = 0\ vs\ H_a: \sigma_c^2 > 0$	$F_c = MSCenter/MSE$
$H_0: \tau_i = 0\ vs\ H_a: at\ least\ one\ \tau_i \neq 0$	$F_t = MSTreatment/MSE$

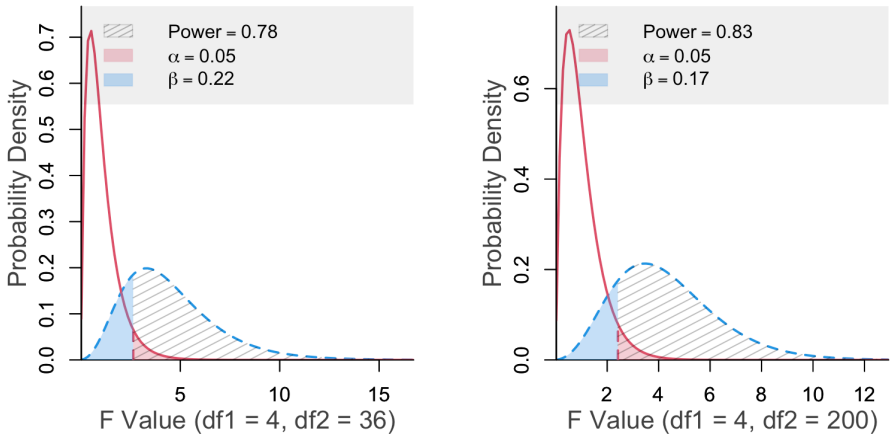


Figure 2.1. Comparison of power for the balanced data using Type III sums of squares when $\boldsymbol{\sigma_{tc}^2}$=0.

2.1.4 Preliminary Simulation Study

2.1.4.1 Correlated Outcomes

Our first preliminary simulation study investigated the correlation between two treatment groups under different center effect sizes (Figure 2.2). The number of each center was 20, with 10 randomly assigned to treatment group 1 and the other 10 to treatment group 2. We used 2000 simulations for each scenario to estimate the correlation between the two treatment groups. Figure 2.2 shows the results of 1000 simulations, which we chose to maintain the clarity of the graph. We simulated a continuous response Y_{ijk} under model 2.13 with homogeneous scenario ($\sigma_{tc}^2 = 0$).

Figure 2.2 shows that when center effects are smaller than 0.3, the correlation between $\bar{Y}_1$ and $\bar{Y}_2$ are harder to notice. An unadjusted analysis under these scenarios could give valid results. However, as center effects increased, a correlation between $\bar{Y}_1$ and $\bar{Y}_2$ increased sharply. This demonstrates that the correlation is introduced by the process of randomization.

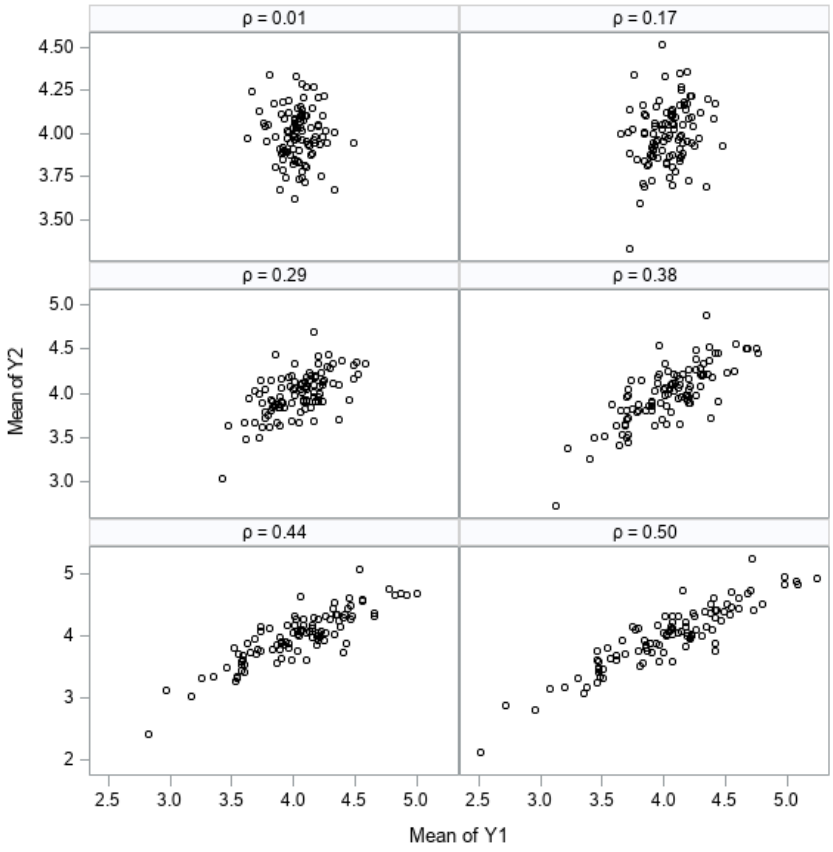


Figure 2.2. Correlation between outcomes from two treatment groups under different center effect sizes.

2.1.4.2 Model Comparisons

a) Data Generating

In the second preliminary study, we used Monte Carlo simulation to assess the performance of six statistical models to analyze a multicenter study, only considering the center effect with a continuous outcome under homogeneity and heterogeneity when the center level is small (n_c = 4). We generated continuous outcomes, *Y*, using the mixed-effects linear regression model given by Eq. 2.13. We generated the random error, ε_{ijk}, from *N* (0, $\sigma_\varepsilon^2 = 0.5$). To simulate center effects, we employed the relationship between ρ_c and σ_c^2:

For homogeneity cases ($\sigma_{tc}^2 = 0$), as Eq. 2.2

$$\rho_c = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_\varepsilon^2}$$

(2.36)

For heterogeneity cases ($\sigma_{tc}^2 \geq 0$), as Eq. 2.3

$$\rho_c = \frac{\sigma_c^2 - \sigma_{tc}^2/(k-1)}{\sigma_c^2 + \sigma_{tc}^2 + \sigma_\varepsilon^2}$$

(2.37)

To comprehensively study the behavior of candidate models at various ρ_c levels under both cases, we considered the following values of ρ_c for completeness: 0.00, 0.25, 0.50, 0.75 and 0.99. This in turn set the corresponding σ_c^2 values that are to be shown in Table 2.5. However, we focused interpretation of the results on lower values of ICC and σ_{tc}^2 as they were more likely to occur in practice.

Table 2.5. The catalog of simulation designs

Scenario	σ_{tc}^2	τ	ρ_c	σ_c^2
1 - 25	0	0, 0.4, 0.8, 1.2, 1.6	0, 0.25, 0.5, 0.75, 0.99	0, 0.167, 0.5, 1.5, 49.5
26 - 50	0.1			0.1, 0.3, 0.7, 1.9, 59.5
51 - 75	0.2			0.2, 0.25, 0.9, 2.3, 69.5
75 - 100	0.3			0.3, 0.30, 1.1, 2.7, 79.5

b) Performance Measures

For each scenario, we generated 1000 hypothetical trial datasets. Then, we applied six statistical models (simple linear regression, Random-effects, Fixed-effects without interaction, Fixed-effects with interaction, Fixed-effects with correction, and Random-effects with correction shown in Table 2.6) using MoMs (Type III sum of square) or REML with or without KR option to each simulated dataset. For each simulation scenario, model and method, we estimated the following measures:

For center-related effects, only Type I error rate and Power.

For treatment effect,

1) Type I error rate (when the true $\tau = 0$) and Power (when the true $\tau > 0$)
2) Average Point estimated τ and estimated standard error (SE) for τ, with simulation standard deviation (SD) for $\hat{\tau}$
3) Bias and mean squared error (MSE) of the point estimator of treatment effect (*i.e.,* τ)
4) The empirical coverage rate of the 95% confidence intervals (CIs) around τ.

Type I error rate and power were calculated as the proportion of the simulation results with a statistically significant center/treatment effect with a two-sided significance level of 5%. The mean value of the estimated treatment effect was calculated as τ, and the simulation or empirical SD was calculated as the standard deviation of the estimated τ across simulations, indicating the precision of the estimator. Bias is estimated as the difference between the average estimate of τ over 1000 simulated datasets and the true effect and is given by

$$Bias\ (\hat{\tau}) = E\ (\hat{\tau}) - \tau = E\ (\hat{\tau} - \tau).$$

Negatively or positively biased estimators lead to an under- or over-estimation of the true τ. Therefore, if an estimator for τ satisfies $E\ (\hat{\tau}) = \tau$, that is, it is unbiased. A good estimator should not only be unbiased but also remain unaffected as much as possible by sampling fluctuation. The overall error rate of the point estimator was captured by estimating by MSE, which is the average squared distance between the estimator and its true value across the 1000 datasets:

$$MSE\ (\hat{\tau}) = E\ (\hat{\tau} - \tau)^2 = Var\ (\hat{\tau}) + (Bias\ (\hat{\tau}))^2.$$

If $MSE\ (\hat{\tau}_1) < MSE\ (\hat{\tau}_2)$, then $\hat{\tau}_1$ is said to be more efficient than $\hat{\tau}_2$. Furthermore, we reported the performance of 95% CIs in different methods for τ in terms of coverage rate (*i.e.,* the proportion of the simulation results in which the estimated 95% CIs contained the true value of the τ). All datasets were simulated and analyzed in SAS software (version 9.4).

Table 2.6. Descriptions of Six Models

Item	Model[1]	Note
Unadjusted	$Y_{ijk} = \mu + \tau_i + \varepsilon_{ijk}$	Only τ_i, as fixed
Random-effects	$Y_{ijk} = \mu + \tau_i + b_j + \varepsilon_{ijk}$	b_j is random
Fixed-effects	$Y_{ijk} = \mu + \tau_i + b_j + \varepsilon_{ijk}$	b_j is fixed
Fixed-effects with interaction	$Y_{ijk} = \mu + \tau_i + b_j + (\tau b)_{ij} + \varepsilon_{ijk}$	b_j and $(tb)_{ij}$ are fixed
Random-effects with correction	$Y_{ijk} = \mu + \tau_i + b_j + (\tau b)_{ij} + \varepsilon_{ijk}$	b_j and $(tb)_{ij}$ are random
Fixed-effects with correction	$Y_{ijk} = \mu + \tau_i + b_j + (\tau b)_{ij} + \varepsilon_{ijk}$	b_j is fixed and $(tb)_{ij}$ is random

[1]where Y_{ijk} represents a continuous outcome measured for the *k*th subject in the *j*th center receiving the *i*th treatment (control or treatment group).

c) Results and Discussion

Here, to test whether the treatment effect is homogeneous across treatment groups, we first explored the Type I error rate and power analysis for the interaction effect. There are three models accounting for the treatment-by-center interaction effect, including Fixed-effects with interaction (τ_{tc}, fixed), Random-effects with correction (σ_{tc}^2, random) and Fixed-effects with correction (σ_{tc}^2, random). We compared these three models' performance for the hypothesis test of $\sigma_{tc}^2 = 0$ or $\tau_{tc} = 0$ in terms of Type I error and power (shown in Figure 2.3 and Figure 2.4). As the results shown from the type III sums of squares analysis, whatever center effect specified as a fixed or random term in the model, these three models performed well in terms of the nominal Type I error level (~ 5%) and great power (~80% for $\sigma_{tc}^2 = 0.3$). On the contrary, the results from REML showed that the test could lead to a severe reduction in type I error rate and power except when the interaction effect is fixed. In this simulation study, the interaction effects were generated at small value ($\sigma_{tc}^2 \leq 0.3$). Thus, the Type III estimates could turn out to be negative. As we know, it needs to be adjusted to zero for a compromise of its unbiasedness. With iterative algorithms, the REML estimator was as stated to be better than the Type III estimator for point estimation purposes (Milliken & Johnson, 2009). Unlike estimation, *F*-statistics for Type III was constructed to test the hypothesis by a ratio of two mean squares. No matter what the solution value for σ_{tc}^2 (negative or

nonnegative), these expected mean squares are unbiased estimators for certain mean squares. However, the REML approach relies on the Wald *Z*-test, which is asymptotically distributed as a normal random variable. In these cases, the number of centers associated with each of the σ_{tc}^2 are small. They were also, owing to REML's restriction that $\sigma_{tc}^2 \geq 0$, it is unsurprising that this resulted in a highly skewed distribution, which is not valid for a Wald Z-test. Unless the center size is large enough, the information associated with the Wald *Z*-test does not help test hypotheses. When the number of centers is small, the more appropriate way is to incorporate τ_{tc} as a fixed effect in the model (*i.e.* Fixed-effects with interaction). As expected, no significant difference is observed in Type I error rate and power among either different treatment effects (τ_t) groups or center effects (ICC values) groups.

The second step is to test the hypothesis of $\sigma_c^2 = 0$ or $\tau_c = 0$ (*i.e.,* center effects), only the unadjusted model did not consider the center effect in the model. Here, we compared the rest of the five models' performance (τ_c: as fixed in Fixed-effects, Fixed-effects with interaction, and Fixed-effects with correction; or σ_c^2: as random in Random-effects, and Random-effects with correction) in terms of Type I error and power (shown in Figure 2.5 and Figure 2.6). For the type III sums of squares analysis, the results remained consistent as the previous proof in part 2.1.3. That is, when $\sigma_{tc}^2 = 0$, all the models using Type II could achieve the nominal Type I error rate. However, since the F_c statistic for the situation should be $\frac{MSCenter}{MSE}$, instead of $\frac{MSCenter}{MST \times C}$, models without random $T \times C$ term (e.g., Fixed-effects, Fixed-effects with interaction and Random-effects) provided greater power than other models (e.g., Fixed-effects with correction and Random-effects with correction). Furthermore, when $\sigma_{tc}^2 > 0$, only $\frac{MSCenter}{MST \times C}$ computed the correct F_c statistic for the hypothesis test of $\sigma_c^2 = 0$ or $\tau_c = 0$, so models without $T \times C$ random term caused an inflated Type I error rate. As σ_{tc}^2 value increased, this inflation increased markedly. This

indicates when heterogeneity exists (nonzero σ_{tc}^2), $MST \times C$ should be the correct term to test the center effect to validate inference.

For REML procedure, also since the number of centers is small, models including center as a random term (*e.g.* Random-effect and Random-effect with correction) are not suitable, and it caused a deficient type I error rate and loss of power. When incorporating the center as a fixed in the model, but without $T \times C$ random term, it could cause the similar issues (an inflated Type I error rate with falsely high power) as Type III approach for nonzero σ_{tc}^2 situation. For this scenario (σ_{tc}^2 has existed), only Fixed-effects with correction, along with Satterthwaite approximation adjustment (In *SAS-Mixed* procedure with the *DDFM = KR* option) could achieve the nominal Type I error rate and provide true power for hypothesis that no center effect existed. As we discussed before the main reason is, as we discussed before, that for small sample inference of fixed center effect from REML procedure, the corresponding estimated standard errors (*i.e.* $a'(X'\hat{\Sigma}^{-1}X)^{-}a$) is too small without adjustment.

The analysis for fixed treatment effect was included two parts, point estimation and hypothesis testing. The point estimates of treatment effect τ were unbiased in six models for either Type III sum of squares or REML procedures across all ICC values (Figure 2.11 - 2.16). Upon review, it was unexpected to observe that the point estimates in the unadjusted model (simple regression model), neglecting the center, remained unchanged regardless of ICC values. When treatments are distributed proportionally across all centers, the center exhibits no association with the treatment allocation. Consequently, whether accounting for the center effect in the model or not has little impact on the point estimate of the treatment-response relationship, especially when dealing with a continuous response variable. Therefore, various methods of integrating between-center information resulted in identical estimates of treatment contrast in a balanced design. These

consistent point estimates are further translated into uniform empirical standard deviation and the estimator's overall error rates (measured by MSE) across all six models regardless of ICC values, except for fixed effects with the interaction model (Higher simulation SD and MSE). When heterogeneous exist, empirical SD and MSE of Fixed-effects with interaction model increased dramatically as σ_{tb}^2 value increased, while other models were changed slightly. Across different ICC values, Fixed-effects and Random-effects Models with or without correction yielded the smallest average estimates of the standard error of τ, while Fixed-effects with the interaction model led to a larger estimated SE of τ. As expected, the unadjusted analyses produced larger average estimated SE values as larger ICC values.

Type I error rate and power analysis for treatment effect τ were shown in Figure 2.7 - 2.10. Under homogeneous scenarios (σ_{tb}^2 = 0), results from the type III sums of squares analysis show that all the models could hold the nominal Type I error rate (~ 5%) regardless of ICC values, except the unadjusted model. Ignoring the center effect (unadjusted model), the Type I error rate was decreasing dramatically as ICC values increased, which was in agreement with the finding by Parzen et al. (1998). Meanwhile, models incorporating treatment-by-center term could provide less efficiency (power < 80% for τ = 0.4). As it was discussed before, for most balanced designs, the point estimator of β could be simplified to $\hat{\beta}_{BLUE} = \hat{\beta}_{LS} = (X'X)^{-}X'y$, which does not depend on the variance components. REML estimator of fixed effect β were the same as Type III sum of squares estimators. However, the hypothesis test and 95% confidence interval for fixed effect rely on estimating variance components and types of statistics for hypothesis testing. *F*-statistics for Type III sum of squares, constructed to test hypothesis by a ratio of two mean squares, always have good properties. For the REML procedure, the estimated standard errors (*i.e.* $a'(X'\hat{\Sigma}^{-1}X)^{-}a$) for small sample inference of fixed treatment effect is too small without adjustment. It could further

cause a deficient Type I error rate and loss of power (shown in Fig 2.6 and 2.7). For this situation, Kackar and Harville (1984) and Kenward and Roger (1997) have provided an adjustment to the estimated standard errors of the fixed effects. Also, it needs to assess the degrees of freedom to associate with a standard error by using the generalization of the Satterthwaite approximation (Geisbrecht & Burns, 1985). In the *SAS-Mixed* procedure, REML using the *DDFM = KR* option could provide better results than without adjustment in lifting Type I error rate to nominal level and more efficient power.

Under heterogeneous scenarios ($\sigma_{tb}^2 > 0$), no matter the type III sums of squares or REML applied, fixed-effects or random-effects model without correction led to inflated type I error rates and falsely high power. Furthermore, this issue got worse as σ_{tb}^2 value got larger. No significant difference was observed in type I error rates across ICC levels for all models except the simple regression model.

In this study, datasets were generated that the treatment effects were homogeneous and heterogeneous among centers, both fixed-effects and random-effects with correction models gave great coverage rate. In contrast, other models failed to get close to the nominal level, especially with large σ_{tb}^2 values. The coverage rates were not affected by ICC, except for the unadjusted model.

d) Conclusion

We used simulations to investigate the performance of six statistical approaches advocated to analyze continuous outcomes in multicenter RCTs only considering the center factor. Our simulation study demonstrated that it is crucial to investigate the random effect initially as a prerequisite for the final analysis of the fixed effect. In summary, when the number of centers is small, hypothesis testing for random effect using Type III sum of squares provided valid and

reliable results if the under model is not mis-specified (e.g., not missing nonzero interaction effect). REML is not a good option if the model includes center or treatment-by-center interaction as a random effect. When treatment-by-center interaction effect was significantly present, including a random interaction term is essential. The absence of a random interaction term can lead to a failure to account for those potentials adequately. Moreover, it potentially leads to incorrect inferences and hypothesis tests.

All six models produced unbiased estimates of treatment effect in all the scenarios. Ignoring the center as fixed or random effects resulted in a less efficient treatment effect hypothesis test in all scenarios as ICC values increased. When REML is applied in RCTs with a small number of centers, the KR adjustment approach can effectively address low Type I error rates and low power concerns. Fixed-effects with an interaction model is an alternative way to adjust for the center (when the number of centers is small). However, it could be less efficient than the fixed/random-effects model including random interaction effects under most circumstances.

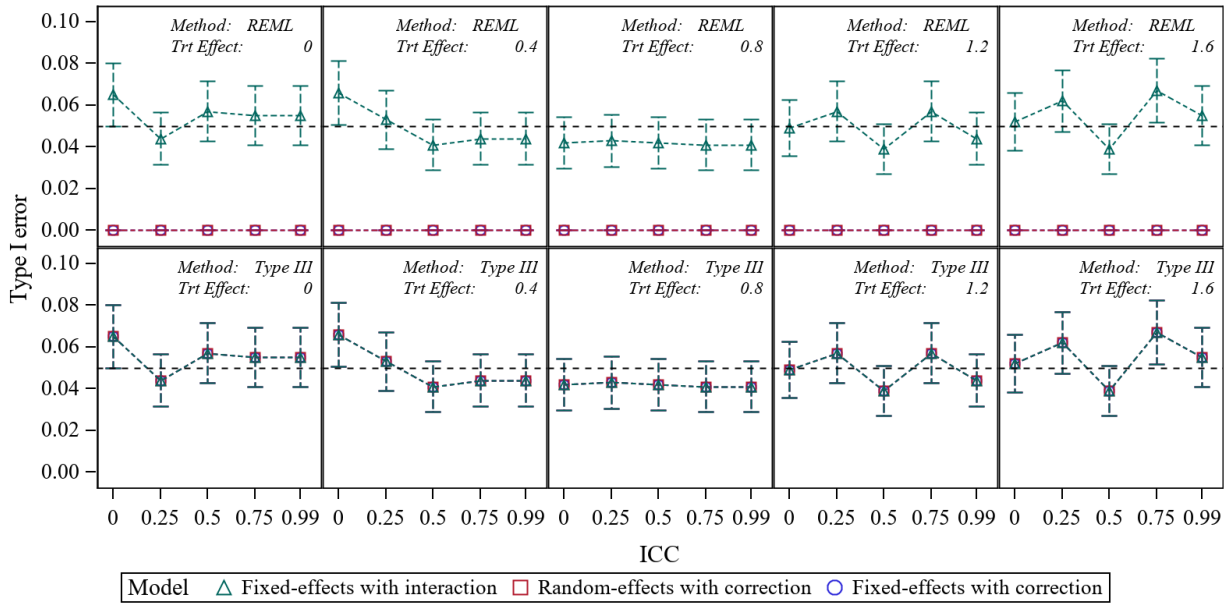


Figure 2.3. Type I error rate for hypothesis test of $\boldsymbol{\sigma_{tc}^2 = 0}$ or $\boldsymbol{\tau_{tc} = 0}$ under different scenarios for 'appropriate' and 'inappropriate' methods of analysis.

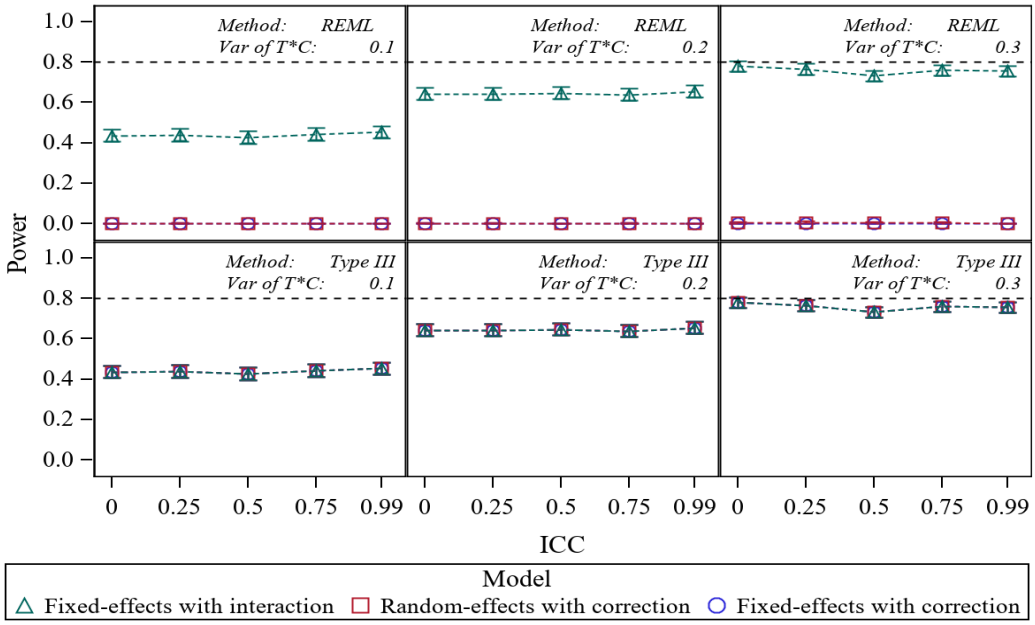


Figure 2.4. Power for hypothesis test of $\boldsymbol{\sigma_{tc}^2 = 0}$ or $\boldsymbol{\tau_{tc} = 0}$ under different scenarios for 'appropriate' and 'inappropriate' methods of analysis.

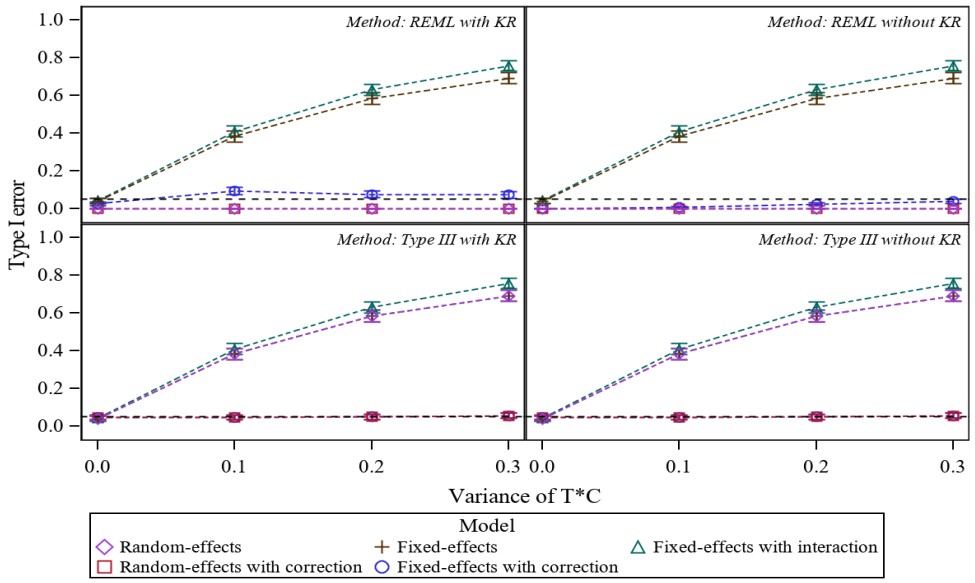


Figure 2.5. Type I error rate for hypothesis test of $\boldsymbol{\sigma_c^2 = 0}$ or $\boldsymbol{\tau_c = 0}$ under different scenarios for 'appropriate' and 'inappropriate' methods of analysis.

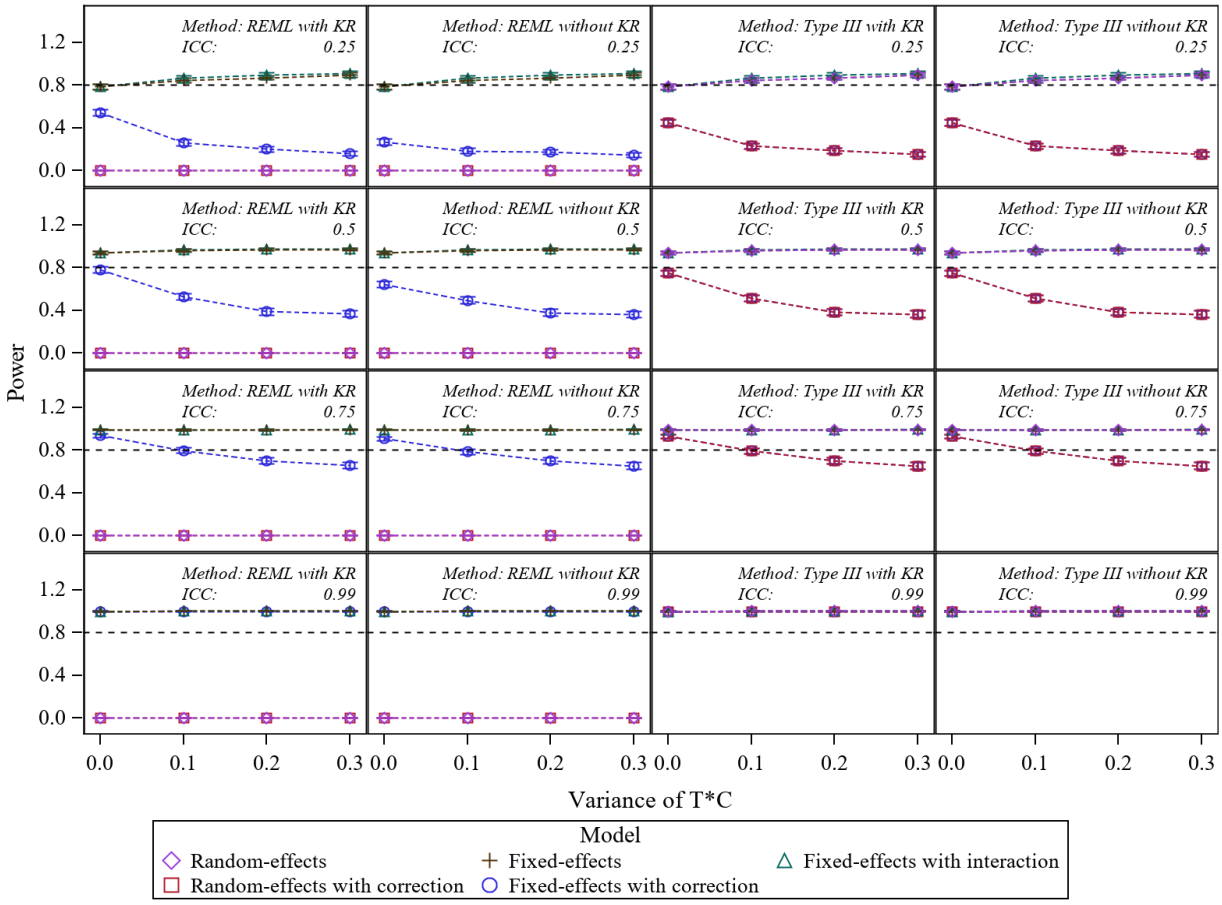


Figure 2.6. Power for hypothesis test of $\boldsymbol{\sigma_c^2 = 0}$ or $\boldsymbol{\tau_c = 0}$ under different scenarios for 'appropriate' and 'inappropriate' methods of analysis.

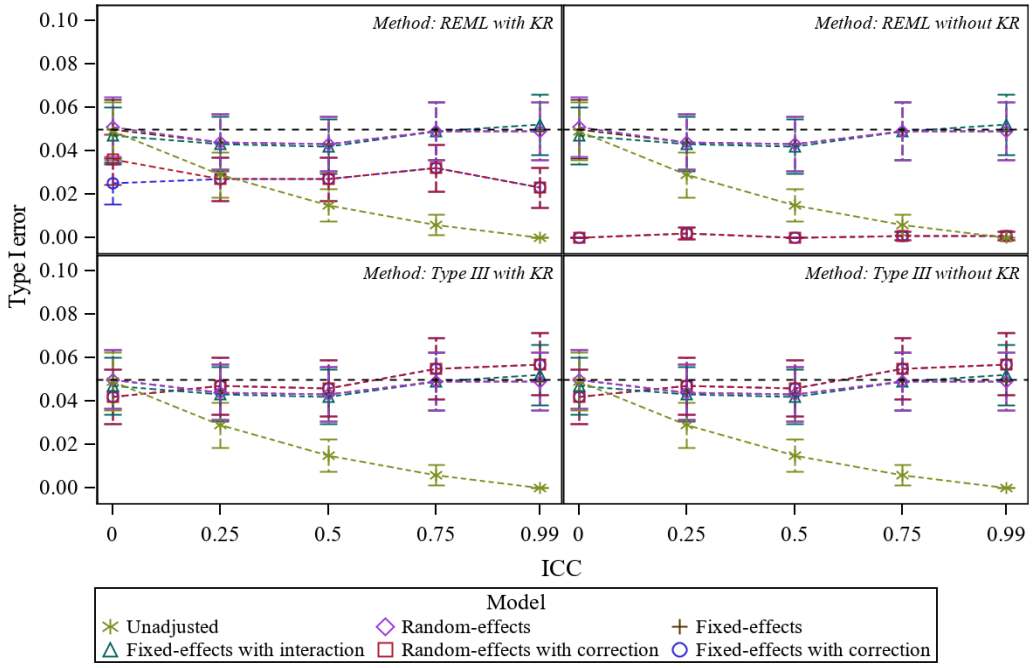


Figure 2.7. Type I error rate for hypothesis test of $\boldsymbol{\tau} = 0$ when treatment effect is homogeneous across treatment groups ($\boldsymbol{\sigma_{tc}^2} = 0$) under different scenarios for 'appropriate' and 'inappropriate' methods of analysis.

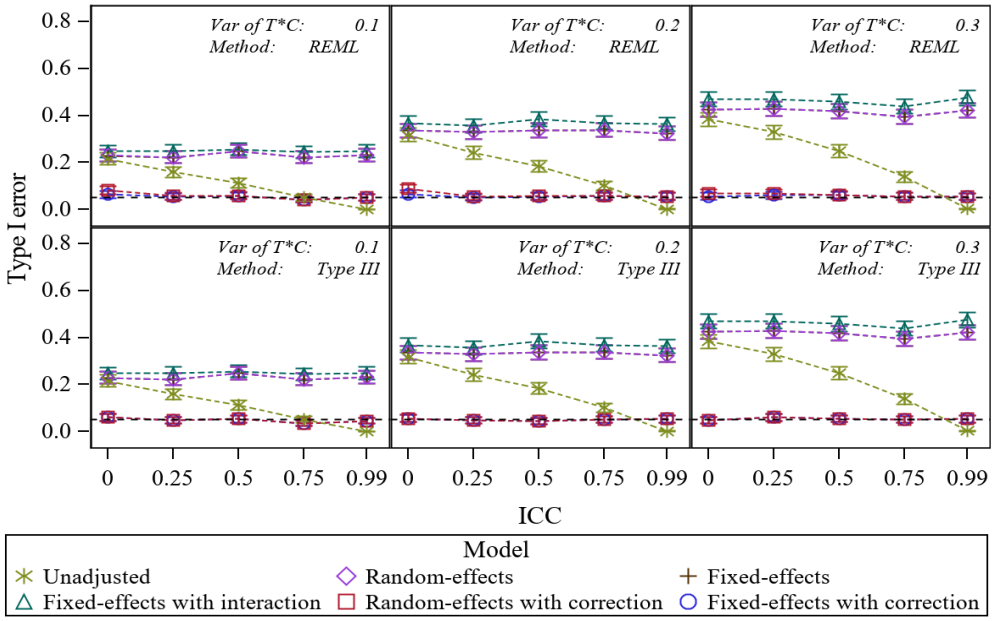


Figure 2.8. Type I error rate for hypothesis test of $\boldsymbol{\tau} = 0$ when treatment effect is heterogeneous across treatment groups ($\boldsymbol{\sigma_{tc}^2} > 0$) under different scenarios for 'appropriate' and 'inappropriate' methods of analysis.

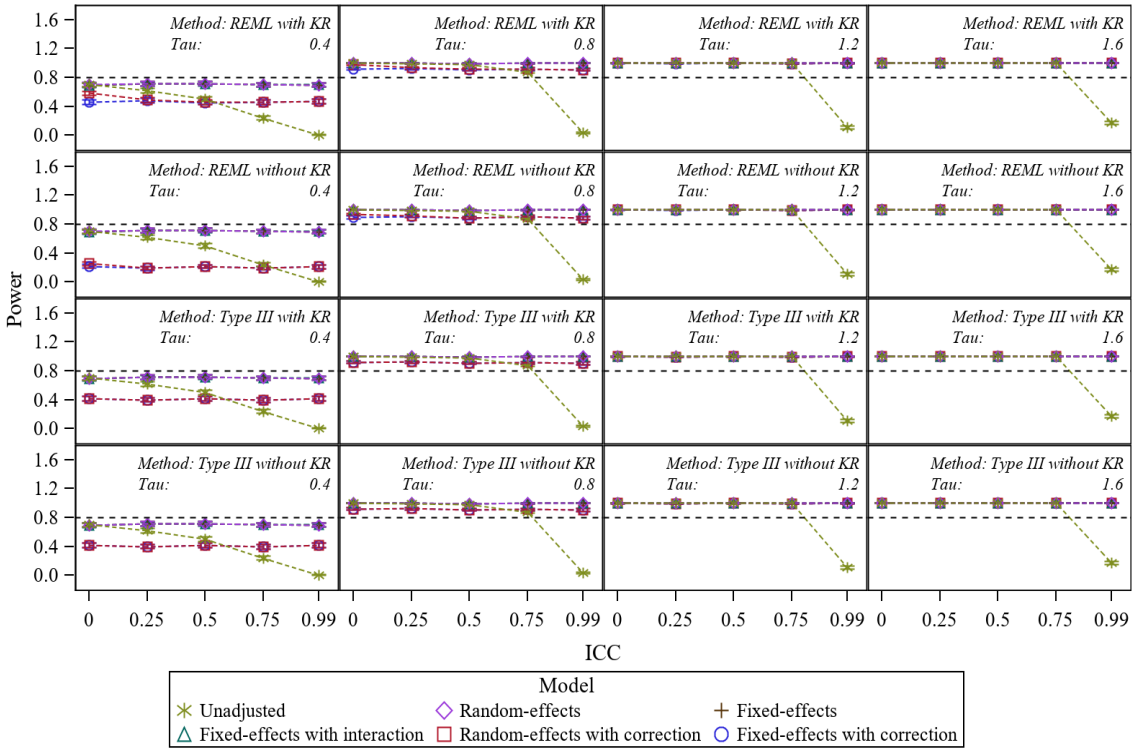


Figure 2.9. Empirical power for hypothesis test of $\boldsymbol{\tau} = 0$ when treatment effect is homogeneous across treatment groups ($\boldsymbol{\sigma^2_{tc}} = 0$) under different scenarios for 'appropriate' and 'inappropriate' methods of analysis.

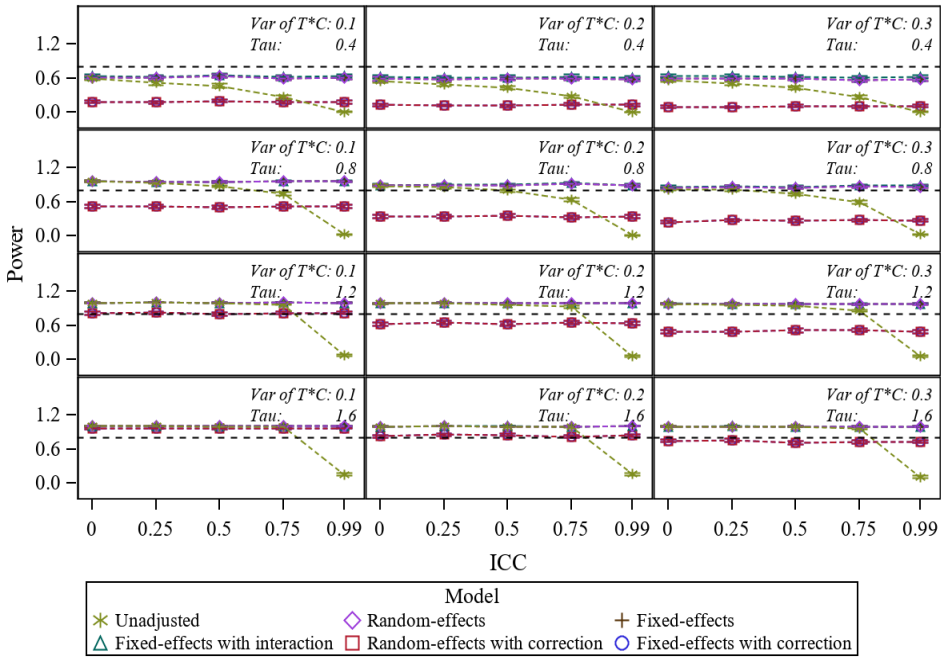


Figure 2.10. Empirical power for hypothesis test of $\boldsymbol{\tau} = 0$ when treatment effect is heterogeneous across treatment groups ($\boldsymbol{\sigma^2_{tc}} > 0$) under different scenarios using Type III sum of squares for 'appropriate' and 'inappropriate' methods of analysis.

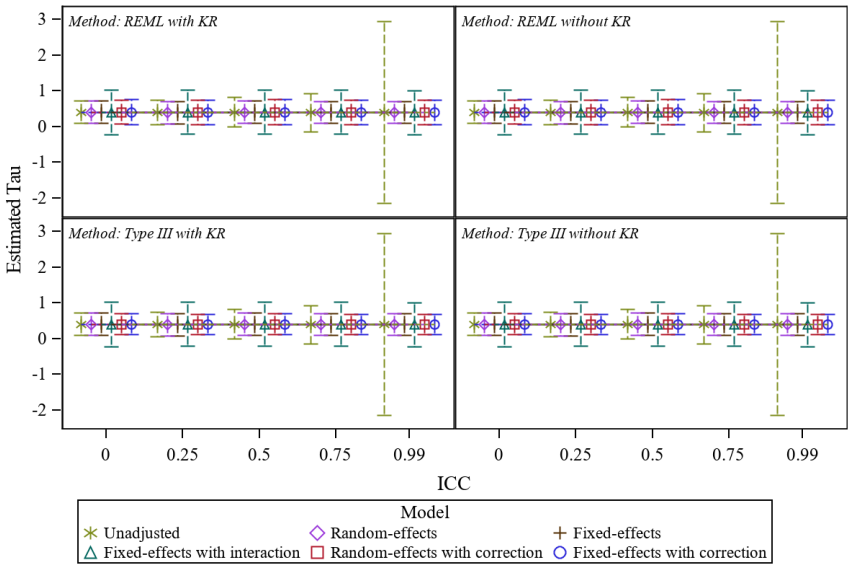


Figure 2.11. Average of Estimated $\boldsymbol{\tau}$ (true value is 0.4) with 95% Confident Intervals across 1000 simulations by ICC when treatment effect is homogeneous across treatment groups ($\boldsymbol{\sigma^2_{tc}} = 0$) under different scenarios for 'appropriate' and 'inappropriate' methods of analysis.

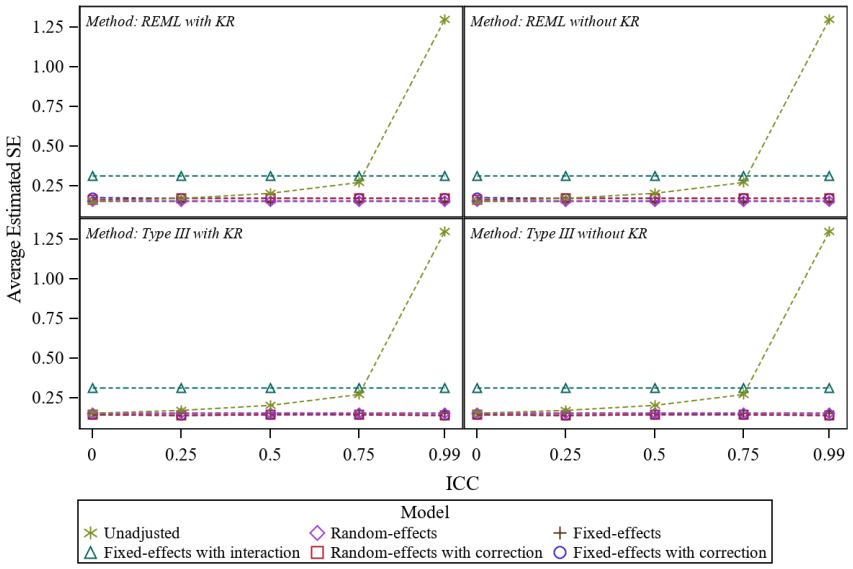


Figure 2.12. Average of the estimated standard error (SE) for treatment effect $\boldsymbol{\tau}$ (true value is 0.4) across 1000 simulations by ICC when treatment effect is homogeneous across treatment groups ($\boldsymbol{\sigma^2_{tc}} = 0$) under different scenarios for 'appropriate' and 'inappropriate' methods of analysis.

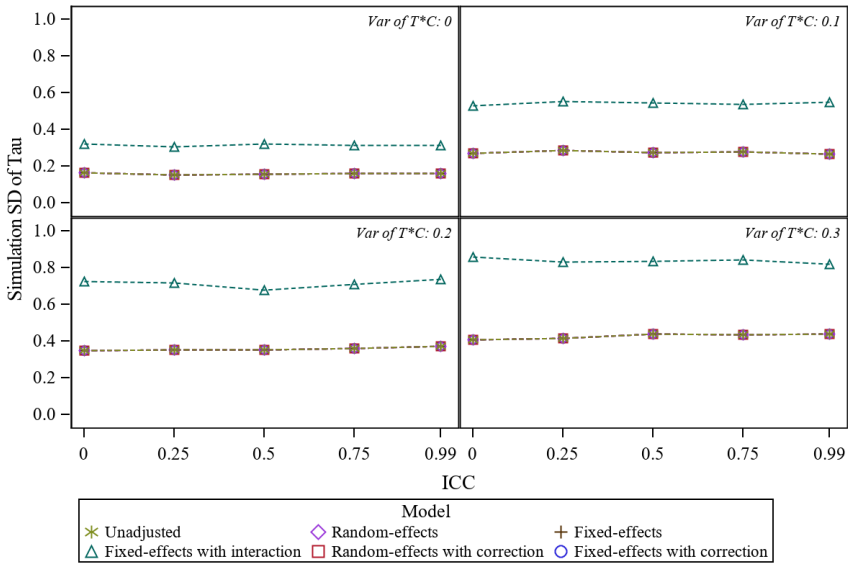


Figure 2.13. The simulation or empirical standard deviation (SD) of the estimated $\boldsymbol{\tau}$ (true value is 0.4) across 1000 simulations by ICC using Type III sum of squares under homogeneous and heterogeneous scenarios for 'appropriate' and 'inappropriate' methods of analysis.

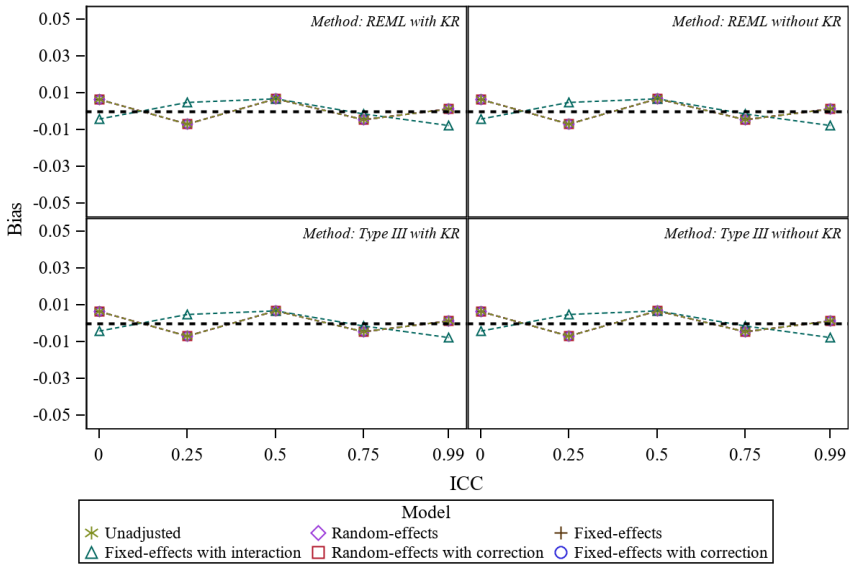


Figure 2.14. Bias of the estimated $\boldsymbol{\tau}$ (true value is 0.4) across 1000 simulations by ICC when treatment effect is homogeneous across treatment groups ($\boldsymbol{\sigma_{tc}^2 = 0}$) under different scenarios for 'appropriate' and 'inappropriate' methods of analysis.

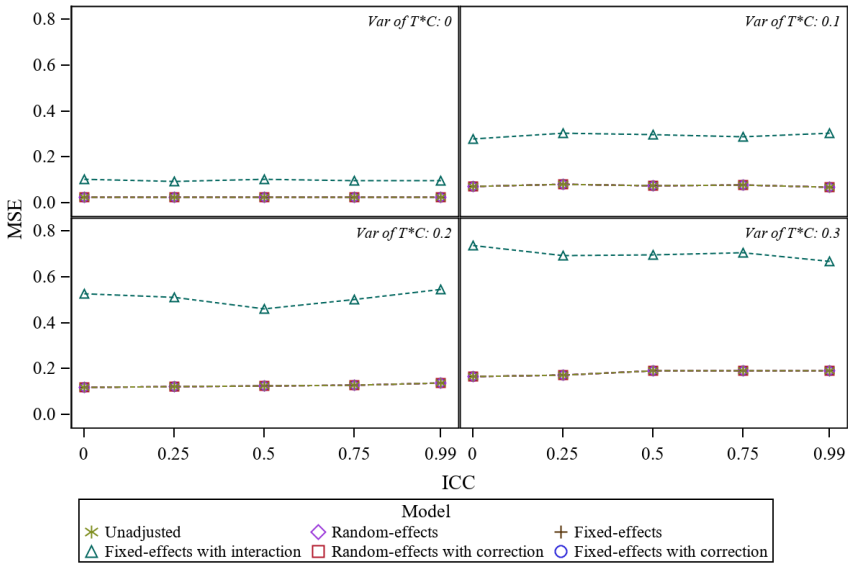


Figure 2.15. MSE of the estimated $\boldsymbol{\tau}$ (true value is 0.4) across 1000 simulations by ICC using Type III sum of squares under homogeneous and heterogeneous scenarios for 'appropriate' and 'inappropriate' methods of analysis.

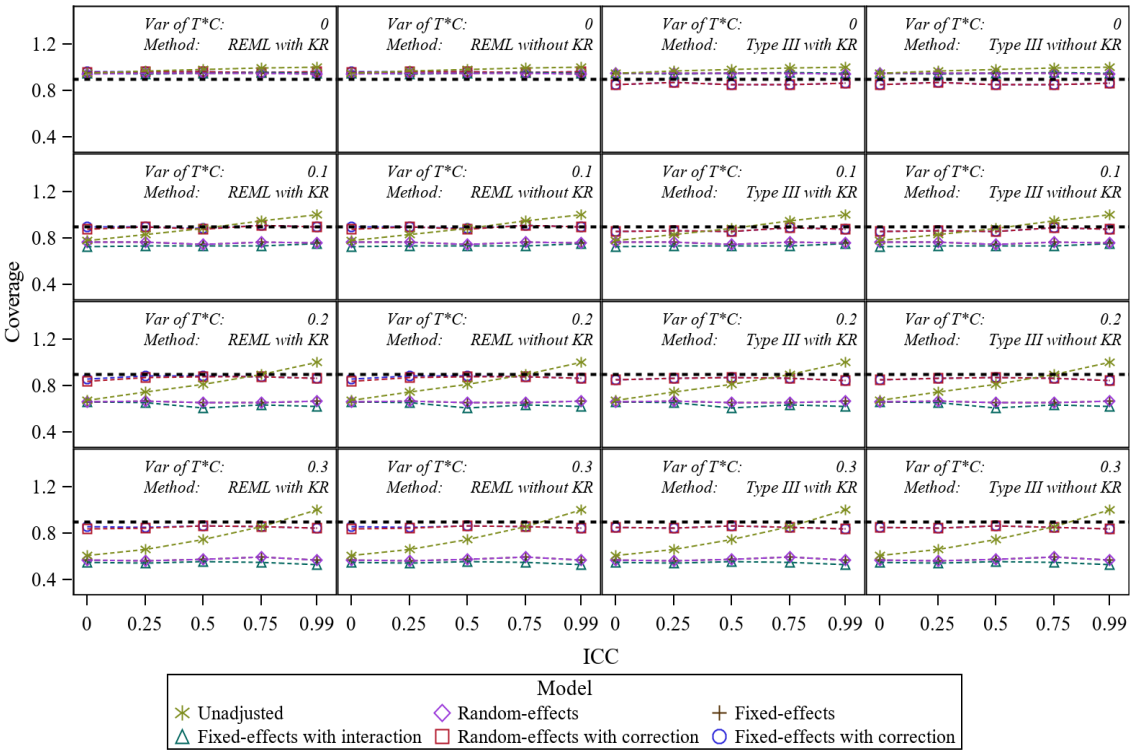


Figure 2.16. The coverage rate of the 95% confidence interval for the estimated $\boldsymbol{\tau}$ (true value is 0.4) across 1000 simulations by ICC under different scenarios for 'appropriate' and 'inappropriate' methods of analysis.

2.2 Center Effect with Other Covariates

We illustrate some general adjustment approaches using the typical example, where there were two covariates of interest, center and litter. For simplicity, we start with two treatment groups.

$$Y_{ijkl} = \mu + \tau_i + \beta_j + \gamma_{k(j)} + \delta_{ik(j)} + e_{ijkl},$$
$$i = 1, 2, j = 1, \ldots, c, k = 1, \ldots, r, l = 1, \ldots, n \tag{2.38}$$

where Y_{ijkl} represents a continuous outcome measured for the *l*th subject in the *j*th center and *k*th litter receiving the ith treatment (control or treatment group). μ represents an intercept term (the overall response mean). The treatment effects τ_i's are fixed, with $\sum_i \tau_i = 0$. The additive effect of the jth center, β_j, the kth litter, $\gamma_{k(j)}$, and the ith treatment in the kth litter, $\delta_{ik(j)}$, are random variables. And then, β_j s, $\gamma_{k(j)}$ s, and e_{ijk} s are all independent with $\beta_j \sim$ N(0, σ_c^2), $\gamma_{k(j)} \sim$ N(0, σ_r^2), $\delta_{ik(j)} \sim$ N(0, σ_δ^2), and $e_{ijkl} \sim$ (0, σ_ε^2).

Our preliminary study shows that an inappropriate approach for hypothesis tests of center and interaction effects could lead to incorrect decisions for removing the center/interaction term from the full model. Furthermore, it will result in the misspecification of the final model for the inference of the treatment effect. For instance, applying the REML method on a mixed model when the number of centers is small, it could accidentally drop the interaction random terms based on their significant level of test. These failures to do so can lead to an invalid inference for treatment effect, including (*i*) missing interaction random term: inflated Type I error rate and falsely high power; (*ii*) missing center random term: SEs for treatment effect biased upwards, resulting in too wide confident intervals and a power reduction. In this section, we explored the effect of missing random effects on more complex scenarios, and which approaches are more robust.

2.2.1 Simulation Study

a) Data Generating

In this simulation study, we used Monte Carlo simulation to assess the performance of three statistical models to analyze a multicenter study considering center and litter effects with a continuous outcome when the litter level is small (number of litters is only 5), and the number of centers varies (i.e., 5, or 25). We generated continuous outcomes, *Y*, using the mixed-effects linear regression model as Eq. 2.38. We generated the random error, ε_{ijkl}, from $N(0, \sigma_\varepsilon^2 = 0.5)$.

To investigate the impact of missing random effects, we generated a three-way interaction random effect σ_δ^2 (with 20 or 100 levels), a litter within center random effect σ_r^2 (with 25 or 125 levels), and a center random effect σ_c^2 (with 5 or 25 levels). A catalog of simulation designs is shown in Table 2.7. To comprehensively study the behavior of candidate models for un-modeling random effects, we mainly focused on the interpretation of the results on three scenarios as follows,

(a) Scenario 1: $\sigma_c^2 = 0.17$, $\sigma_r^2 = 0$, and $\sigma_\delta^2 = 0$,

(b) Scenario 2: $\sigma_c^2 = 0.23$, $\sigma_r^2 = 0.2$, and $\sigma_\delta^2 = 0$,

(c) Scenario 3: $\sigma_c^2 = 0.3$, $\sigma_r^2 = 0.2$, and $\sigma_\delta^2 = 0.2$.

Table 2.7. The catalog of simulation designs

Scenario	σ_c^2	σ_r^2	σ_δ^2	τ	Litter	Center Level, N/litter
1-8	0, 0.17, 0.23	0	0	0, 0.4	5	C=5, N=20 C=25, N=4
9-16		0.2				
17-24	0, 0.23, 0.30	0	0.2			
25-32		0.2				

b) Performance Measures

For each scenario, we generated 10,000 hypothetical trial datasets, expecting small errors due to the large number of simulations. Then, we applied three statistical models (shown in Table 2.7) with Type III or REML to each simulated dataset. We purposefully mis-specified models in some scenarios by ignoring additional data structure.

For each simulation scenarios and model, we evaluated the parameters as follows,

1) For model 1, the estimated $\hat{\tau}$ for the fixed treatment effect τ, $\hat{\sigma}_c^2$ for the random center effect σ_c^2, $\widehat{\gamma_k}$ for the fixed litter effect γ_k, and $\hat{\sigma}_\varepsilon^2$ for the residual σ_ε^2.
2) For model 2, the estimated $\hat{\tau}$ for the fixed treatment effect τ, $\hat{\sigma}_c^2$ for the random center effect σ_c^2, $\hat{\sigma}_r^2$ for the random litter (center) effect σ_r^2, and $\hat{\sigma}_\varepsilon^2$ for the residual σ_ε^2.
3) For model 3, the estimated $\hat{\tau}$ for the fixed treatment effect τ, $\hat{\sigma}_c^2$ for the random center effect σ_c^2, $\hat{\sigma}_r^2$ for the random litter (center) effect σ_r^2 , $\hat{\sigma}_\delta^2$ for the random $A \times C \times L$ interaction effect σ_δ^2, and $\hat{\sigma}_\varepsilon^2$ for the residual σ_ε^2.

Additionally, we estimated the following measures for treatment effect τ:

1) Type I error rate (when the true $\tau = 0$) and Power (when the true $\tau > 0$).
2) Average of Point estimated τ and estimated SE for τ, with simulation standard deviation (SD) for $\hat{\tau}$.
3) Bias and mean squared error (*MSE*) of the point estimator of treatment effect τ.
4) Empirical coverage rate of the 95% confidence intervals (CIs) around τ.

The details of explanation as the same as section 2.1.4.2 - part b). All datasets were simulated and analyzed in SAS software (version 9.4).

Table 2.8. Descriptions of Three Models

Item	Model[1]	Note
Model 1	$Y_{ijkl} = \mu + \tau_i + \beta_j + \gamma_k + e_{ijkl}$	β_j is random, and γ_k is fixed
Model 2	$Y_{ijkl} = \mu + \tau_i + \beta_j + \gamma_{k(j)} + e_{ijkl}$	β_j and $\gamma_{k(j)}$ are random
Model 3	$Y_{ijkl} = \mu + \tau_i + \beta_j + \gamma_{k(j)} + \delta_{ik(j)} + e_{ijkl}$	$\beta_j, \gamma_{k(j)}$ and $\delta_{ik(j)}$ are random

[1]where y_{ijkl} represents a continuous outcome measured for the *l*th subject in the *k*th litter and *j*th center receiving the *i*th treatment (control or treatment group).

c) Results and Discussion

Missing random effect predictors had little effect on the fixed effect estimates but had systematic impacts on the estimates of random effects (Figure 2.17-2.22). The variance due to the un-modelled random effect was almost entirely absorbed by the nested random effect variance of interest. The residual variance absorbed most of the variance arising from the unmodelled random effect, and only a small fraction appeared in the random effect of other interests. As the number of centers got larger (n_c = 25), the variance of residual has absorbed more the unmodelled random components. Under scenario 3, both model 1 and model 2 were mis-specified models that were missing random effects. Since model 2 included the litter effect as random instead of fixed (in model), variance of litter effect also was absorbed some fraction unmodelled random component. It was resulting in the smaller estimates of residual in model 2 than model 1.

Table 2.12 – 2.14 summarized descriptive statistics of the point estimator of treatment effect in all Models for selected scenarios in the balanced design. Overall, the point estimates of τ were unbiased in all three of the Type III and REML methods models. Surprisingly, point estimates of fixed effect for all models were identical in scenarios with the same center size, even for some scenarios with missing random effects as described above. When the design is balanced, the point estimate of τ doesn't depend on the random effect estimate. In other words, when treatments are allocated in the same proportion in all center-by-litter groups, all random effects such as center, litter (center), or three-way interaction have no association with the treatment allocation, hence

adjusting for these random effects or not have little impact on the point estimate of the treatment - response relationship given a continuous response variable. Thus, different approaches to incorporate these covariates' information resulted in the same treatment contrast estimates in a balanced design.

Furthermore, the same point estimates across 10,000 Monte Carlo simulations led to the estimator's identical empirical SD and overall error rate (measured by MSE) for $\hat{\tau}$ in all Models (true model or missing random term model) for scenarios with the same center size. When the center number was only 5, the miss-modeling random effect (Model 1 and 2) tended to slightly underestimate the standard error of $\hat{\tau}$, compared to Model 3. This difference reduced as the center number increased.

In this study, some datasets were generated that the treatment effects were heterogeneous among center-by-litter groups (i.e., nonzero $A \times C \times L$ three-way interaction), the fixed-effects analysis (Model 1, litter as fixed effect) performed better than random-effects analysis (Model 2, litter as random effect), especially when the center number was larger. However, when the treatment effects were homogeneous among center-by-litter groups (i.e., no $A \times C \times L$ three-way interaction) along with nonzero litter (center) effect, random-effects analysis (Model 2, litter as random effect) outperformed the fixed-effects analysis (Model 1, litter as fixed effect) regardless of center size. Model 3 considered the observed "heterogeneity" for obtaining the precise point estimate and the associated standard error for the treatment effect τ, thereby ensuring the validity of hypothesis testing.

The statistical Type I error rate, power analysis, and the empirical coverage of 95% confidence in balanced studies are displayed in Figure. 2.22 - 2.24. Regardless of Type III or REML applied, the models under fulfilled assumptions produced a nominal value of 95% for

coverage rate, a nominal value of 5% for Type I error rate and > 80% power. When missing random effect in models (Model 1 and Model 2) for scenarios with nonzero σ_δ^2 (shown in (c) of Figure. 2.22 - 2.24) and small center size (n_c = 5), Model 1 and Model 2 provided the inflated Type I error, false power, and low coverage rate of the 95% confidence interval for the estimated τ, due to underestimated SE for $\hat{\tau}$. As center size increased (n_c = 25), this issue improved in terms of the nominal value of 5% for Type I error rate, reliable power, and above 95% coverage rate. Since some datasets that the litter (center) effects were generated as random (shown in (b) of Figure. 2.22 - 2.24), random-effects analysis (Model 2, litter as random effect) outperformed the fixed-effects analysis (Model 1, litter as fixed effect) regardless of the number of centers in terms of the nominal value of 5% for Type I error rate.

d) Conclusion

In this study, we investigated the performance of three statistical models using Type III or REML, which have been commonly recommended for analyzing continuous outcomes in multicenter randomized clinical trials. This simulation study revealed that all three models produced unbiased estimates of treatment effect in balanced cases, regardless of modeling or un-modeling non-zero random effect. Despite this, if treatment effects were heterogeneous among center-by-litter groups, inappropriate models (i.e., missing random effects) could underestimate the standard error of the effect estimates when a small number of centers were involved, resulting in incorrect inference.

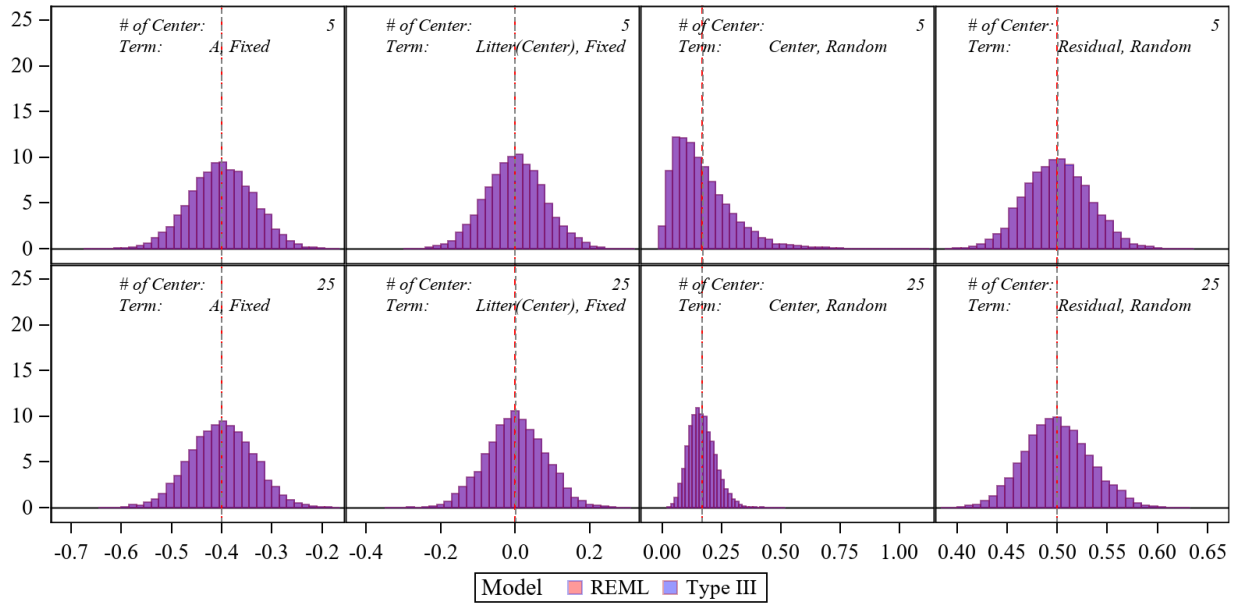


Figure 2.17. Parameter estimates of interest for Model 1 under fulfilled assumptions ($\boldsymbol{\sigma_c^2}$ = 0.17, $\boldsymbol{\sigma_r^2}$ = 0, and $\boldsymbol{\sigma_\delta^2}$ = 0). The three columns show the distribution of point estimates across 10,000 replicated simulation runs for the fixed effect slope, the center variance, and the residual variance. The simulated true value is shown as a red dashed line, and the mean of the estimated values is shown as a grey dashed line.

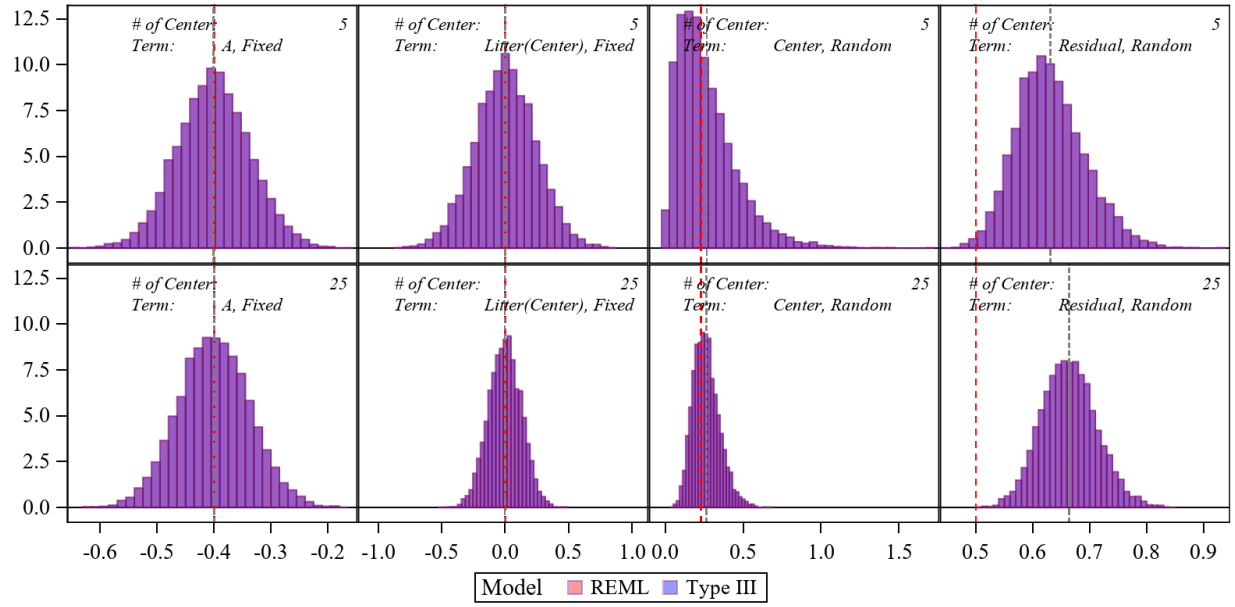


Figure 2.18. Effect of missing random effects on Parameter estimates of interest for Model 1 ($\boldsymbol{\sigma_c^2}$ = 0.23, $\boldsymbol{\sigma_r^2}$ = 0.2, and $\boldsymbol{\sigma_\delta^2}$ = 0). The three columns show the distribution of point estimates across 10,000 replicated simulation runs for the fixed effect slope, the center variance, and the residual variance. The simulated true value is shown as a red dashed line, and the mean of the estimated values is shown as a grey dashed line.

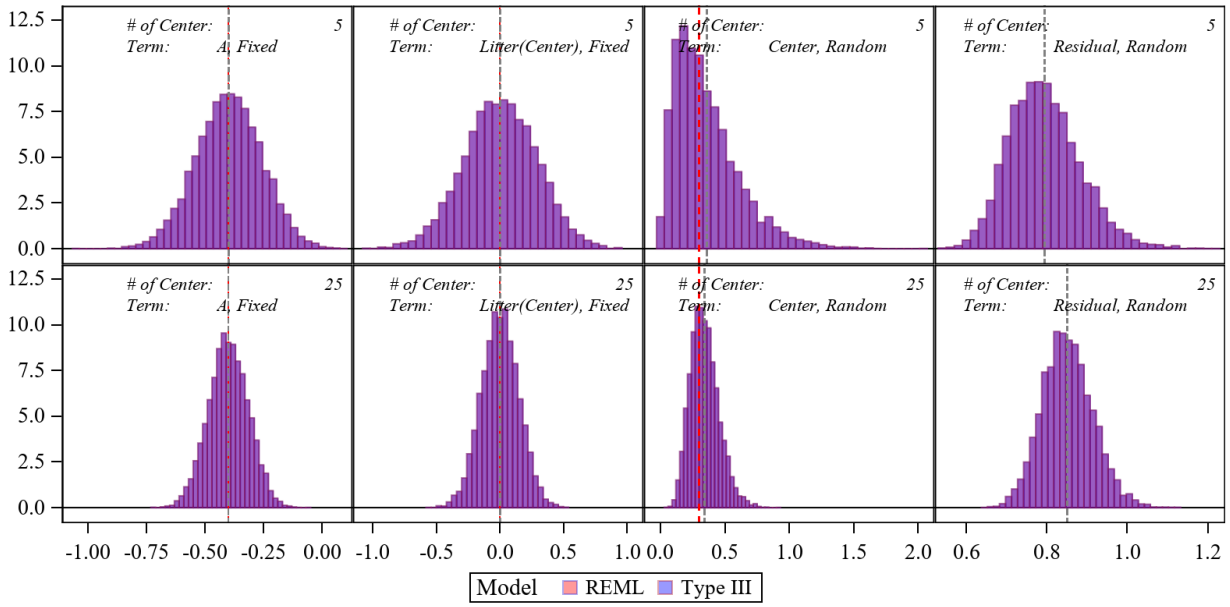


Figure 2.19. Effect of missing random effects on Parameter estimates of interest for Model 1 ($\boldsymbol{\sigma_c^2}$ = 0.3, $\boldsymbol{\sigma_r^2}$ = 0.2, and $\boldsymbol{\sigma_\delta^2}$ = 0.2). The three columns show the distribution of point estimates across 10,000 replicated simulation runs for the fixed effect slope, the center variance, and the residual variance. The simulated true value is shown as a red dashed line, and the mean of the estimated values is shown as a grey dashed line.

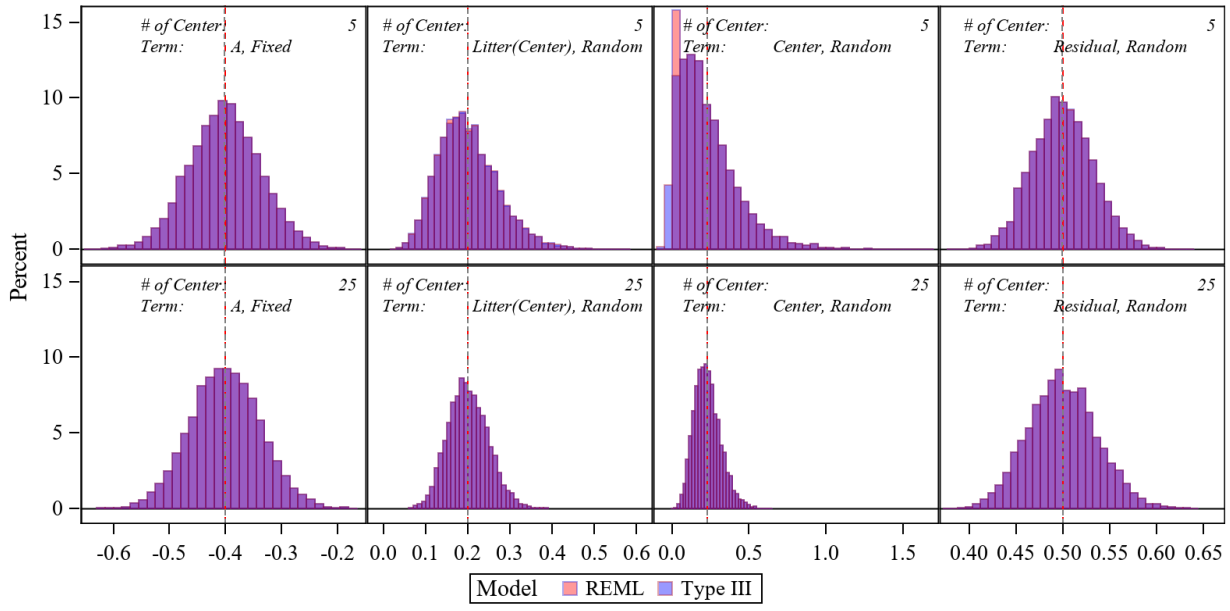


Figure 2.20. Parameter estimates of interest for Model 2 under fulfilled assumptions ($\boldsymbol{\sigma_c^2}$ = 0.23, $\boldsymbol{\sigma_r^2}$ = 0.2, and $\boldsymbol{\sigma_\delta^2}$ = 0). The four columns show the distribution of point estimates across 10,000 replicated simulation runs for the fixed effect slope, the center variance, the litter (center) variance, and the residual variance. The simulated true value is shown as a red dashed line, and the mean of the estimated values is shown as a grey dashed line.

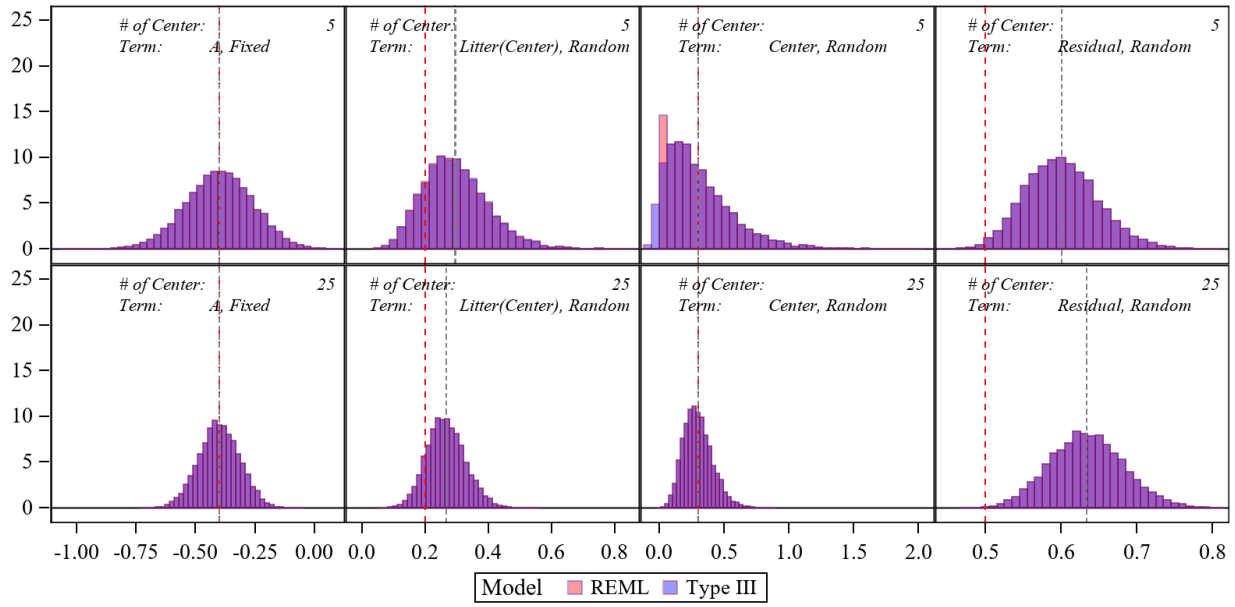


Figure 2.21. Effect of missing random effects on Parameter estimates of interest for Model 2 ($\boldsymbol{\sigma_c^2}$ = 0.3, $\boldsymbol{\sigma_r^2}$ = 0.2, and $\boldsymbol{\sigma_\delta^2}$ = 0.2). The four columns show the distribution of point estimates across 10,000 replicated simulation runs for the fixed effect slope, the center variance, the litter (center) variance, and the residual variance. The simulated true value is shown as a red dashed line, and the mean of the estimated values is shown as a grey dashed line.

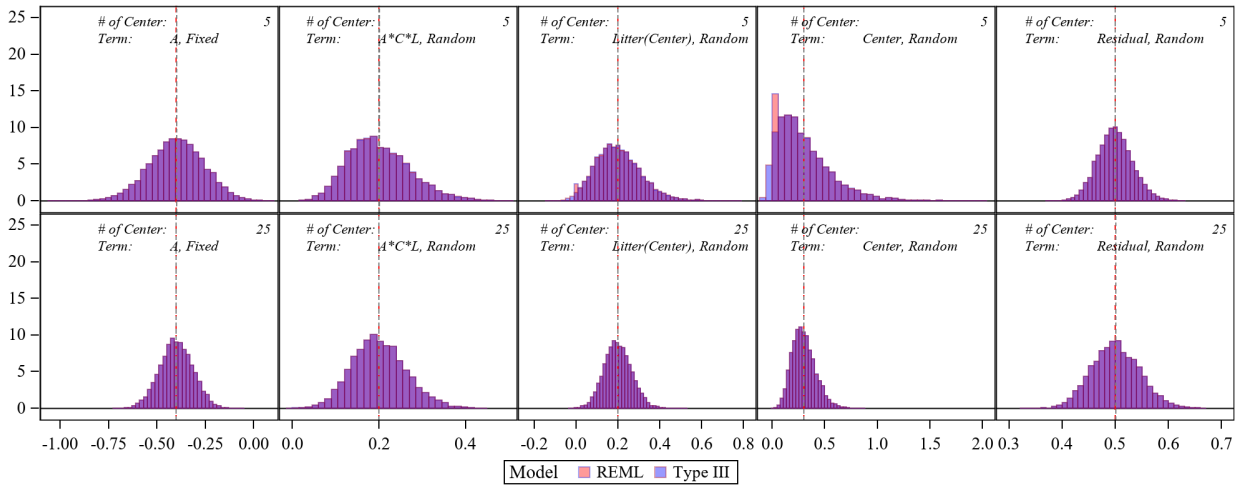


Figure 2.22. Parameter estimates of interest for Model 3 under fulfilled assumptions ($\boldsymbol{\sigma_c^2}$ = 0.3, $\boldsymbol{\sigma_r^2}$ = 0.2, and $\boldsymbol{\sigma_\delta^2}$ = 0.2). The five columns show the distribution of point estimates across 10,000 replicated simulation runs for the fixed effect slope, the center variance, the litter (center) variance, the $\boldsymbol{A \times C \times L}$ variance, and the residual variance. The simulated true value is shown as a red dashed line, and the mean of the estimated values is shown as a grey dashed line.

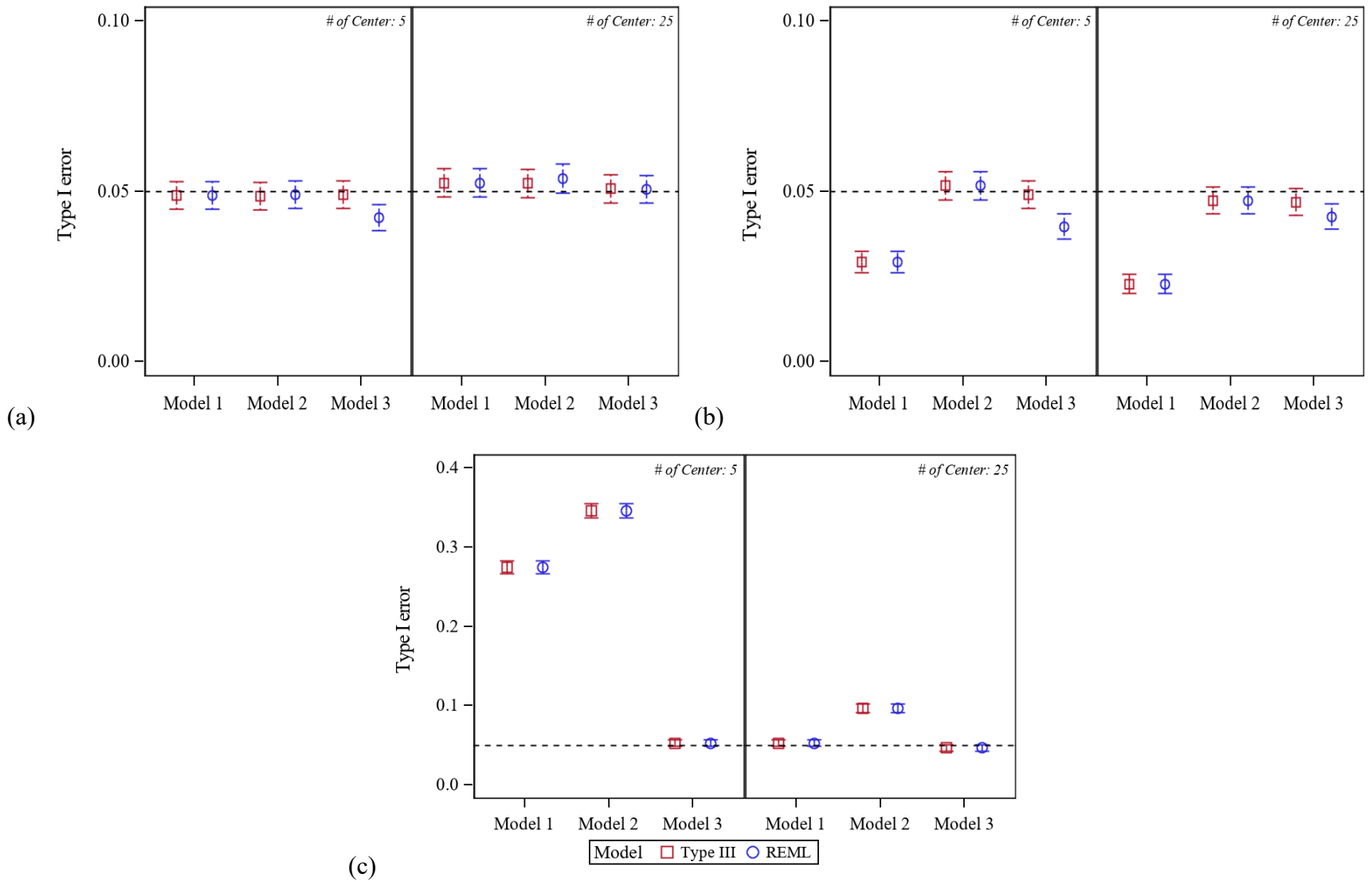


Figure 2.23. Type I error rate for hypothesis test of $\boldsymbol{\tau} = 0$ for model under scenarios with small or medium center number (5 *vs* 25) using Type III or REML. (a) Scenario 1: $\boldsymbol{\sigma_c^2}$ = 0.17, $\boldsymbol{\sigma_r^2}$ = 0, and $\boldsymbol{\sigma_\delta^2}$ = 0; (b) Scenario 2: $\boldsymbol{\sigma_c^2}$ = 0.23, $\boldsymbol{\sigma_r^2}$ = 0.2, and $\boldsymbol{\sigma_\delta^2}$ = 0; (c) Scenario 3: $\boldsymbol{\sigma_c^2}$ = 0.3, $\boldsymbol{\sigma_r^2}$ = 0.2, and $\boldsymbol{\sigma_\delta^2}$ = 0.2.

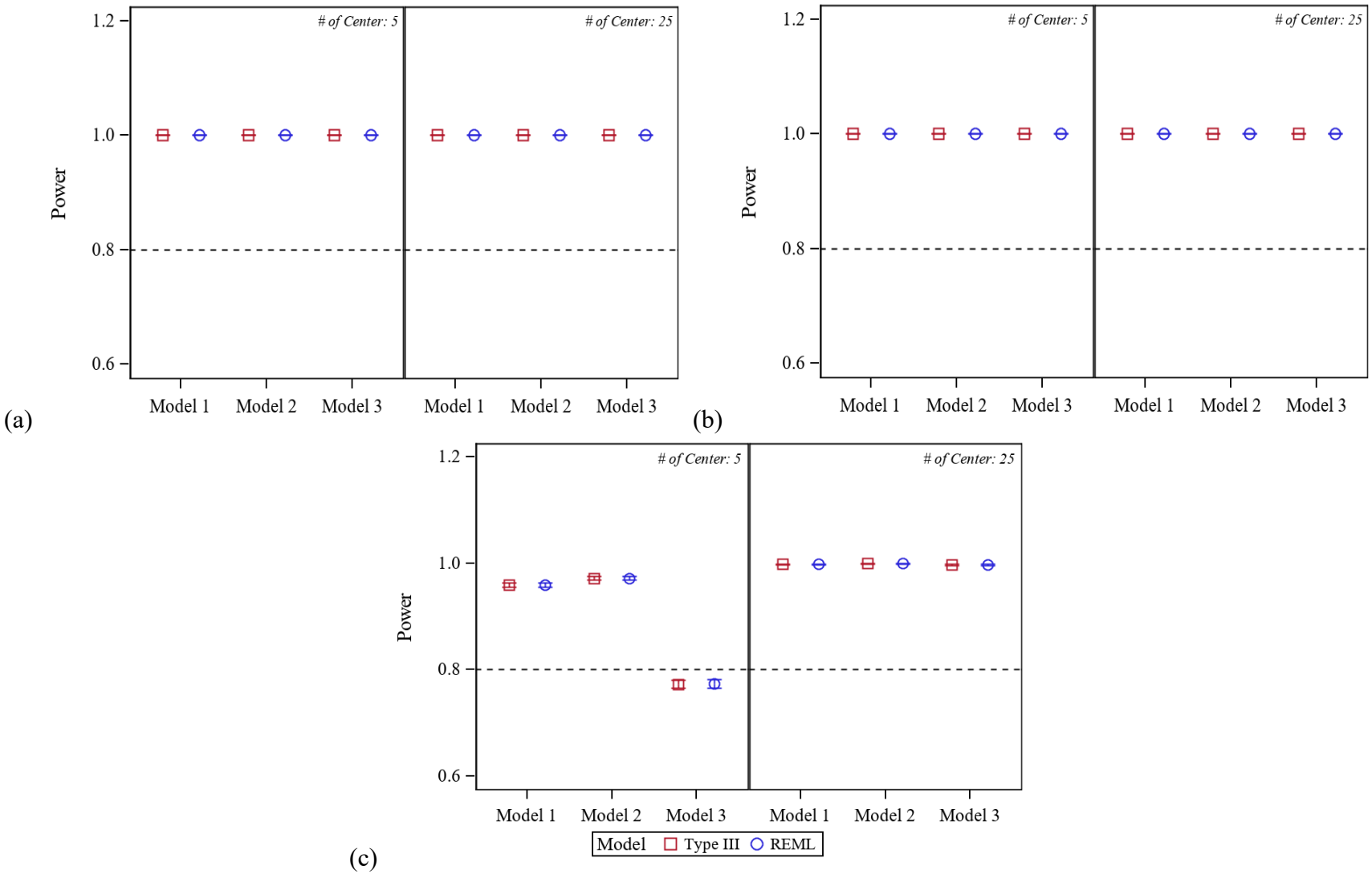


Figure 2.24. Power analysis for hypothesis test of $\boldsymbol{\tau} = 0$ for model under scenarios with small or medium center number (5 *vs* 25) using Type III or REML. (a) Scenario 1: $\boldsymbol{\sigma_c^2}$ = 0.17, $\boldsymbol{\sigma_r^2}$ = 0, and $\boldsymbol{\sigma_\delta^2}$ = 0; (b) Scenario 2: $\boldsymbol{\sigma_c^2}$ = 0.23, $\boldsymbol{\sigma_r^2}$ = 0.2, and $\boldsymbol{\sigma_\delta^2}$ = 0; (c) Scenario 3: $\boldsymbol{\sigma_c^2}$ = 0.3, $\boldsymbol{\sigma_r^2}$ = 0.2, and $\boldsymbol{\sigma_\delta^2}$ = 0.2.

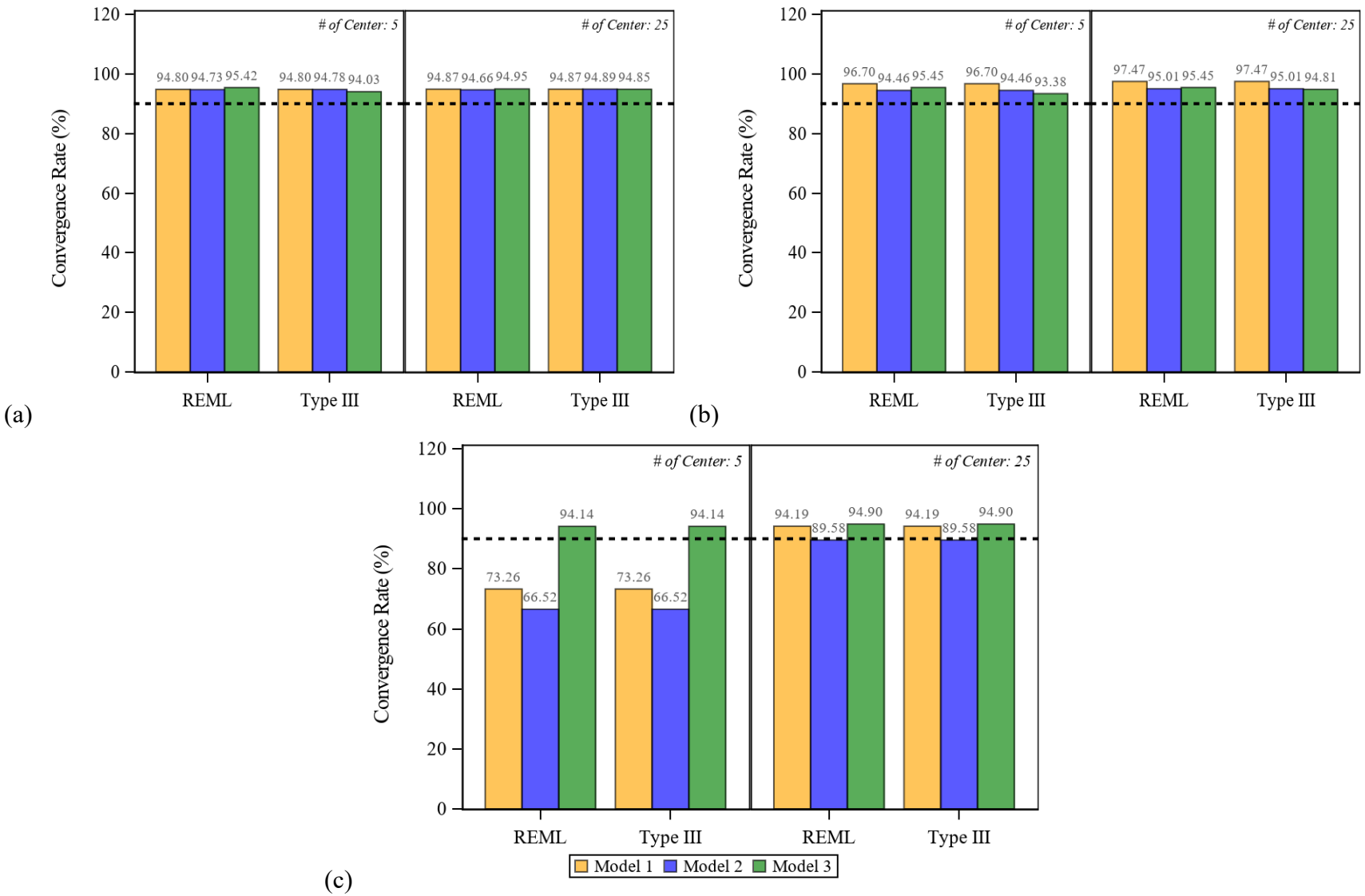


Figure 2.25. The coverage rate of the 95% confidence interval for the estimated $\boldsymbol{\tau} = 0.4$ across 10,000 simulations for model under scenarios with small or medium center number (5 *vs* 25) using Type III or REML. (a) Scenario 1: $\boldsymbol{\sigma_c^2} = 0.17$, $\boldsymbol{\sigma_r^2} = 0$, and $\boldsymbol{\sigma_\delta^2} = 0$; (b) Scenario 2: $\boldsymbol{\sigma_c^2} = 0.23$, $\boldsymbol{\sigma_r^2} = 0.2$, and $\boldsymbol{\sigma_\delta^2} = 0$; (c) Scenario 3: $\boldsymbol{\sigma_c^2} = 0.3$, $\boldsymbol{\sigma_r^2} = 0.2$, and $\boldsymbol{\sigma_\delta^2} = 0.2$.

Table 2.9. Properties of point estimates of the treatment effect $\boldsymbol{\tau}$ for models using Type III or REML under scenario ($\boldsymbol{\sigma_c^2}$ = 0.17, $\boldsymbol{\sigma_r^2}$ = 0, and $\boldsymbol{\sigma_\delta^2}$ = 0) with small or medium center size (5 *vs* 25).

			REML				Type III			
τ	Center size	Model[1]	Mean effect ± SD	Bias	MSE	Avg.SE	Mean effect ± SD	Bias	MSE	Avg.SE
0	5	Model 1	-0.002 ± 0.063	-0.002	0.004	0.063	-0.002 ± 0.063	-0.002	0.004	0.063
		Model 2	-0.002 ± 0.063	-0.002	0.004	0.063	-0.002 ± 0.063	-0.002	0.004	0.063
		Model 3	-0.002 ± 0.063	-0.002	0.004	0.065	-0.002 ± 0.063	-0.002	0.004	0.063
	25	Model 1	-0.001 ± 0.063	-0.001	0.004	0.063	-0.001 ± 0.063	-0.001	0.004	0.063
		Model 2	-0.001 ± 0.063	-0.001	0.004	0.063	-0.001 ± 0.063	-0.001	0.004	0.063
		Model 3	-0.001 ± 0.063	-0.001	0.004	0.064	-0.001 ± 0.063	-0.001	0.004	0.063
0.4	5	Model 1	0.400 ± 0.063	-0.000	0.004	0.063	0.400 ± 0.063	-0.000	0.004	0.063
		Model 2	0.400 ± 0.063	-0.000	0.004	0.063	0.400 ± 0.063	-0.000	0.004	0.063
		Model 3	0.400 ± 0.063	-0.000	0.004	0.065	0.400 ± 0.063	-0.000	0.004	0.063
	25	Model 1	0.400 ± 0.063	-0.000	0.004	0.063	0.400 ± 0.063	-0.000	0.004	0.063
		Model 2	0.400 ± 0.063	-0.000	0.004	0.063	0.400 ± 0.063	-0.000	0.004	0.063
		Model 3	0.400 ± 0.063	-0.000	0.004	0.064	0.400 ± 0.063	-0.000	0.004	0.063

[1] For this scenario, all three models were over-specified models.

Table 2.10. Properties of point estimates of the treatment effect $\boldsymbol{\tau}$ for models using Type III or REML under scenario ($\boldsymbol{\sigma_c^2}$ = 0.23, $\boldsymbol{\sigma_r^2}$ = 0.2, and $\boldsymbol{\sigma_\delta^2}$ = 0) with small or medium center size (5 *vs* 25).

			REML				Type III			
τ	Center size	Model[1]	Mean effect $\pm$ SD	Bias	MSE	Avg.SE	Mean effect $\pm$ SD	Bias	MSE	Avg.SE
0	5	Model 1	0.000 $\pm$ 0.063	0.000	0.004	0.071	0.000 $\pm$ 0.063	0.000	0.004	0.071
		Model 2	0.000 $\pm$ 0.063	0.000	0.004	0.063	0.000 $\pm$ 0.063	0.000	0.004	0.063
		Model 3	0.000 $\pm$ 0.063	0.000	0.004	0.067	0.000 $\pm$ 0.063	0.000	0.004	0.063
	25	Model 1	-0.001 $\pm$ 0.063	-0.001	0.004	0.073	-0.001 $\pm$ 0.063	-0.001	0.004	0.073
		Model 2	-0.001 $\pm$ 0.063	-0.001	0.004	0.063	-0.001 $\pm$ 0.063	-0.001	0.004	0.063
		Model 3	-0.001 $\pm$ 0.063	-0.001	0.004	0.064	-0.001 $\pm$ 0.063	-0.001	0.004	0.063
0.4	5	Model 1	0.401 $\pm$ 0.063	0.001	0.004	0.071	0.401 $\pm$ 0.063	0.001	0.004	0.071
		Model 2	0.401 $\pm$ 0.063	0.001	0.004	0.063	0.401 $\pm$ 0.063	0.001	0.004	0.063
		Model 3	0.401 $\pm$ 0.063	0.001	0.004	0.066	0.401 $\pm$ 0.063	0.001	0.004	0.063
	25	Model 1	0.400 $\pm$ 0.063	0.000	0.004	0.073	0.400 $\pm$ 0.063	0.000	0.004	0.073
		Model 2	0.400 $\pm$ 0.063	0.000	0.004	0.063	0.400 $\pm$ 0.063	0.000	0.004	0.063
		Model 3	0.400 $\pm$ 0.063	0.000	0.004	0.064	0.400 $\pm$ 0.063	0.000	0.004	0.063

[1] For this scenario, Model 1 was specified litter (center) effect τ_r as fixed, Model 2 was under fulfilled assumptions, and Model 3 was an over-specified model.

Table 2.11. Properties of point estimates of the treatment effect $\boldsymbol{\tau}$ for models using Type III or REML under scenario ($\boldsymbol{\sigma_c^2}$ = 0.3, $\boldsymbol{\sigma_r^2}$ = 0.2, and $\boldsymbol{\sigma_\delta^2}$ = 0.2) with small or medium center size (5 *vs* 25).

			REML				Type III			
τ	Center size	Model[1]	Mean effect $\pm$ SD	Bias	MSE	Avg.SE	Mean effect $\pm$ SD	Bias	MSE	Avg.SE
0	5	Model 1	0.001 $\pm$ 0.143	0.001	0.021	0.080	0.001 $\pm$ 0.142	0.001	0.021	0.080
		Model 2	0.001 $\pm$ 0.143	0.001	0.021	0.069	0.001 $\pm$ 0.142	0.001	0.021	0.069
		Model 3	0.001 $\pm$ 0.143	0.001	0.021	0.140	0.001 $\pm$ 0.142	0.001	0.021	0.140
	25	Model 1	0.001 $\pm$ 0.084	0.001	0.007	0.082	0.001 $\pm$ 0.084	0.001	0.007	0.082
		Model 2	0.001 $\pm$ 0.084	0.001	0.007	0.071	0.001 $\pm$ 0.084	0.001	0.007	0.071
		Model 3	0.001 $\pm$ 0.084	0.001	0.007	0.085	0.001 $\pm$ 0.084	0.001	0.007	0.085
0.4	5	Model 1	0.399 $\pm$ 0.140	-0.001	0.020	0.080	0.399 $\pm$ 0.140	-0.001	0.020	0.080
		Model 2	0.399 $\pm$ 0.140	-0.001	0.020	0.069	0.399 $\pm$ 0.140	-0.001	0.020	0.069
		Model 3	0.399 $\pm$ 0.140	-0.001	0.020	0.140	0.399 $\pm$ 0.140	-0.001	0.020	0.140
	25	Model 1	0.400 $\pm$ 0.085	0.000	0.007	0.082	0.400 $\pm$ 0.085	0.000	0.007	0.082
		Model 2	0.400 $\pm$ 0.085	0.000	0.007	0.071	0.400 $\pm$ 0.085	0.000	0.007	0.071
		Model 3	0.400 $\pm$ 0.085	0.000	0.007	0.085	0.400 $\pm$ 0.085	0.000	0.007	0.085

[1] For this scenario, Model 1 was missing random effects σ_δ^2 with fixed parameter τ_r, Model 2 was missing random effects σ_δ^2, and Model 3 was under fulfilled assumptions.

Chapter 3 - Multicenter Randomized Clinical Trials with Binary Outcomes

In Chapter 2, we have shown that analyses of multicenter RCTs with continuous outcomes often need to account for the center, other covariates and interaction effects to ensure valid inference. This chapter will discuss when and how to account for these variables for binary outcomes appropriately. Since the analysis of multicenter RCTs with binary outcomes accounting for center and other covariates is more complicated, we first explored the common method.

In medical research, when comparing two treatments using binary data, the odds ratio, rate ratio (often called relative risk or risk ratio), and risk difference are frequently specified as primary measures. The risk difference is an absolute measurement of effect, whereas the relative risk and the odds ratio used more in medical research are relative measurements for comparing outcomes. In some studies, the odds ratio is preferred because the inference of it is straightforward since it has a straight relationship with the regression coefficient in a logistic regression model. However, relative risk is used more than the other two parameters in randomized clinical trials, especially since the two relevant proportions are probably both small (Wang & Shan, 2015).

In this chapter, we are interested in the ICC on the logistic scale, which can be written as the proportion of the total outcome variance that is due to between-center variation:

$$\rho_c = \frac{\sigma_c^2}{\sigma_c^2 + \frac{s^2\pi^2}{3}} \tag{3.1}$$

As the same as Multicenter RCTs with continuous outcomes, most studies with binary outcomes also assume that the treatment effects were homogeneous across all centers, that is, no treatment-by-center interaction (Edgar et al., 2021; Kahan, 2014; Kim et al., 2020). However, we

suspect that the constant treatment effect assumption may not be satisfactory in the real world. Thus, an extension from Chapter 2, when $\sigma_{rc}^2 \neq 0$, which means the treatment effects were heterogeneous among 'block' groups, then ρ_c is defined as

$$\rho_c = \frac{\sigma_c^2 - \sigma_{rc}^2/(k-1)}{\sigma_c^2 + \sigma_{rc}^2 + \frac{s^2\pi^2}{3}} \tag{3.2}$$

For this chapter, we investigated (1) the appropriate method for model estimates of interests under balanced and unbalanced cases; and (2) the consequences of missing random effect components on model estimates of interests.

3.1 Single-center RCT Example

Let's consider a single-center RCT first. Suppose we are interested in estimating the overall treatment efficacy. More specifically, in veterinary biologics, a clinical study assesses a vaccine's ability to prevent clinical disease. Generally, the measurement for prevention is estimated through evaluation on an individual basis, a binary outcome associated with a case definition of disease (affected or unaffected) (Wakeland & Fergen, 2016). Then, those binary outcomes were aggregated within treatment groups to estimate the proportion of affected animals for that group. Prevented Fraction (PF) is the primary conclusion criterion suggested by the USDA (Reference VSM 800.202) to support the licensure of a vaccine having efficacy. PF is the complement of the relative risk ($1 - \phi$), where $\phi = p_1/p_0$. Here, p_1 represents the affected fraction in the experimental product group, and p_0 represents the affected fraction in the placebo group.

More often, in randomized controlled efficacy studies, blocking or stratifying factors (e.g., housing, gender, or litter) need to be considered in statistical methods to estimate the PF. The ideal study design utilizes a 1:1 allocation rule for randomizing animals to the treatment and control

groups within blocks. However, in real-world studies, achieving balance in each block is not always possible. Thus, the overall relative risk must be estimated across strata of varying sizes.

3.2 Adjustment Models

We use either a marginal or conditional model to account for blocking or stratified variables. The common conditional models include a mixed-effects logistic regression model with a random intercept for covariates, one of the generalized linear mixed models (GLMMs), and the Cochran Mantel Haenszel approach. The marginal model is recommended for generalized estimating equations. Both marginal and conditional models target to estimate the difference between treatment groups (e.g., odds ratio and relative risk), and generally, the true value of these estimands is different (Hauck et al., 1998; Robinson & Jewell, 1991). In marginal models, these estimands compare the treatment effect difference for subjects across strata, whereas, in conditional models, they compare the treatment effect difference for subject across strata. If no treatment effect exists or the ICC is zero, the estimates for marginal and conditional models are identical (Kim et al., 2020). As the ICC increases, the difference between these two types of estimands' values increases.

3.2.1 Cochran-Mantel-Haenszel (CMH)

When analyzing stratified 2×2 tables, the standard procedure used to estimate overall relative risk was introduced by Cochran (1954), who first proposed a hypothesis test for the difference in proportions across strata. Mantel and Haenszel (1959) proposed a very similar test and introduced an estimator for a common odds ratio, which also approximates the relative risk under most case-control study circumstances with low prevalence. Rothman and Boice (1979) first proposed a relative risk estimator, which is analogous to the Mantel-Haenszel estimator for the odds ratio. We refer to this class of estimation procedures for the relative risk as Cochran-Mantel-

Haenszel (CMH), and the estimate of the common ϕ is also called the Mantel-Haenszel estimator. An example of the type of table required for CMH ϕ estimates is shown in Table 3.1.

Table 3.1. The Contingency table for stratum *h* required CMH relative risk estimates[1].

Treatment Group	Health Status		Total
	Condition Absent (0)	Condition Present (1)	
Vaccine (1)	n_{h11}	n_{h12}	$n_{h1.}$
Placebo (0)	n_{h21}	n_{h22}	$n_{h2.}$
Total	$n_{h.1}$	$n_{h.2}$	n_h

[1] $h = 1, \ldots, H$.

Specifically, the estimator of the common relative risk (ϕ) for Condition Present (column 2) in $h = 1, \ldots, H$ stratified 2×2 tables is given by (Rothman & Boice, 1979)

$$\hat{\phi}_{CMH} = \left(\sum_{i=1}^{h} n_{h12} n_{h2.}/n_h\right) \Big/ \left(\sum_{i=1}^{h} n_{h22} n_{h1.}/n_h\right) \tag{3.3}$$

which is a within-stratum comparison, where the treatment effect is calculated within each stratum and then combined for an overall effect.

This estimator remains consistent both in sparse-stratum (*h* increases with *n*) or large-stratum (*h* fixed but *n* increases) asymptotes (Agresti & Hartzel, 2000). Then, it provides an excellent approximation to the maximum likelihood estimator, and this approach does not assume a large sample size compared to the number of strata. The strata size does not need to be equal, and even if some cell counts are extremely small or zero, the CMH estimator remains well-defined. Therefore, it should give valid results even in some situations with small numbers in strata (Kim et al., 2020). Under standard asymptotic, the CMH estimator is preferred since it's very efficient and practical to use.

To construct the confidence interval for the common ϕ, Greenland and Robins (1985) variance estimate for log (ϕ) is used. That is,

$$\hat{\sigma}^2 = \widehat{Var}(\log(\hat{\phi}_{CMH})) = \frac{\sum_{i=1}^{h}(n_{h1.}n_{h2.}n_{h.2} - n_{h12}n_{h22}n_h)/n_h^2}{(\sum_{i=1}^{h} n_{h12}n_{h2.}/n_h)/(\sum_{i=1}^{h} n_{h22}n_{h1.}/n_h)}$$

(3.4)

Therefore, the 100(1- α)% confidence limits for the common relative risk is

$$(\hat{\phi}_{CMH} \times \exp(-z\hat{\sigma}), \hat{\phi}_{CMH} \times \exp(z\hat{\sigma}))$$

(3.5)

The CMH approach is not recommended for multi-way tables, which have more than two levels of the variables being analyzed. For this case, CMH can account for these covariates by forming strata from all combinations of covariates (including center) in the analysis, then estimating the treatment effect within each stratum. However, including many covariates may reduce power. Furthermore, this can easily lead to large strata levels, which increases the chances of some strata being dropped from the analysis because only one type of event is observed in a particular stratum. Another limitation is the CMH test assumes that the treatment effect is the same for all levels of the stratifying variable, which may not always be the case, especially since the stratifying variable is a combination of several covariates. Meanwhile, the test also assumes that the subjects are independent, which may not be true in certain situations.

3.2.2 Generalized Linear Mixed Models (GLMMs) with Delta Method

Although logistic regression is suitable for analyzing common outcomes with adjustment of other covariates, in public health, the focus is often on estimating the relative risk. Logistic regression for binary outcome data typically gives an adjusted odds ratio (OR) estimate, not a relative risk. Only for studies of rare disease (< 10 %), logistic regression is designed to provide an adjusted odds ratio that approximates the adjusted relative risk (McNutt, 2003a). Sometimes, it could be incorrect when applied to the common disease or intervention study in animal health.

Zhang and Yu (1998) have proposed a method to convert *ORs* to *RRs* by equation $RR = OR/([1 - p_0] + [p_0 \times OR])$, where *OR* and p_0 are estimated directly from the logistic regression model. They also applied this equation to the lower and upper confidence limits of ORs to correct the lower and upper limits of the confidence interval for *RRs*. However, some authors argued that the confidence interval proposed by this method would be too narrow (Localio et al., 2001; McNutt, 2003b). Another option is the log-binomial regression model, which replaces the logit link with a log link in GLMM. This model estimates relative risk directly by exponentiation of the estimated regression coefficient.

In this chapter, our interest is a conditional model of estimating PF (i.e., 1 - *RR*), and the corresponding confidence intervals are based on the delta method with logistic regression. For a binary response variable *Y* and treatment variable *X* (0, control; 1, treated), let $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$. Then, a random intercept in a mixed effects logistic regression model can be written as:

$$logit\left(\pi_{ij}\right) = log\frac{\pi_{ij}}{1 - \pi_{ij}} = \beta x_{ij} + u_j \tag{3.6}$$

where u_j is the effect of the j^{th} stratum generally assumed to follow a normal distribution with mean 0 and variance σ^2. The parameter σ^2 provides a summary of the variation among strata, and we assume a common β across strata.

Generally, random effects assume a large sample size and have not been evaluated in settings with unbalanced cases of small sample sizes. Also, it assumes that the stratum effects follow a normal distribution, which may be an unrealistic assumption.

3.2.3 Generalized estimating equations (GEEs)

The GEE log-binomial regression model is the suited marginal model applied in RCTs with binary outcomes, and it produces an unbiased estimate of the adjusted relative risk in general (McNutt, 2003b). However, since the probability of the outcomes must fall within [0,1], in some situations, the log-binomial regression model does not converge to provide parameter estimates (Localio et al., 2007). Another potential cause of failure to converge may be software programs utilizing insufficient default convergence criteria. This issue can be addressed by demanding more iterations during the modeling fitting process (McNutt et al., 2003; Skove et al., 1998).

In contrast to the log-binomial regression model, the Poisson regression model has no difficulty with convergence (McNutt, 2003b). Generally, the GEE log-binomial and log Poisson models take the same form, assuming either a binomial or Poisson distribution for outcomes, which is written as:

$$log\left(\pi_{ij}\right) = \beta x_{ij} + u_j \tag{3.7}$$

The relative risk is then given by $\exp(\hat{\beta})$. Poisson regression without robust error variances is commonly not suggested when applied to binomial data because it may result in a conservative confidence interval due to overestimated error for relative risk (Fang, 2011). Zou (2004) has proposed a "modified Poisson" method to estimate the *RR* using a robust error variance procedure known as sandwich estimation. Typically, the first step is to make a misspecification of the distribution for Poisson regression when the underlying data are binomially distributed. Then, robust sandwich SEs for the appropriate correction could be given by misspecification of a working correlation structure in the model (e.g., independent, exchangeable). In SAS software, sandwich error estimation can be implemented using the PROC GENMOD procedure with the REPEATED statement. Since it produces unbiased SE estimates, GEE models with robust

variance estimators are more likely to be employed. The only problem would be that this model relies on asymptotic theory that needs to assume a large number of strata. Thus, when the number of strata is small, GEEs with robust SEs lead to an inflated type I error rate (Kahan et al., 2016; Mancl & DeRouen, 2001).

Table 3.2. Working Correlation Matrix for robust sandwich error estimator.

Type	Working Correlation Structure	Estimator[1]
Independent	$Corr(y_{ij}, y_{ik}) = \begin{cases} 1, j = k \\ 0, j \neq k \end{cases}$	The working correlation is not estimated in this case.
Exchangeable	$Corr(y_{ij}, y_{ik}) = \begin{cases} 1, j = k \\ \alpha, j \neq k \end{cases}$	$\hat{\alpha} = \frac{1}{(N^* - p)\phi} \sum_{i=1}^{K} f_i \sum_{j<k} e_{ij}\, e_{ik}$ $N^* = 0.5 \sum_{i=1}^{K} f_i e_{ij} e_{ik}$

[1] The dispersion parameter is ϕ estimated by $\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^{K} f_i \sum_{j=1}^{n_i} e_{ij}^2$, where $N = \sum_{i=1}^{K} f_i n_i$ is the total number of measurements and p is the number of regression parameters.

3.3 Preliminary Study

The aims of this preliminary study were: (*i*) to compare the relative performance of three methods for estimating PF (*i.e.*, 1 – *RR*); and (*ii*) to make recommendations about which method(s) should be used in practice. In this study, we considered a single-site RCT with one stratifying variable (*e.g.,* litter) where subjects were randomized to a treatment or control. To cover a variety of study designs, we varied several parameters, including the number of strata, strata sizes, SD of strata, treatment effect (PF), and treatment group allocation ratio, as summarized in Table 3.3.

Table 3.3. Catalogue of preliminary study

Design	Scenario	True PF[1]	Strata		Treatment ratio		Total Sample size
			(Size, Level)	SD[2]	Global	Strata	
Balanced	1-9	0, 1/3, 2/3	(7, 4)	0, 0.25, 0.5	3:4	3:4	28
	19-27		(5, 4)		2:3	2:3	20
	37-45		(3, 6)		1:2	1:2	18
	55-63		(5, 6)		2:3	2:3	30
Unbalanced	10-18		(7, 4)		1:1	3:4	28
	28-36		(5, 4)			2:3	20
	46-54		(3, 6)			1:2	18
	64-72		(5, 6)			2:3	30
Balanced	73-81		(8, 4)		1:1	1:1	32
	81-90		(6, 6)				36

[1] PF: prevented fraction = 1- relative risk.
[2] SD: standard deviation of strata.

3.3.1 Data Generating Mechanism

For each scenario, we generated 2000 hypothetical trial datasets for preliminary study. The datasets for each combination of parameters were generated from the model (3.5), which only considers stratum variance without individual-level error.

3.3.2 Performance Measures

We applied three statistical models (CMH, GLMM and GEE) to each simulated dataset. For each simulation scenarios and model, we estimated the following measures:

1) Convergence rate
2) Type I error rate (when the true PF is 0) and Power (when the true PF is > 0)
3) Average Point estimated PF with simulation standard deviation (*SD*)
4) Bias and mean squared error (*MSE*) of the point estimator of PF
5) Empirical coverage rate and length of the 95% confidence intervals (*CIs*) around PF

The model was classified as a convergence failure when we received an error or warning message indicating the analysis did not converge. In calculating performance, we included only the simulations in which the model successfully converged. Type I error rate and power were calculated as the proportion of the simulation results with a statistically significant treatment effect with a two-sided significance level of 5%. The mean value of the estimated treatment effect was calculated as PF (1 - *RR*). Bias is the difference between the expected value of the estimator and its true values, and is

$$Bias\ (\widehat{PF}) = E\ (\widehat{PF}) - PF = E\ (\widehat{PF} - PF).$$

(3.8)

Negatively or positively biased estimators lead to an under- or over-estimation of the true PF. Therefore, if an estimator for PF satisfies $E\ (\widehat{PF}) = PF$. That is, it is unbiased. A good estimator should not only be unbiased, but also remain unaffected as much as possible by sampling fluctuation. MSE is the squared distance between the estimator and its true value:

$$MSE\ (\widehat{PF}) = E\ (\widehat{PF} - PF)^2 = Var\ (\widehat{PF}) + (Bias\ (\widehat{PF}))^2.$$

(3.9)

If $MSE\ (\widehat{PF}_1) < MSE\ (\widehat{PF}_2)$, then $\widehat{PF}_1$ is said to be more efficient than $\widehat{PF}_2$. At last, 95% CIs produced by different methods for PF are compared in terms of coverage rate (i.e. the proportion of the simulation results in which the estimated 95% CI contained the true value of the PF) and length. Methods that provide narrower CIs with coverage probability close to the nominal level are preferable. All datasets were simulated and analyzed in SAS software (version 9.4).

3.3.3 Results and Discussion

Results for convergence rate, estimated PF with one SE, type I error, power, and coverage and length of 95% CIs are shown in Table 3.4-3.6 and Figure 3.1-3.6, respectively. Here, we summarize the key results of the preliminary study.

We encountered issues with model convergence with the GLMM approach in most scenarios, especially when the sample size is extremely small (total sample size only 18 or 20). Even for a bigger sample size (32 or 36) with a balanced treatment ratio, the convergence rate could only reach approximately 0.6 ~ 0.8. In general, CMH and GEE had high convergence rates. However, in some cases, when the event proportion is high (*i.e.,* 0.9) and the sample size is small with no treatment effect, CMH tends to encounter convergence issues (no solution for PF). Kim et al. (2020) found the same results as ours, and they stated that the reason for this issue was a skewed subject distribution across strata and when both the number of strata and total subjects were small.

Overall, the CMH approach performed well in most scenarios and performed as well or better than GLMM and GEE methods in terms of the nominal type I error rate, power, and unbiased estimates with fairly CIs in most scenarios. The only exception was when there were four strata with an inconsistent treatment ratio of 3:4 (total sample size is 28), the CMH approach gave a slightly inflated Type I error rate and low coverage rate. Our results kept consistence with the finding by Tian et al. (2019), that CMH did not perform well when treatment allocations were imbalanced across strata. Overall, the GLMM logistic regression model provided unbiased estimates of the treatment effect, but with low type I error rate and power. In the contrary, studies by Kahan (2014), and Pedroza and Truong (2017) have shown that mixed-effects logistic regression models provide unbiased estimates of the treatment effect with good power.

When the treatment ratio was consistent across strata (1:1, 1:2, 2:3, or 3:4), GEE with robust SEs could always provide an unbiased estimate with a narrow 95% CI. However, the robust SEs led to the inflated Type I error and low coverage rate since we only have a small number of strata. Also, the imbalanced data with varied treatment ratios may mess up the *RR* and *SEs* estimates, resulting in a biased value with a negative infinity lower bound.

3.3.4 Conclusion

In this preliminary study, all models were investigated only under homogeneous scenarios. For this case, CMH is always recommended as the first and best approach. For small sample with unbalanced arms, the GLMM model has structural problems (*e.g.,* probabilities outside the [0, 1] interval), so the estimates are particularly useful for logit link functions. Therefore, given its good performance and their ease of computation, one might consider always using CMH instead of other two estimators.

However, for large-stratum asymptotes (fixed h), CMH estimators may lose some efficiency compared to ML estimator (GLMM) (Agresti & Hartzel, 2000). Moreover, when interaction effect exists, a complexity of model fitting should be considered. We will discuss more in next section.

Table 3.4. Type I error and power for scenarios when strata number is 4 and 6 of balanced and unbalanced design[1].

Strata Number	Design		Type I error			Power					
			0.000			0.333			0.667		
			CMH	GEE	GLMM	CMH	GEE	GLMM	CMH	GEE	GLMM
4	Balanced	2:3	0.015	0.160	0.000	0.393	0.480	0.082	0.797	0.794	0.535
		3:4	0.030	0.124	0.001	0.477	0.584	0.230	0.930	0.897	0.801
		4:4	0.030	0.161	0.001	0.462	0.571	0.294	0.947	0.905	0.848
	Unbalanced	2:3	0.018	0.128	0.001	0.273	0.458	0.069	0.661	0.736	0.297
		3:4	0.136	0.202	0.013	0.471	0.556	0.207	0.841	0.811	0.678
6	Balanced	1:2	0.019	0.181	0.000	0.454	0.432	0.038	0.777	0.801	0.485
		2:3	0.036	0.116	0.000	0.491	0.537	0.236	0.945	0.922	0.867
		3:3	0.041	0.121	0.005	0.536	0.589	0.377	0.971	0.951	0.926
	Unbalanced	1:2	0.017	0.080	0.003	0.209	0.329	0.040	0.586	0.638	0.231
		2:3	0.045	0.168	0.013	0.424	0.559	0.218	0.856	0.853	0.723

[1] Methods: CMH, Cochran Mantel Haenszel; GLMM, generalized linear mixed model; GEE, Generalized estimating equation.

Table 3.5. Coverage rate and length of 95% CIs for scenarios when strata number is 4 of balanced and unbalanced design[1].

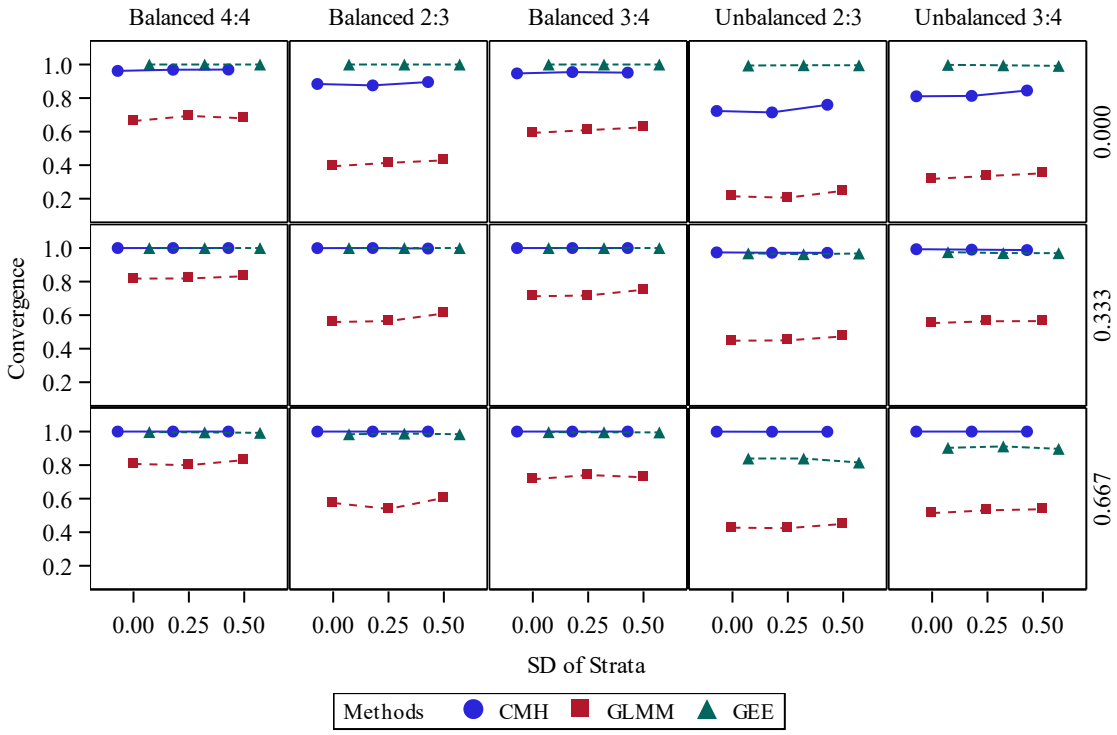
True PF	Design		CMH		GEE		GLMM	
			Coverage	Length	Coverage	Length	Coverage	Length
0.000	Balanced	2:3	0.985	0.700	0.840	0.552	1.000	0.892
		3:4	0.970	0.544	0.876	0.446	0.999	0.618
		4:4	0.970	0.491	0.839	0.408	0.999	0.533
	Unbalanced	2:3	0.982	0.813	0.872	1.040×10^{95}	0.999	1.063
		3:4	0.864	0.614	0.798	1.190×10^{74}	0.987	0.758
0.333	Balanced	2:3	0.948	0.746	0.847	0.651	0.932	0.948
		3:4	0.934	0.616	0.827	0.531	0.938	0.719
		4:4	0.936	0.597	0.828	0.513	0.933	0.671
	Unbalanced	2:3	0.890	0.820	0.644	4.030×10^{95}	0.900	1.094
		3:4	0.816	0.656	0.527	4.500×10^{153}	0.830	0.812
0.667	Balanced	2:3	0.929	0.703	0.870	0.616	0.938	0.938
		3:4	0.941	0.571	0.825	0.494	0.948	0.704
		4:4	0.937	0.560	0.836	0.488	0.945	0.674
	Unbalanced	2:3	0.750	0.816	0.733	5.860×10^{95}	0.835	1.058
		3:4	0.808	0.629	0.738	1.900×10^{130}	0.863	0.770

[1]Methods: CMH, Cochran Mantel Haenszel; GLMM, generalized linear mixed model; GEE, Generalized estimating equation.

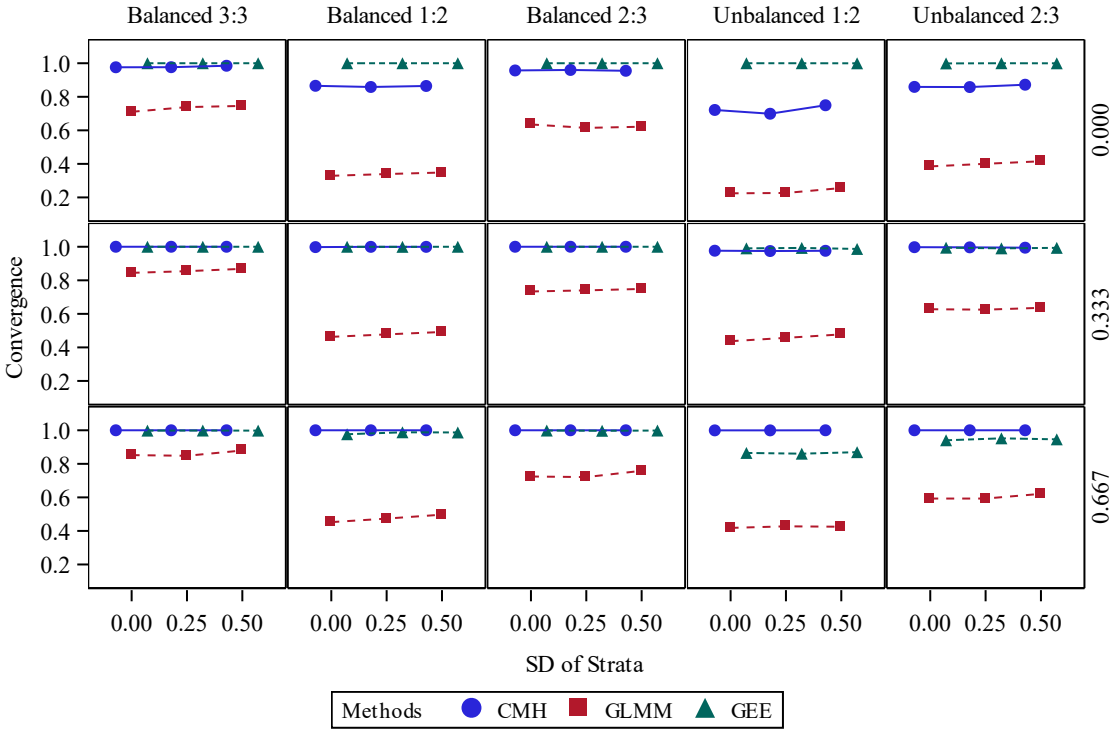
Table 3.6. Coverage rate and length of 95% CIs for scenarios when strata number is 6 of balanced and unbalanced design[1].

True PF	Design		CMH		GEE		GLMM	
			Coverage	Length	Coverage	Length	Coverage	Length
0.000	Balanced	1:2	0.959	0.455	0.880	0.407	0.995	0.484
		2:3	0.981	0.814	0.819	0.686	1.000	1.230
		3:3	0.964	0.527	0.884	0.468	1.000	0.602
	Unbalanced	1:2	0.983	0.942	0.920	0.643	0.997	1.132
		2:3	0.955	0.568	0.832	0.724	0.987	0.676
0.333	Balanced	1:2	0.942	0.558	0.880	0.515	0.942	0.613
		2:3	0.955	0.797	0.920	0.759	0.949	1.121
		3:3	0.942	0.592	0.891	0.551	0.952	0.675
	Unbalanced	1:2	0.858	0.945	0.759	0.855	0.881	1.201
		2:3	0.844	0.639	0.707	2.730×10^{85}	0.830	0.745
0.667	Balanced	1:2	0.947	0.524	0.914	0.490	0.953	0.592
		2:3	0.931	0.727	0.932	0.692	0.954	0.988
		3:3	0.942	0.537	0.910	0.499	0.944	0.631
	Unbalanced	1:2	0.751	0.964	0.825	0.847	0.838	1.145
		2:3	0.829	0.607	0.808	1.318	0.856	0.710

[1]Methods: CMH, Cochran Mantel Haenszel; GLMM, generalized linear mixed model; GEE, Generalized estimating equation.

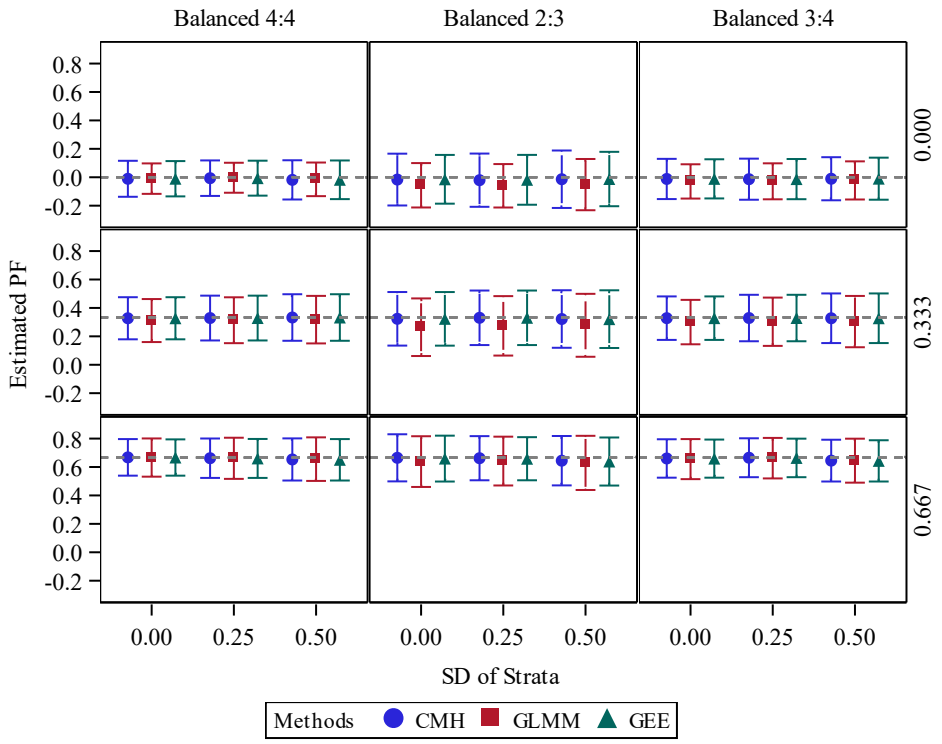


(A)

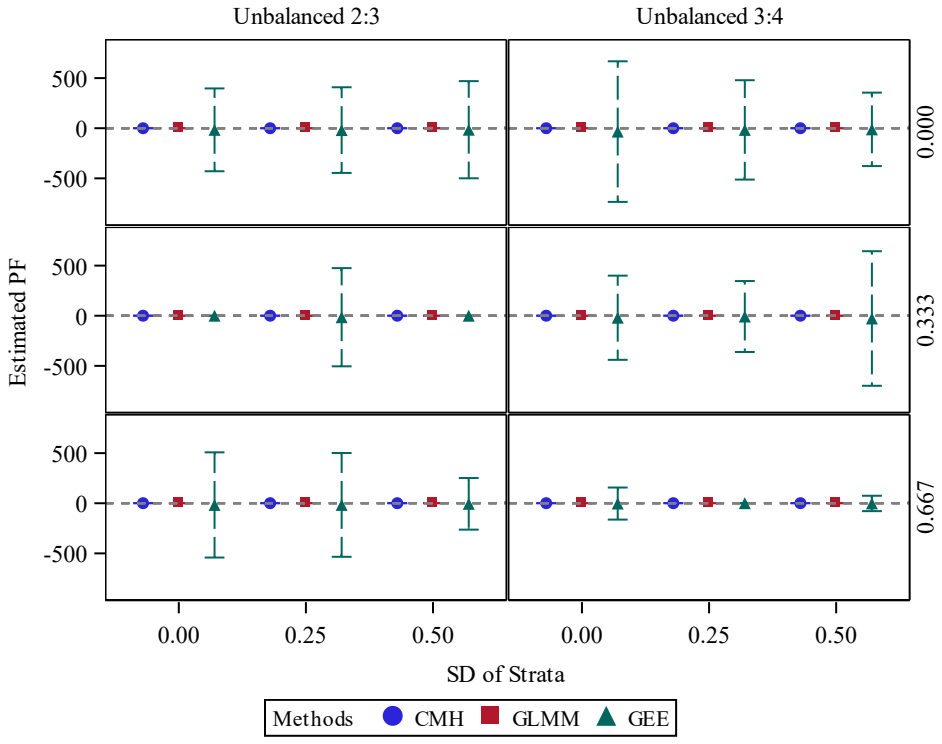


(B)

Figure 3.1. Convergence for scenarios when the true PF was 0, 1/3, and 2/3; the standard deviation (SD) of strata is 0, 0.25, and 0.50. (A) strata number of 4; (B) strata number of 6.



(A)



(B)

Figure 3.2. Estimated PF with one standard deviation for scenarios when the true PF was 0, 1/3, and 2/3; the standard deviation (SD) of strata is 0, 0.25, and 0.50; and strata number of 4. (A) balanced; (B) unbalanced.

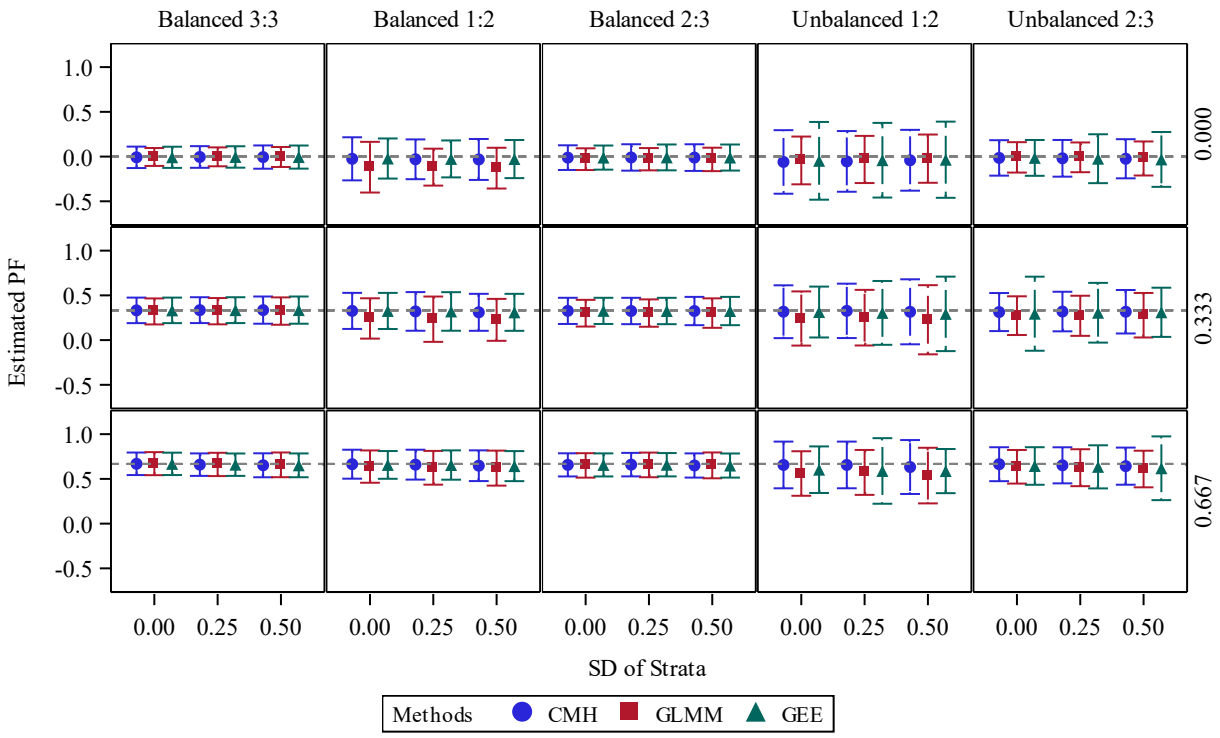


Figure 3.3. Estimated PF with 95% CIs for scenarios when the true PF was 0, 1/3, and 2/3; the standard deviation (SD) of strata is 0, 0.25, and 0.50; and strata number of 6.

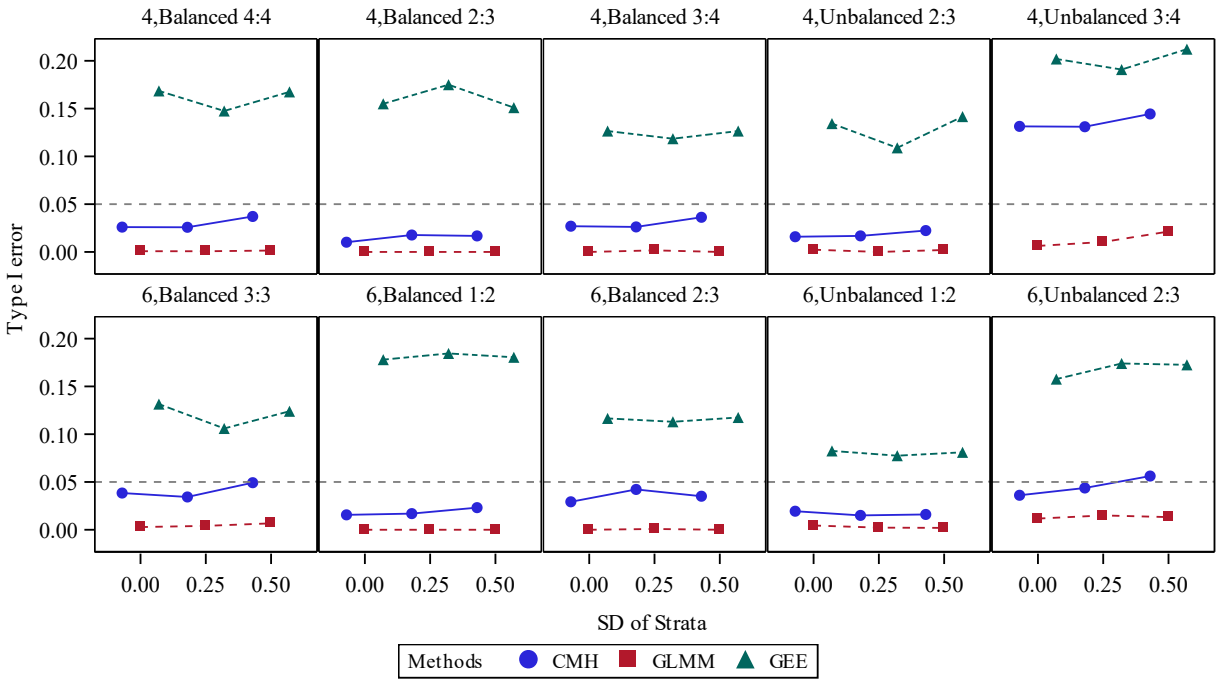
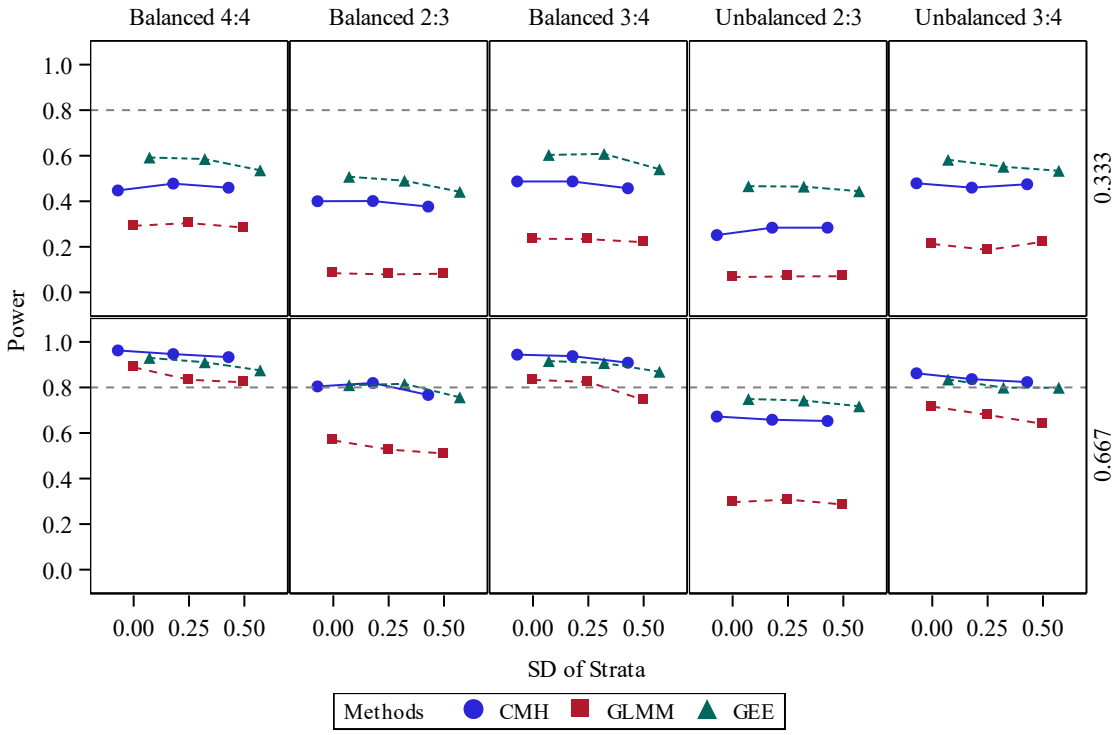
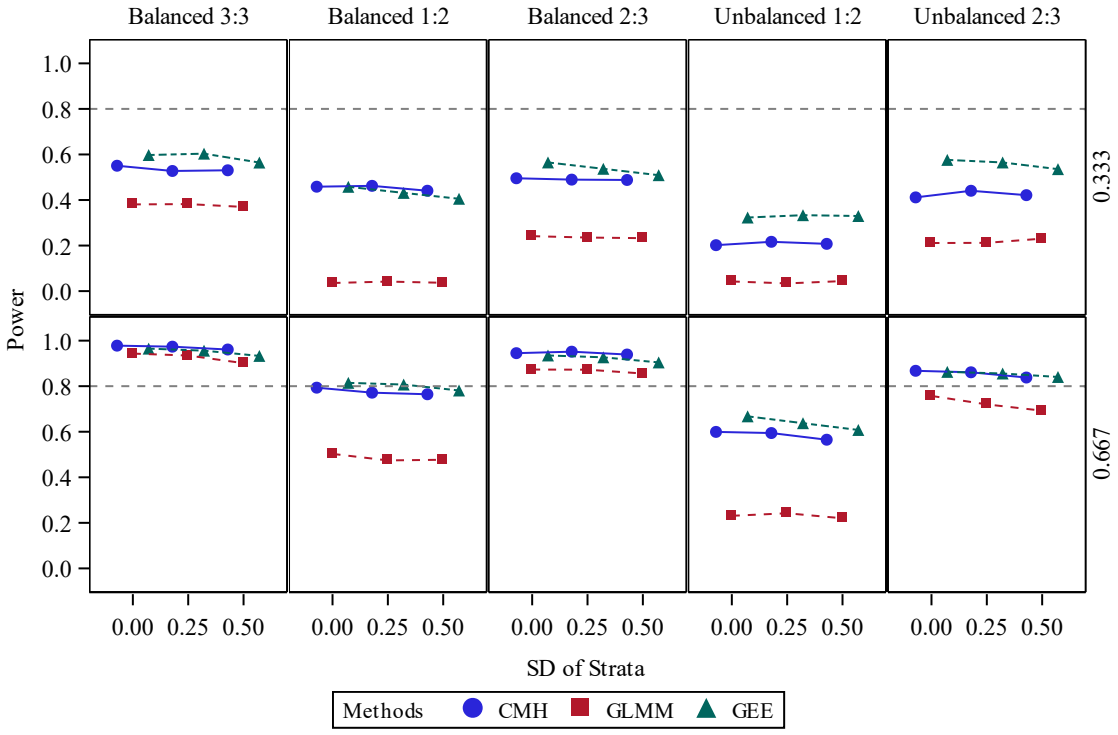


Figure 3.4. Type I error rate for scenarios when the standard deviation (SD) of strata is 0, 0.25, and 0.50; and strata number of 4 and 6.

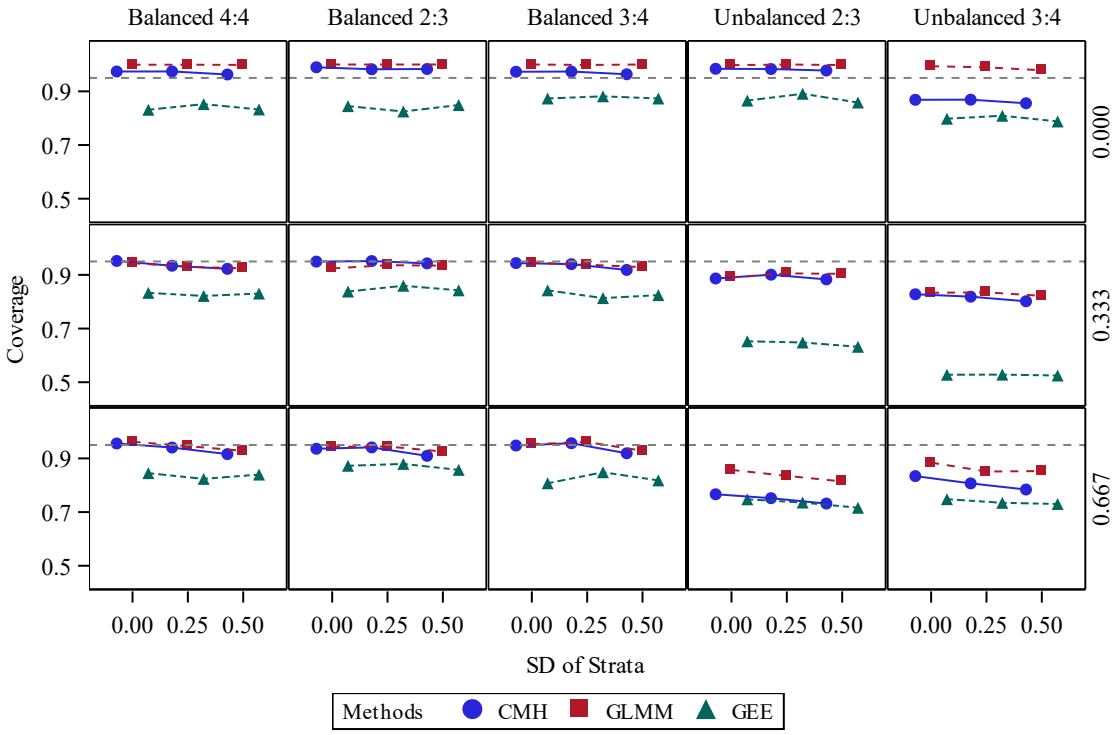


(A)

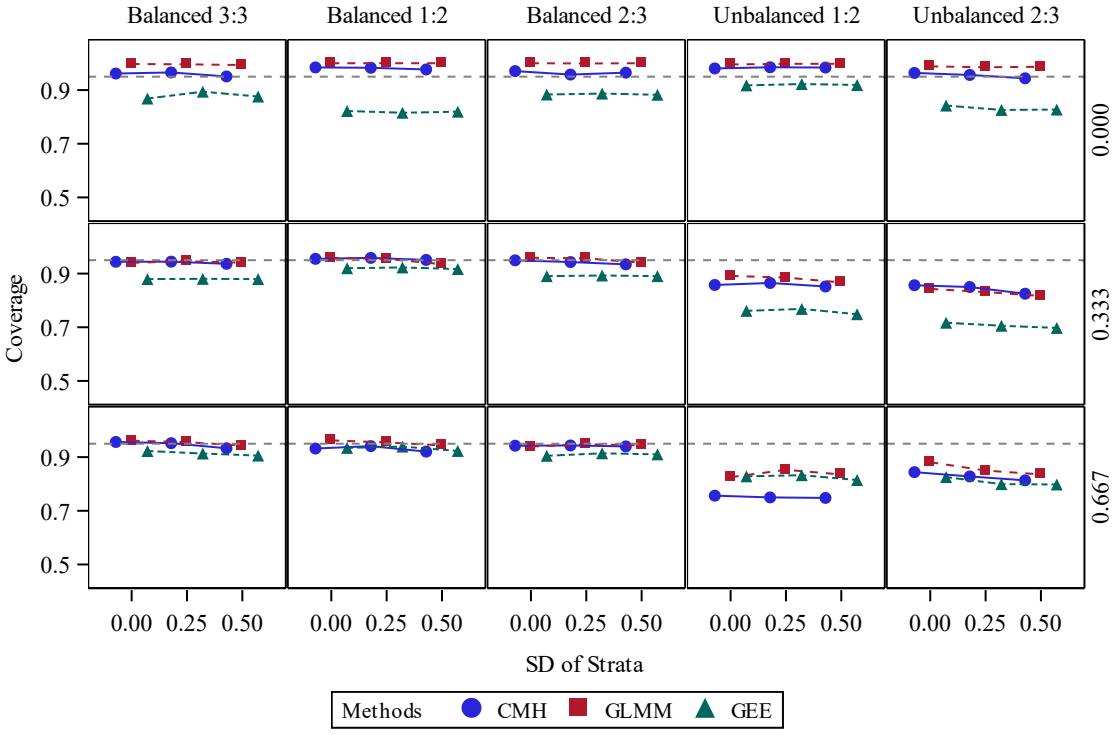


(B)

Figure 3.5. Power for scenarios when the true PF was 1/3 and 2/3; the standard deviation (SD) of strata is 0, 0.25, and 0.50. (A) strata number of 4; (B) strata number of 6.



(A)



(B)

Figure 3.6. Coverage for scenarios when the true PF was 0, 1/3, and 2/3; and the standard deviation (SD) of strata is 0, 0.25, and 0.50. (A) strata number of 4; (B) strata number of 6.

Table 3.7. Bias and MSE of point estimated PF for scenarios when strata number is 4 of balanced and unbalanced design[1].

True PF[2]	Design		CMH		GEE		GLMM	
			Bias	MSE	Bias	MSE	Bias	MSE
0.000	Balanced	2:3	-0.017	0.037	-0.015	0.032	-0.056	0.030
		3:4	-0.012	0.021	-0.011	0.020	-0.027	0.017
		4:4	-0.012	0.017	-0.012	0.016	-0.009	0.012
	Unbalanced	2:3	-0.038	0.088	-16.54	1.961×10^5	-0.020	0.049
		3:4	-0.026	0.054	-20.53	2.908×10^5	-0.018	0.035
0.333	Balanced	2:3	-0.008	0.038	-0.008	0.038	-0.061	0.048
		3:4	-0.005	0.027	-0.005	0.027	-0.031	0.030
		4:4	-0.004	0.025	-0.004	0.025	-0.019	0.026
	Unbalanced	2:3	-0.012	0.084	-5.122	7.998×10^4	-0.093	0.098
		3:4	-0.014	0.067	-18.73	2.510×10^5	-0.065	0.079
0.667	Balanced	2:3	-0.010	0.027	-0.015	0.026	-0.030	0.033
		3:4	-0.010	0.020	-0.011	0.019	-0.013	0.022
		4:4	-0.006	0.019	-0.007	0.019	-0.006	0.021
	Unbalanced	2:3	-0.021	0.062	-13.77	2.032×10^5	-0.107	0.075
		3:4	-0.010	0.044	-2.150	1.049×10^4	-0.058	0.048

[1] Methods: CMH, Cochran Mantel Haenszel; GLMM, generalized linear mixed model; GEE, Generalized estimating equation.
[2] PF: prevented fraction = 1- relative risk.

Table 3.8. Bias and MSE of point estimated PF for scenarios when strata number is 6 of balanced and unbalanced design[1].

True PF[2]	Design		CMH		GEE		GLMM	
			Bias	MSE	Bias	MSE	Bias	MSE
0.000	Balanced	1:2	-0.029	0.054	-0.025	0.047	-0.122	0.073
		2:3	-0.011	0.021	-0.010	0.020	-0.030	0.017
		3:3	-0.006	0.015	-0.006	0.015	-0.004	0.011
	Unbalanced	1:2	-0.052	0.122	-0.041	0.183	-0.033	0.072
		2:3	-0.020	0.043	-0.023	0.070	-0.013	0.031
0.333	Balanced	1:2	-0.014	0.044	-0.015	0.044	-0.099	0.066
		2:3	-0.008	0.023	-0.008	0.023	-0.032	0.025
		3:3	0.001	0.021	0.001	0.021	-0.011	0.022
	Unbalanced	1:2	-0.013	0.104	-0.030	0.128	-0.094	0.121
		2:3	-0.017	0.051	-0.029	0.120	-0.059	0.056
0.667	Balanced	1:2	-0.010	0.028	-0.015	0.027	-0.039	0.037
		2:3	-0.011	0.018	-0.012	0.017	-0.014	0.020
		3:3	-0.007	0.017	-0.007	0.016	-0.004	0.018
	Unbalanced	1:2	-0.018	0.076	-0.074	0.093	-0.110	0.086
		2:3	-0.014	0.040	-0.034	0.078	-0.043	0.042

[1] Methods: CMH, Cochran Mantel Haenszel; GLMM, generalized linear mixed model; GEE, Generalized estimating equation.
[2] PF: prevented fraction = 1- relative risk.

3.4 Multicenter RCTs

Previous researches have shown that when the center factor was considered in the randomization process, it should always be accounted for in the final analysis (Kahan, 2014; Kahan et al., 2016; Liang & Zeger, 1986; Parzen et al., 1998). For binary outcomes, adjusting for center effects in the analysis is impractical when there are too few patients or events per center. Meanwhile, most researchers paid attention to analysis methods under homogeneous scenarios, which did not consider the issue of treatment-by-center interaction. This simulation study aims to clarify (*i*) under what circumstances it is beneficial to account for center effects in the analysis and (*ii*) the best method of adjustment for the center in RCTs with a binary outcome.

3.4.1 Data Generating Mechanism

We generated 10,000 hypothetical trial datasets for each setting, expecting small errors due to the large number of repetitions. The choice of 10,000 repetitions was justified by sufficiently low Monte Carlo errors. The datasets for each combination of parameters were generated from the following model,

$$y_{ijk} \sim Bernoulli(\pi_{ijk})$$

$$logit\left(\pi_{ijk}\right) = log\frac{\pi_{ijk}}{1-\pi_{ijk}} = \beta x_{ijk} + u_j + s_{ij} + \varepsilon_{ijk} \tag{3.10}$$

where y_{ijk} is a binary outcome generated from a Bernoulli distribution (π_{ijk}). Meanwhile, u_j is the random effect of the j^{th} stratum (*i.e.,* Center), which is generally assumed to follow a normal distribution with mean 0 and variance σ_c^2. The parameter σ_c^2 provides a summary of the variation among strata, and we assume a common β across strata. s_{ij} represents a random center-by-treatment interaction effect, which was generated from a normal distribution with a mean of 0 and standard deviation σ_{tc}^2, and ε_{ijk} was a random error from the standard logistic distribution.

To thoroughly study the performance of candidate models at various ICC levels, we considered the following values of ICC for completeness: 0.00, 0.25, 0.50, and 0.75. However, we focused interpretation of the results on lower values of ICC and σ_{tc}^2 as they were more likely to occur in practice. Additionally, to cover a variety of study designs, we varied several other parameters, including the number of centers, center sizes, treatment effect (PF), and treatment-by-center random effect (homogeneity and heterogeneity), as summarized in Table 3.9.

Table 3.9. Catalogue of simulation study

Design	Scenario	σ_{tc}^2	True PF[1]	Center		ICC
				(Size, Level)	σ_c^2	
Homogeneity (i.e. $\sigma_{tc}^2 = 0$)	1-4	0	0, 0.45	(20, 5), (10, 25)	0	0
	5-8				1.10	0.25
	9-12				3.27	0.50
	13-16				9.86	0.75
Heterogeneity (i.e. $\sigma_{tc}^2 > 0$)	17-20	0.4	0, 0.45	(20, 5), (10, 25)	0.40	0
	21-24				1.63	0.25
	25-28				4.09	0.50
	29-32				11.46	0.75
	33-36	0.8	0, 0.45	(20, 5), (10, 25)	0.80	0
	37-40				2.16	0.25
	41-45				4.89	0.50
	46-48				13.06	0.75

[1] PF: prevented fraction = 1- relative risk.

3.4.2 Performance Measures

We applied seven statistical models (shown in Table 3.10) to each simulated dataset. For the center effect, both unadjusted and adjusted analyses were performed. For GLMM, an additional model incorporating interaction random effect was applied (GLMM-ALL). GEE with robust variance estimates (GEE-Center) was under log Poisson regression using REPEATED statement with SUBJECT = CENTER. Three additional GEE models, including the center effect as fixed, were also explored without robust variance estimates, or with robust variance estimates

(Center/Center-by-Treatment) in the first 2000 simulated dataset. Since the performances of these three models were not as well as expected, the results are only shown in Appendix A.

For each simulation scenario and model, we estimated the following measures:

1) Average Point estimated PF with simulation standard deviation (*SD*)
2) Bias and mean squared error (*MSE*) of the point estimator of PF
3) Type I error rate (when the true PF is 0) and Power (when the true PF is > 0)
4) Empirical coverage rate and length of the 95% confidence intervals (*CIs*) around PF.

The details of explanation as the same as section 3.3.2. All datasets were simulated and analyzed in SAS software (version 9.4).

Table 3.10. Descriptions of ten models.

Method[2]	Model[1]	Note[3]
Simple	-	
CMH	-	
Unadjusted GLMM	$logit\left(\pi_{ij}\right) = \beta x_{ij}$	
GLMM-Center	$logit\left(\pi_{ij}\right) = \beta x_{ij} + u_j$	u_j are random
GLMM-All	$logit\left(\pi_{ij}\right) = \beta x_{ij} + u_j + s_{ij}$	u_j and s_{ij} are random
Unadjusted GEE	$log\left(\pi_{ij}\right) = \beta x_{ij}$	without WCS
GEE-Center	$log\left(\pi_{ij}\right) = \beta x_{ij} + u_j$	u_j are random, with WCS=center
GEE* without Robust SEs	$log\left(\pi_{ij}\right) = \beta x_{ij} + u_j$	u_j are fixed, without WCS
GEE*-Center	$log\left(\pi_{ij}\right) = \beta x_{ij} + u_j$	u_j are fixed, with WCS=center
GEE*-Center by Treatment	$log\left(\pi_{ij}\right) = \beta x_{ij} + u_j + s_{ij}$	u_j are fixed, with WCS=center-by-treatment

[1] $y_{ijk} \sim Bernoulli(\pi_{ijk})$, where y_{ijkl} represents a binary outcome measured for the *l*th subject in the *k*th litter and *j*th center receiving the *i*th treatment (control or treatment group).
[2] Methods: CMH, Cochran Mantel Haenszel; GLMM, generalized linear mixed model; GEE, Generalized estimating equation.
[3] WCS, working correlation structure.

3.4.3 Results and Discussion

Simulation results are shown in Table 3.11. – 3.18. the bias of all point estimated PF decreased with a larger center level (even with a similar total sample size, 200 vs. 250), especially for GLMM methods. Furthermore, the more precise point estimates across 10,000 Monte Carlo simulations led to the smaller empirical *SD* and overall error rate (measured by *MSE*) of the estimator PF in all scenarios. Moreover, all four of simple, the CMH and GEE analysis methods gave identical estimates of treatment effect PF matter adjusted for center effect or not. Thus, these same point estimates across 10,000 simulations result in the identical empirical *SD* and overall error rate (measured by *MSE*) of the estimator PF in these four Models within scenarios. When true PF was zero, four of these estimators were also unbiased estimates of treatment effect PF, regardless of σ_{tc}^2 values. However, as ICC increased, the estimates PF for all four methods slightly decreased.

On the contrary to nonzero true PF scenarios, as ICC and σ_{tc}^2 increased, four CMH and GEE methods underestimated the point estimates of PF; in this case, all these techniques were slightly and downwardly biased. When GLMM approaches were applied, results from the model ignoring center and interaction random effect (unadjusted GLMM) were shown that the point estimate of PF was severely underestimated for true PF was zero and contrarily overestimated for true PF was nonzero. Similarly, the other two GLMMs only account for the center effect or account for center and interaction effects, which have this under/overestimated issue for treatment effect PF, but only slightly.

Results for Type I error rate and power analysis of hypothesis testing for treatment effect PF under both homogeneous and heterogeneous scenarios are shown in Figure 3.7. – 3.8. Since the length of 95% confidence intervals was expected to narrow as the number of centers increased,

it indicates that no matter how each model performs, larger the number of centers could improve overall performance for inference of treatment effect. As expected, the log-Poisson regression without robust sandwich SEs (unadjusted GEE) produces very conservative confidence intervals for PF, leading to failure to maintain the nominal level of Type I error rate and not power efficiency. When the treatment effect was homogeneous across centers (σ_{tc}^2 was zero), CMH performed better than other models in terms of close to the nominal value of 5% for Type I error rate and 80% power, especially when the center number is small (only 5). However, when there was heterogeneity existed ($\sigma_{tc}^2 > 0$), CMH was not the best option for this kind of data since it could lead to the inflated Type I error rate unless there were enough large center number (> 25) in the trials. In this case, GLMM incorporating with center and interaction term was unexpected to be the most appropriate model for hypothesis testing for treatment effect PF. However, a slight decrease was observed in Type I error and power as ICC increases. If the center level was large enough (> 25), either GLMM only incorporating with center effect or GEE with center effect as Robust *SEs* estimator could be the second option for these heterogeneous scenarios.

Simulation studies have been conducted to investigate the impact of accounting for between-center variation in type I error rate and power of treatment effect under homogeneous cases. For the Type I error rate, our simulation study based on balanced scenarios found that it failed to maintain at the nominal level (~ 5%) as ICC increased. However, they disagree with those reported by Kahan and Morris (2013b) and Edgar et al. (2021), that for binary outcomes, type 1 error rates were substantially inflated when between-center variance was ignored in the analysis, particularly if there was unequally sized treatment allocation across centers. Moreover, the results showed that neglecting the center effect could lead to loss of power as ICC increased, which was consistent with the findings from Kahan and Morris (2013b). Conversely, when the

number of centers was large enough (> 25), there was only a slight statistical difference among models with or without adjustment (σ_c^2 < 4.89). Agreement can be found from studies by Edgar et al. (2021) and Lingsma et al. (2011) that accounting for the sizeable between-center variation (σ_c^2, 0.21 - 4.81) had no clinically meaningful impact on the estimated treatment effect for large, multi-center RCTs (n_c = 237).

3.4.4 Conclusion

In this chapter, seven models were investigated under homogeneous and heterogeneous scenarios. For multicenter RCTs with binary outcomes, it is crucial to consider center and center-by-treatment effects to enhance statistical power and minimize bias in treatment effect estimates and standard errors. Especially if stratified randomization by the center has been adopted in the design, the analysis should adjust for the center. Unless an exploration analysis has proven no statistical heterogeneous treatment effect across the center. When the treatment effect was homogeneous across centers, Cochran-Mantel-Haenszel incorporating the center effect as a stratification factor should always be the best choice, even when the number of centers is small. Otherwise, once heterogeneous treatment effect across centers was confirmed, GLMM incorporating with center and interaction term could be the acceptable method for a small center level (~5), while GLMM only incorporating with center effect or GEE with center effect as Robust SEs estimator could be the alternative option for a large number of centers (≥ 25).

Table 3.11. Average point estimated PF and empirical simulation standard deviation (SD) for true PF was 0 using seven models under homogeneous scenarios ($\boldsymbol{\sigma_{tc}^2}$ = 0) with different ICC values (0, 0.25, 0.5, and 0.75) and center sizes (5 *vs* 25).

		Center (size, level)			
		(20, 5)		(10, 25)	
ICC	Method[1]	Avg. $\widehat{PF}$	SD	Avg. $\widehat{PF}$	SD
0	Simple	-0.005	0.096	-0.002	0.061
	CMH	-0.005	0.096	-0.002	0.061
	Unadjusted GEE	-0.005	0.096	-0.002	0.061
	GEE-Center	-0.005	0.096	-0.002	0.061
	Unadjusted GLMM	-42.48	4236.4	-0.039	0.292
	GLMM-Center	-0.114	0.584	-0.039	0.294
	GLMM-All	-0.115	0.585	-0.039	0.294
0.25	Simple	-0.005	0.100	-0.002	0.062
	CMH	-0.005	0.100	-0.002	0.062
	Unadjusted GEE	-0.005	0.100	-0.002	0.062
	GEE-Center	-0.005	0.100	-0.002	0.062
	Unadjusted GLMM	-0.118	0.610	-0.034	0.276
	GLMM-Center	-0.117	0.604	-0.034	0.275
	GLMM-All	-0.005	0.100	-0.002	0.062
0.5	Simple	-0.007	0.108	-0.002	0.064
	CMH	-0.007	0.108	-0.002	0.064
	Unadjusted GEE	-0.007	0.108	-0.002	0.064
	GEE-Center	-0.007	0.108	-0.002	0.064
	Unadjusted GLMM	-118.7	5323.9	-0.025	0.224
	GLMM-Center	-0.129	0.640	-0.034	0.269
	GLMM-All	-0.128	0.634	-0.034	0.268
0.75	Simple	-0.005	0.112	-0.002	0.065
	CMH	-0.005	0.112	-0.002	0.065
	Unadjusted GEE	-0.005	0.112	-0.002	0.065
	GEE-Center	-0.005	0.112	-0.002	0.065
	Unadjusted GLMM	-301.1	5910.7	-0.014	0.169
	GLMM-Center	-0.141	0.817	-0.036	0.280
	GLMM-All	-0.139	0.803	-0.036	0.279

[1]Methods: Simple, based on crude RR; CMH, Cochran Mantel Haenszel-stratified by center; Unadjusted GEE, log-Poisson regression without adjustment for center and robust SEs; Unadjusted GLMM, generalized linear mixed model without random effect; GLMM-Center, generalized linear mixed model with random center effect; GLMM-All, generalized linear mixed model with random center and interaction effect.

Table 3.12. Average point estimated PF and empirical simulation standard deviation (SD) for true PF was 0.4 using six models under homogeneous scenarios ($\boldsymbol{\sigma_{tc}^2}$ = 0) with different ICC values (0, 0.25, 0.5, and 0.75) and center sizes (5 *vs* 25).

		Center (size, level)			
		(20, 5)		(10, 25)	
ICC	Method	Avg. $\widehat{PF}$	SD	Avg. $\widehat{PF}$	SD
0	Simple	0.383	0.096	0.386	0.060
	CMH	0.383	0.096	0.386	0.060
	Unadjusted GEE	0.383	0.096	0.386	0.060
	GEE-Center	0.383	0.096	0.386	0.060
	Unadjusted	0.625	0.127	0.629	0.078
	GLMM-Center	0.628	0.127	0.630	0.078
	GLMM-All	0.629	0.127	0.631	0.078
0.25	Simple	0.370	0.110	0.371	0.065
	CMH	0.370	0.110	0.371	0.065
	Unadjusted GEE	0.370	0.110	0.371	0.065
	GEE-Center	0.370	0.110	0.371	0.065
	Unadjusted	0.587	0.136	0.588	0.081
	GLMM-Center	0.612	0.136	0.607	0.081
	GLMM-All	0.613	0.136	0.608	0.081
0.5	Simple	0.346	0.122	0.345	0.068
	CMH	0.346	0.122	0.345	0.068
	Unadjusted GEE	0.346	0.122	0.345	0.068
	GEE-Center	0.346	0.122	0.345	0.068
	Unadjusted	0.529	0.153	0.526	0.085
	GLMM-Center	0.599	0.154	0.593	0.087
	GLMM-All	0.599	0.153	0.592	0.087
0.75	Simple	0.305	0.139	0.294	0.073
	CMH	0.305	0.139	0.294	0.073
	Unadjusted GEE	0.305	0.139	0.294	0.073
	GEE-Center	0.305	0.139	0.294	0.073
	Unadjusted	0.433	0.176	0.417	0.093
	GLMM-Center	0.578	0.194	0.574	0.105
	GLMM-All	0.576	0.195	0.572	0.105

[1]Methods: Simple, based on crude RR; CMH, Cochran Mantel Haenszel-stratified by center; Unadjusted GEE, Log-Poisson regression without adjustment for center and robust SEs; Unadjusted GLMM, generalized linear mixed model without random effect; GLMM-Center, generalized linear mixed model with random center effect; GLMM-All, generalized linear mixed model with random center and interaction effects.

Table 3.13. Average point estimated PF and empirical simulation standard deviation (SD) for true PF was 0 using six models under heterogeneous scenarios ($\boldsymbol{\sigma_{tc}^2}$ = 0.4) with different ICC values (0, 0.25, 0.5, and 0.75) and center sizes (5 *vs* 25).

		Center (size, level)			
		(20, 5)		(10, 25)	
ICC	Method	Avg. $\widehat{PF}$	SD	Avg. $\widehat{PF}$	SD
0	Simple	-0.007	0.113	-0.002	0.066
	CMH	-0.007	0.113	-0.002	0.066
	Unadjusted GEE	-0.007	0.113	-0.002	0.066
	GEE-Center	-0.007	0.113	-0.002	0.066
	Unadjusted GLMM	-38.27	3812.7	-0.040	0.308
	GLMM-Center	-0.149	0.692	-0.040	0.311
	GLMM-All	-0.150	0.691	-0.040	0.311
0.25	Simple	-0.006	0.119	-0.002	0.068
	CMH	-0.006	0.119	-0.002	0.068
	Unadjusted GEE	-0.006	0.119	-0.002	0.068
	GEE-Center	-0.006	0.119	-0.002	0.068
	Unadjusted GLMM	-63.67	3859.2	-0.031	0.268
	GLMM-Center	-0.146	0.765	-0.036	0.288
	GLMM-All	-0.145	0.739	-0.036	0.286
0.5	Simple	-0.007	0.127	-0.002	0.069
	CMH	-0.007	0.127	-0.002	0.069
	Unadjusted GEE	-0.007	0.127	-0.002	0.069
	GEE-Center	-0.007	0.127	-0.002	0.069
	Unadjusted GLMM	-220.4	6828.0	-0.022	0.227
	GLMM-Center	-0.159	0.787	-0.033	0.283
	GLMM-All	-0.154	0.764	-0.033	0.280
0.75	Simple	-0.008	0.135	-0.002	0.070
	CMH	-0.008	0.135	-0.002	0.070
	Unadjusted GEE	-0.008	0.135	-0.002	0.070
	GEE-Center	-0.008	0.135	-0.002	0.070
	Unadjusted GLMM	-360.4	7886.0	-0.012	0.173
	GLMM-Center	-0.194	0.982	-0.037	0.301
	GLMM-All	-0.182	0.898	-0.036	0.297

[1]Methods: Simple, based on crude RR; CMH, Cochran Mantel Haenszel-stratified by center; Unadjusted GEE, Log-Poisson regression without adjustment for center and robust SEs; Unadjusted GLMM, generalized linear mixed model without random effect; GLMM-Center, generalized linear mixed model with random center effect; GLMM-All, generalized linear mixed model with random center and interaction effects.

Table 3.14. Average point estimated PF and empirical simulation standard deviation (SD) for true PF was 0.4 using six models under heterogeneous scenarios ($\boldsymbol{\sigma_{tc}^2}$ = 0.4) with different ICC values (0, 0.25, 0.5, and 0.75) and center sizes (5 *vs* 25).

		Center (size, level)			
		(20, 5)		(10, 25)	
ICC	Method	Avg. $\widehat{PF}$	SD	Avg. $\widehat{PF}$	SD
0	Simple	0.378	0.114	0.380	0.067
	CMH	0.378	0.114	0.380	0.067
	Unadjusted GEE	0.378	0.114	0.380	0.067
	GEE-Center	0.378	0.114	0.380	0.067
	Unadjusted GLMM	0.605	0.148	0.613	0.085
	GLMM-Center	0.611	0.148	0.616	0.085
	GLMM-All	0.614	0.148	0.618	0.085
0.25	Simple	0.293	0.148	0.287	0.075
	CMH	0.293	0.148	0.287	0.075
	Unadjusted GEE	0.364	0.125	0.364	0.070
	GEE-Center	0.364	0.125	0.364	0.070
	Unadjusted GLMM	0.567	0.158	0.570	0.088
	GLMM-Center	0.598	0.160	0.596	0.089
	GLMM-All	0.600	0.159	0.596	0.089
0.5	Simple	0.341	0.135	0.336	0.072
	CMH	0.341	0.135	0.336	0.072
	Unadjusted GEE	0.341	0.135	0.336	0.072
	GEE-Center	0.341	0.135	0.336	0.072
	Unadjusted GLMM	0.512	0.172	0.507	0.093
	GLMM-Center	0.587	0.178	0.583	0.096
	GLMM-All	0.588	0.178	0.582	0.096
0.75	Simple	0.293	0.148	0.287	0.075
	CMH	0.293	0.148	0.287	0.075
	Unadjusted GEE	0.293	0.148	0.287	0.075
	GEE-Center	0.293	0.148	0.287	0.075
	Unadjusted GLMM	0.412	0.192	0.401	0.097
	GLMM-Center	0.561	0.225	0.566	0.112
	GLMM-All	0.558	0.224	0.563	0.112

[1]Methods: Simple, based on crude RR; CMH, Cochran Mantel Haenszel-stratified by center; Unadjusted GEE, Log-Poisson regression without adjustment for center and robust SEs; Unadjusted GLMM, generalized linear mixed model without random effect; GLMM-Center, generalized linear mixed model with random center effect; GLMM-All, generalized linear mixed model with random center and interaction effects.

Table 3.15. Properties of point estimated PF for true PF was 0 using seven models under homogeneous scenarios ($\boldsymbol{\sigma_{tc}^2}$ = 0) with different ICC values (0, 0.25, 0.5, and 0.75) and center sizes (5 *vs* 25).

		Center (size, level)					
		(20, 5)			(10, 25)		
ICC	Method	Bias	MSE	Avg.length	Bias	MSE	Avg.length
0	Simple	-0.005	0.009	0.378	-0.002	0.004	0.237
	CMH	-0.005	0.009	0.378	-0.002	0.004	0.237
	Unadjusted GEE	-0.005	0.009	0.902	-0.002	0.004	0.557
	GEE-Center	-0.005	0.009	0.323	-0.002	0.004	0.230
	Unadjusted GLMM	-42.48	1.79E7	6.7×10^{176}	-0.039	0.087	1.161
	GLMM-Center	-0.114	0.355	2.288	-0.039	0.088	1.165
	GLMM-All	-0.115	0.355	2.480	-0.039	0.088	1.196
0.25	Simple	-0.005	0.010	0.404	-0.002	0.004	0.254
	CMH	-0.005	0.010	0.391	-0.002	0.004	0.244
	Unadjusted GEE	-0.005	0.010	0.916	-0.002	0.004	0.565
	GEE-Center	-0.005	0.010	0.334	-0.002	0.004	0.237
	Unadjusted GLMM	-25.52	6.46E6	4.0×10^{176}	-0.031	0.070	1.074
	GLMM-Center	-0.118	0.386	2.344	-0.034	0.077	1.112
	GLMM-All	-0.117	0.379	2.633	-0.034	0.077	1.158
0.5	Simple	-0.007	0.012	0.447	-0.002	0.004	0.278
	CMH	-0.007	0.012	0.405	-0.002	0.004	0.249
	Unadjusted GEE	-0.007	0.012	0.943	-0.002	0.004	0.577
	GEE-Center	-0.007	0.012	0.353	-0.002	0.004	0.243
	Unadjusted GLMM	-118.7	2.84E7	1.9×10^{177}	-0.025	0.051	0.973
	GLMM-Center	-0.129	0.426	2.509	-0.034	0.073	1.101
	GLMM-All	-0.128	0.418	2.893	-0.034	0.073	1.140
0.75	Simple	-0.005	0.013	0.525	-0.002	0.004	0.321
	CMH	-0.005	0.013	0.417	-0.002	0.004	0.249
	Unadjusted GEE	-0.005	0.013	0.992	-0.002	0.004	0.601
	GEE-Center	-0.005	0.013	0.369	-0.002	0.004	0.243
	Unadjusted GLMM	-301.1	3.5E7	4.8×10^{177}	-0.014	0.029	0.830
	GLMM-Center	-0.141	0.687	3.412	-0.036	0.080	1.153
	GLMM-All	-0.139	0.665	3.869	-0.036	0.079	1.186

[1]Methods: Simple, based on crude RR; CMH, Cochran Mantel Haenszel-stratified by center; Unadjusted GEE, Log-Poisson regression without adjustment for center and robust SEs; Unadjusted GLMM, generalized linear mixed model without random effect; GLMM-Center, generalized linear mixed model with random center effect; GLMM-All, generalized linear mixed model with random center and interaction effects.

Table 3.16. Properties of point estimated PF for true PF was 0.4 using seven models under homogeneous scenarios ($\boldsymbol{\sigma_{tc}^2}$ = 0) with different ICC values (0, 0.25, 0.5, and 0.75) and center sizes (5 *vs* 25).

		Center (size, level)					
		(20, 5)			(10, 25)		
ICC	Method	Bias	MSE	Avg.length	Bias	MSE	Avg.length
0	Simple	-0.017	0.010	0.383	-0.014	0.004	0.240
	CMH	-0.017	0.010	0.383	-0.014	0.004	0.240
	Unadjusted GEE	-0.017	0.010	0.640	-0.014	0.004	0.393
	GEE-Center	-0.017	0.010	0.325	-0.014	0.004	0.233
	Unadjusted GLMM	0.225	0.067	1.2×10^{165}	0.229	0.058	0.311
	GLMM-Center	0.228	0.068	0.524	0.230	0.059	0.312
	GLMM-All	0.229	0.068	0.553	0.231	0.059	0.318
0.25	Simple	-0.030	0.013	0.396	-0.029	0.005	0.249
	CMH	-0.030	0.013	0.383	-0.029	0.005	0.240
	Unadjusted GEE	-0.030	0.013	0.658	-0.029	0.005	0.405
	GEE-Center	-0.030	0.013	0.369	-0.029	0.005	0.247
	Unadjusted GLMM	0.187	0.053	2.3×10^{165}	0.188	0.042	0.327
	GLMM-Center	0.212	0.064	0.571	0.207	0.050	0.330
	GLMM-All	0.213	0.064	0.625	0.208	0.050	0.341
0.5	Simple	-0.054	0.018	0.421	-0.055	0.008	0.264
	CMH	-0.054	0.018	0.385	-0.055	0.008	0.240
	Unadjusted GEE	-0.054	0.018	0.692	-0.055	0.008	0.426
	GEE-Center	-0.054	0.018	0.416	-0.055	0.008	0.262
	Unadjusted GLMM	0.129	0.040	2.3×10^{166}	0.126	0.023	0.351
	GLMM-Center	0.199	0.063	0.666	0.193	0.045	0.361
	GLMM-All	0.199	0.063	0.735	0.192	0.044	0.371
0.75	Simple	-0.095	0.028	0.471	-0.106	0.017	0.296
	CMH	-0.095	0.028	0.383	-0.106	0.017	0.238
	Unadjusted GEE	-0.095	0.028	0.757	-0.106	0.017	0.467
	GEE-Center	-0.095	0.028	0.463	-0.106	0.017	0.276
	Unadjusted GLMM	0.033	0.032	2.0×10^{167}	0.017	0.009	0.388
	GLMM-Center	0.178	0.069	0.947	0.174	0.041	0.435
	GLMM-All	0.176	0.069	1.048	0.172	0.041	0.445

[1]Methods: Simple, based on crude RR; CMH, Cochran Mantel Haenszel-stratified by center; Unadjusted GEE, Log-Poisson regression without adjustment for center and robust SEs; Unadjusted GLMM, generalized linear mixed model without random effect; GLMM-Center, generalized linear mixed model with random center effect; GLMM-All, generalized linear mixed model with random center and interaction effects.

Table 3.17. Properties of point estimated PF for true PF was 0 using seven models under heterogeneous scenarios ($\boldsymbol{\sigma_{tc}^2}$ = 0.4) with different ICC values (0, 0.25, 0.5, and 0.75) and center sizes (5 *vs* 25).

		Center (size, level)					
		(20, 5)			(10, 25)		
ICC	Method	Bias	MSE	Avg.length	Bias	MSE	Avg.length
0	Simple	-0.007	0.013	0.389	-0.002	0.004	0.243
	CMH	-0.007	0.013	0.389	-0.002	0.004	0.243
	Unadjusted GEE	-0.007	0.013	0.909	-0.002	0.004	0.560
	GEE-Center	-0.007	0.013	0.377	-0.002	0.004	0.251
	Unadjusted GLMM	-38.27	1.45E7	6.0×10^{176}	-0.040	0.097	1.133
	GLMM-Center	-0.149	0.502	2.367	-0.040	0.098	1.140
	GLMM-All	-0.150	0.500	2.761	-0.040	0.098	1.207
0.25	Simple	-0.006	0.014	0.418	-0.002	0.005	0.260
	CMH	-0.006	0.014	0.403	-0.002	0.005	0.250
	Unadjusted GEE	-0.006	0.014	0.924	-0.002	0.005	0.568
	GEE-Center	-0.006	0.014	0.397	-0.002	0.005	0.258
	Unadjusted GLMM	-63.67	1.49E7	1.0×10^{177}	-0.031	0.073	1.046
	GLMM-Center	-0.146	0.607	2.455	-0.036	0.084	1.096
	GLMM-All	-0.145	0.567	2.997	-0.036	0.083	1.179
0.5	Simple	-0.007	0.016	0.461	-0.002	0.005	0.285
	CMH	-0.007	0.016	0.416	-0.002	0.005	0.254
	Unadjusted GEE	-0.007	0.016	0.951	-0.002	0.005	0.581
	GEE-Center	-0.007	0.016	0.417	-0.002	0.005	0.264
	Unadjusted GLMM	-220.4	4.67E7	3.5×10^{177}	-0.022	0.052	0.942
	GLMM-Center	-0.159	0.644	2.633	-0.033	0.081	1.090
	GLMM-All	-0.154	0.607	3.277	-0.033	0.080	1.166
0.75	Simple	-0.008	0.018	0.541	-0.002	0.005	0.329
	CMH	-0.008	0.018	0.423	-0.002	0.005	0.252
	Unadjusted GEE	-0.008	0.018	1.005	-0.002	0.005	0.605
	GEE-Center	-0.008	0.018	0.426	-0.002	0.005	0.263
	Unadjusted GLMM	-360.4	6.23E7	5.7×10^{177}	-0.012	0.030	0.807
	GLMM-Center	-0.194	1.002	4.107	-0.037	0.092	1.151
	GLMM-All	-0.182	0.839	4.673	-0.036	0.090	1.216

[1]Methods: Simple, based on crude RR; CMH, Cochran Mantel Haenszel-stratified by center; Unadjusted GEE, Log-Poisson regression without adjustment for center and robust SEs; Unadjusted GLMM, generalized linear mixed model without random effect; GLMM-Center, generalized linear mixed model with random center effect; GLMM-All, generalized linear mixed model with random center and interaction effects.

Table 3.18. Properties of point estimated PF for true PF was 0.4 using seven models under heterogeneous scenarios ($\boldsymbol{\sigma_{tc}^2}$ = 0.4) with different ICC values (0, 0.25, 0.5, and 0.75) and center sizes (5 *vs* 25).

		Center (size, level)					
		(20, 5)			(10, 25)		
ICC	Method	Bias	MSE	Avg.length	Bias	MSE	Avg.length
0	Simple	-0.022	0.013	0.389	-0.020	0.005	0.243
	CMH	-0.022	0.013	0.389	-0.020	0.005	0.243
	Unadjusted GEE	-0.022	0.013	0.648	-0.020	0.005	0.398
	GEE-Center	-0.022	0.013	0.390	-0.020	0.005	0.255
	Unadjusted GLMM	0.205	0.064	3.0×10^{165}	0.213	0.053	0.318
	GLMM-Center	0.211	0.066	0.541	0.216	0.054	0.319
	GLMM-All	0.214	0.068	0.616	0.218	0.055	0.336
0.25	Simple	-0.036	0.017	0.404	-0.036	0.006	0.253
	CMH	-0.036	0.017	0.390	-0.036	0.006	0.244
	Unadjusted GEE	-0.036	0.017	0.668	-0.036	0.006	0.411
	GEE-Center	-0.036	0.017	0.426	-0.036	0.006	0.266
	Unadjusted GLMM	0.167	0.053	8.2×10^{165}	0.170	0.037	0.334
	GLMM-Center	0.198	0.065	0.592	0.196	0.046	0.338
	GLMM-All	0.200	0.066	0.699	0.196	0.046	0.360
0.5	Simple	-0.059	0.022	0.428	-0.064	0.009	0.269
	CMH	-0.059	0.022	0.390	-0.064	0.009	0.243
	Unadjusted GEE	-0.059	0.022	0.702	-0.064	0.009	0.432
	GEE-Center	-0.059	0.022	0.461	-0.064	0.009	0.279
	Unadjusted GLMM	0.112	0.042	2.6×10^{166}	0.107	0.020	0.358
	GLMM-Center	0.187	0.067	0.691	0.183	0.043	0.372
	GLMM-All	0.188	0.067	0.825	0.182	0.042	0.392
0.75	Simple	-0.107	0.034	0.485	-0.113	0.018	0.302
	CMH	-0.107	0.034	0.388	-0.113	0.018	0.240
	Unadjusted GEE	-0.107	0.034	0.775	-0.113	0.018	0.473
	GEE-Center	-0.107	0.034	0.491	-0.113	0.018	0.290
	Unadjusted GLMM	0.012	0.037	2.2×10^{167}	0.001	0.009	0.393
	GLMM-Center	0.161	0.076	1.022	0.166	0.040	0.448
	GLMM-All	0.158	0.075	1.180	0.163	0.039	0.468

[1] Methods: Simple, based on crude RR; CMH, Cochran Mantel Haenszel-stratified by center; Unadjusted GEE, log-Poisson regression without adjustment for center and robust SEs; Unadjusted GLMM, generalized linear mixed model without random effect; GLMM-Center, generalized linear mixed model with random center effect; GLMM-All, generalized linear mixed model with random center and interaction effects.

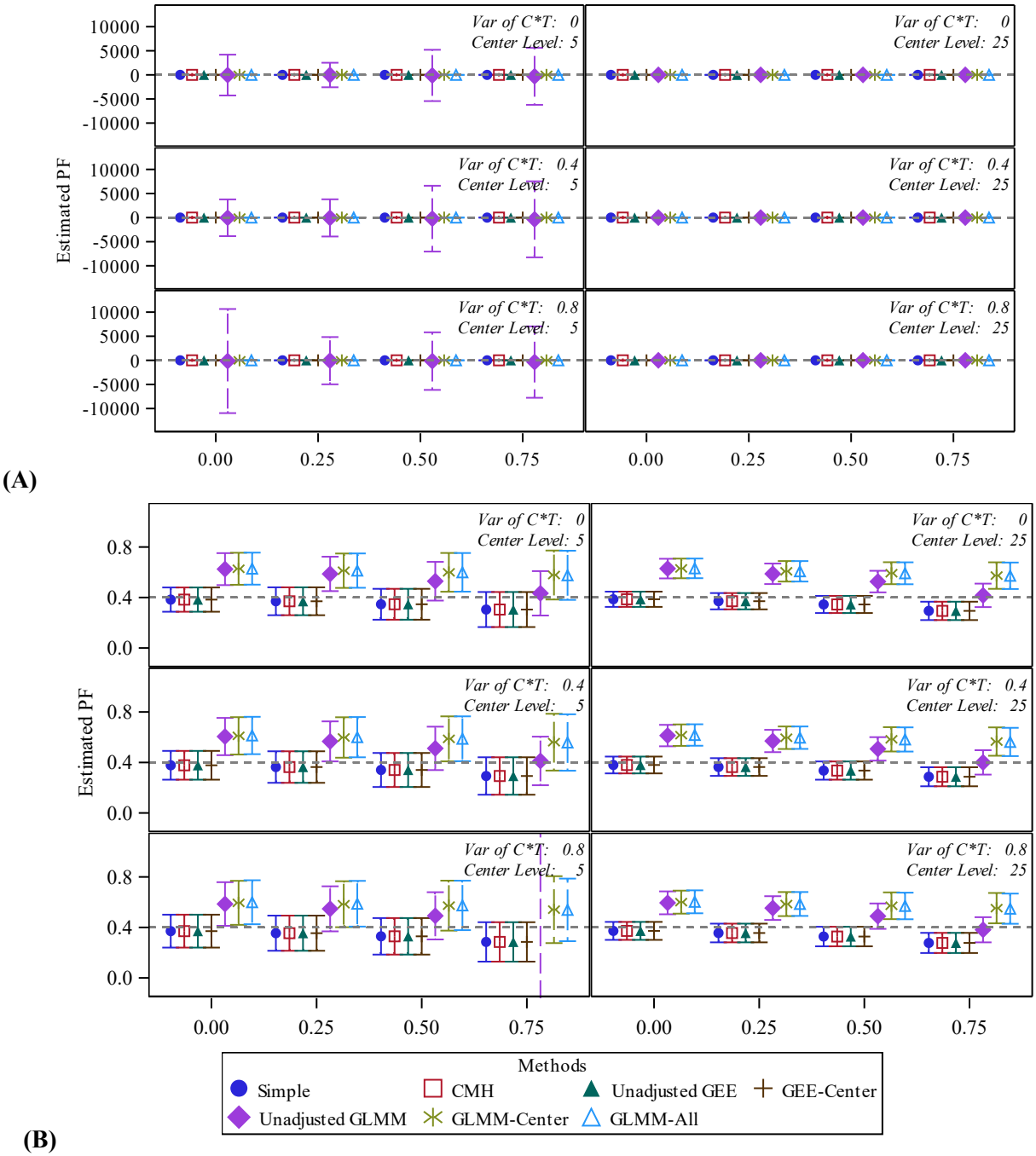
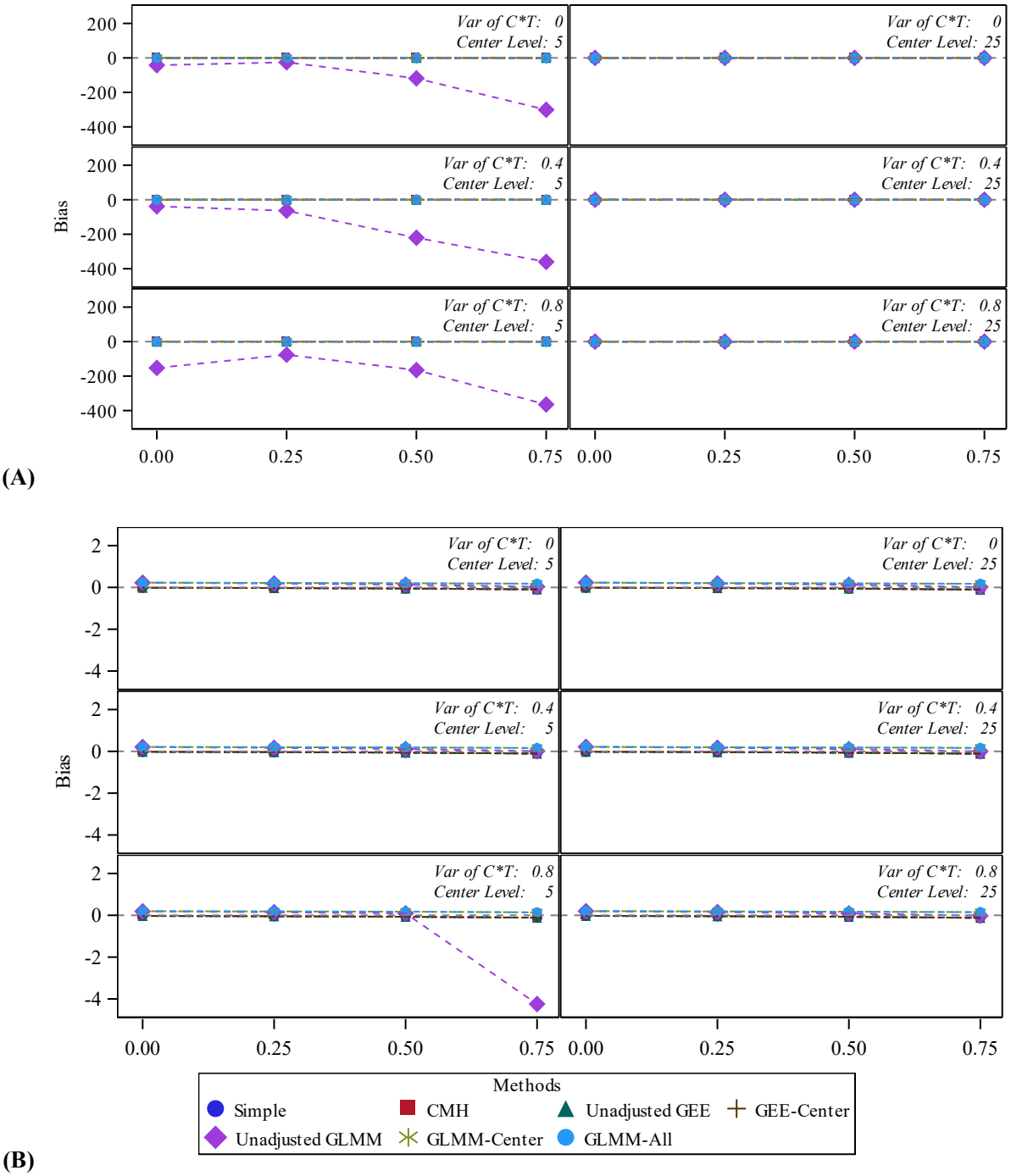


Figure 3.7. Average point estimated PF with empirical simulation standard deviation (*SD*) from seven models under both homogeneous or heterogeneous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and center numbers (5 *vs* 25). (A) Scenarios: true PF was 0. (B) Scenarios: true PF was 0.4.



(A)

(B)

Figure 3.8. Bias of point estimated PF from seven models under both homogeneous or heterogeneous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and center sizes (5 *vs* 25). (A) Scenarios: true PF was 0. (B) Scenarios: true *PF* was 0.4.

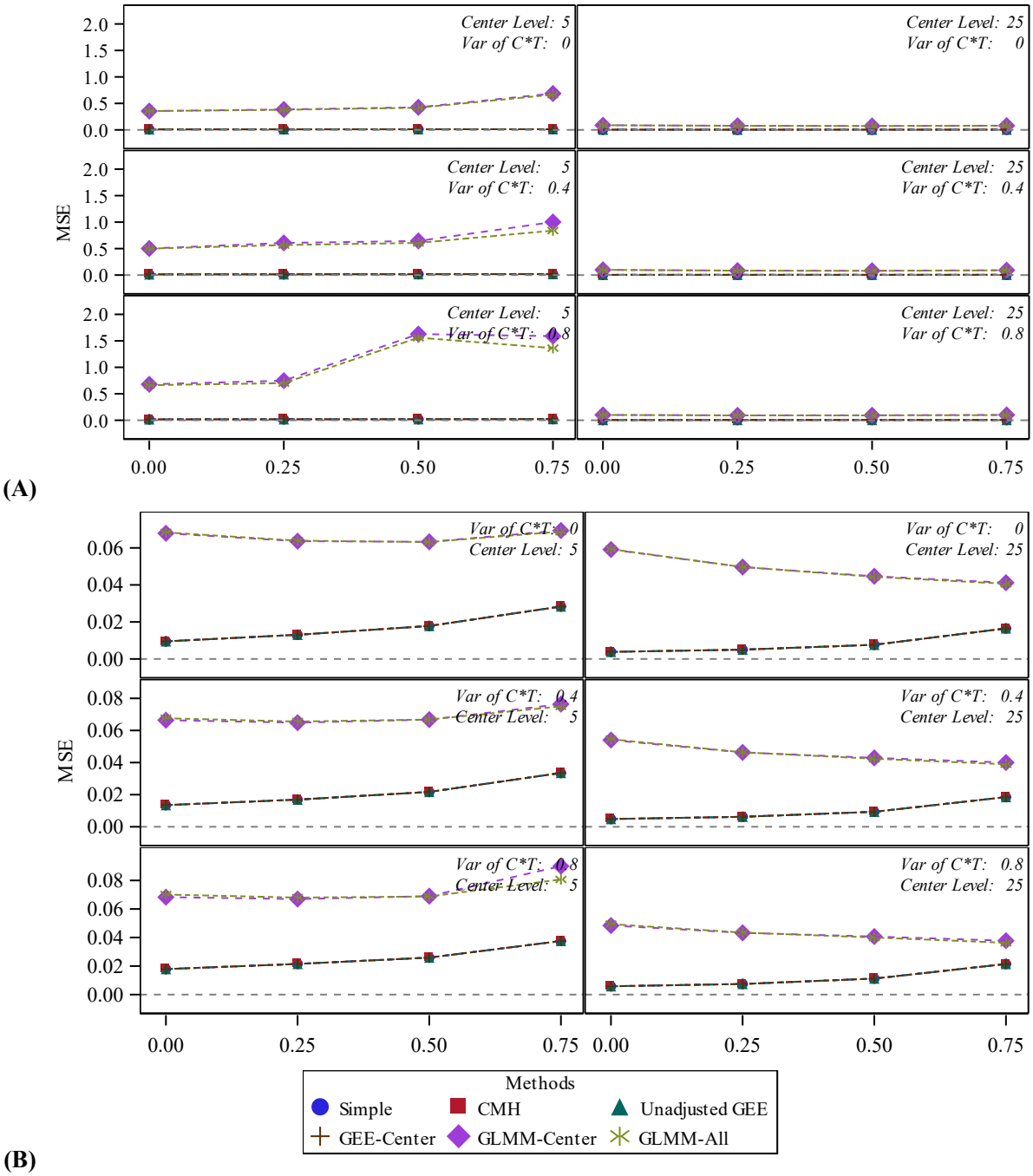


Figure 3.9. MSE of point estimated PF from six models (except Unadjusted GLMM) under both homogeneous or heterogeneous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and center sizes (5 *vs* 25). (A) Scenarios: true PF was 0. (B) Scenarios: true PF was 0.4.

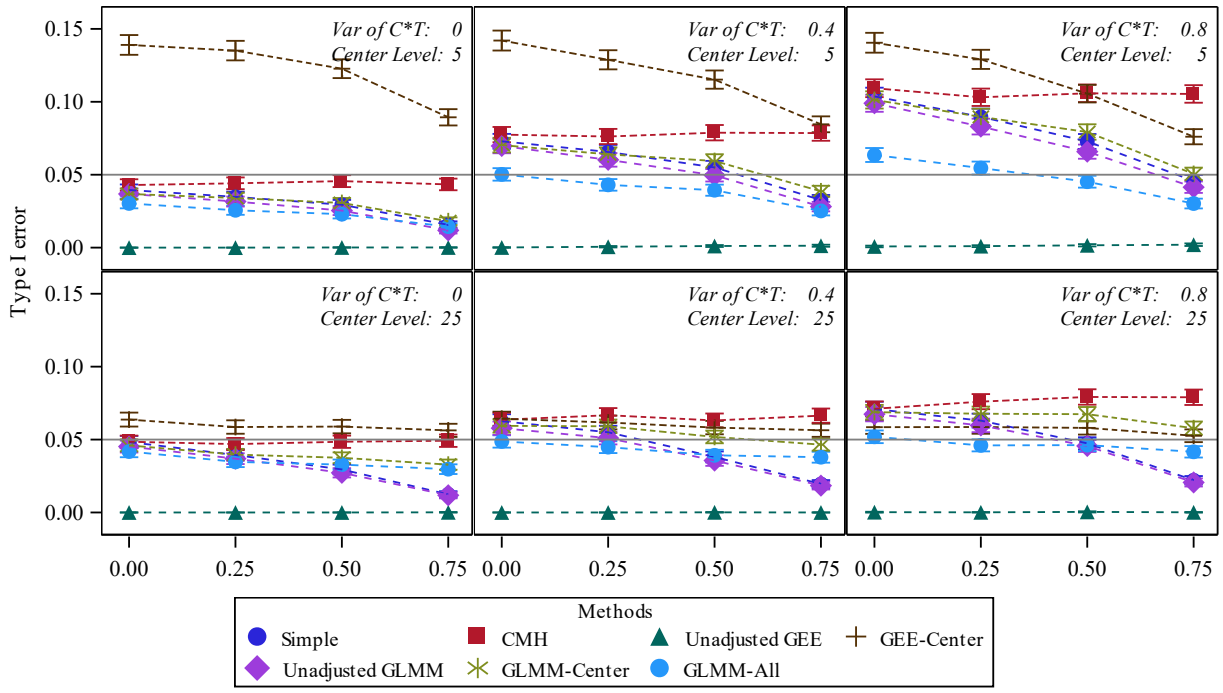


Figure 3.10. Type I error rate of hypothesis testing for treatment effect PF using all seven models under both homogeneous or heterogeneous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and the number of centers (5 vs 25). The grey solid reference line represents the nominal Type I error rate $\boldsymbol{\alpha} = 0.05$.

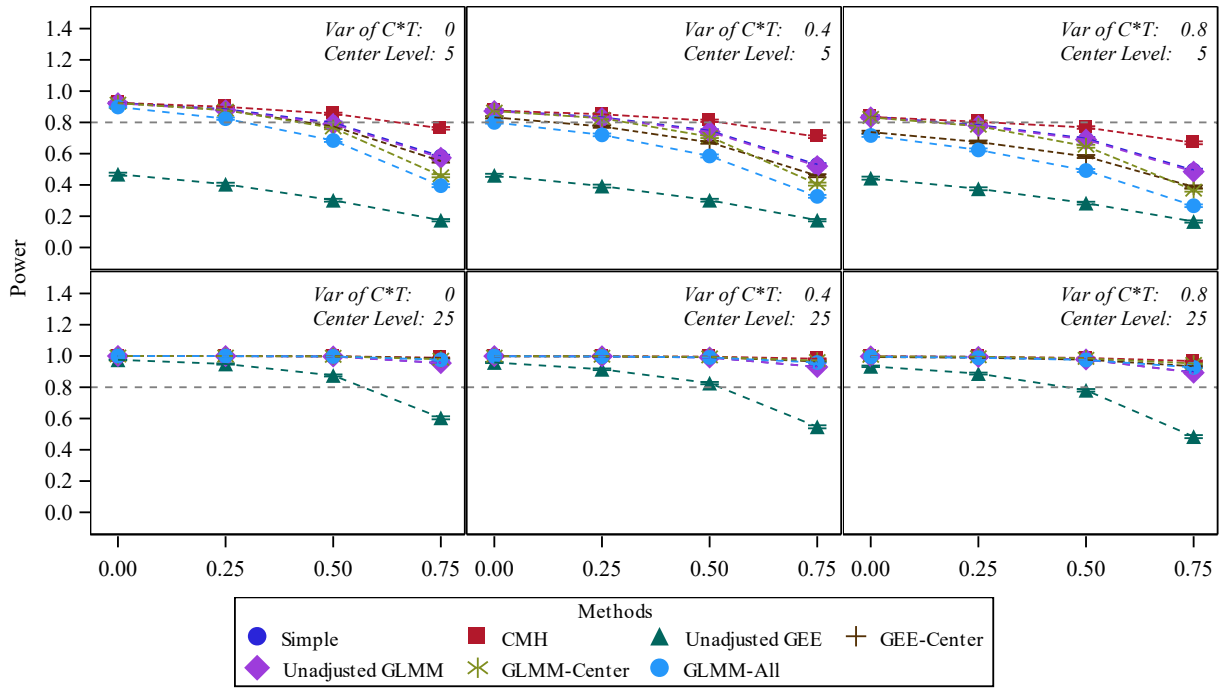
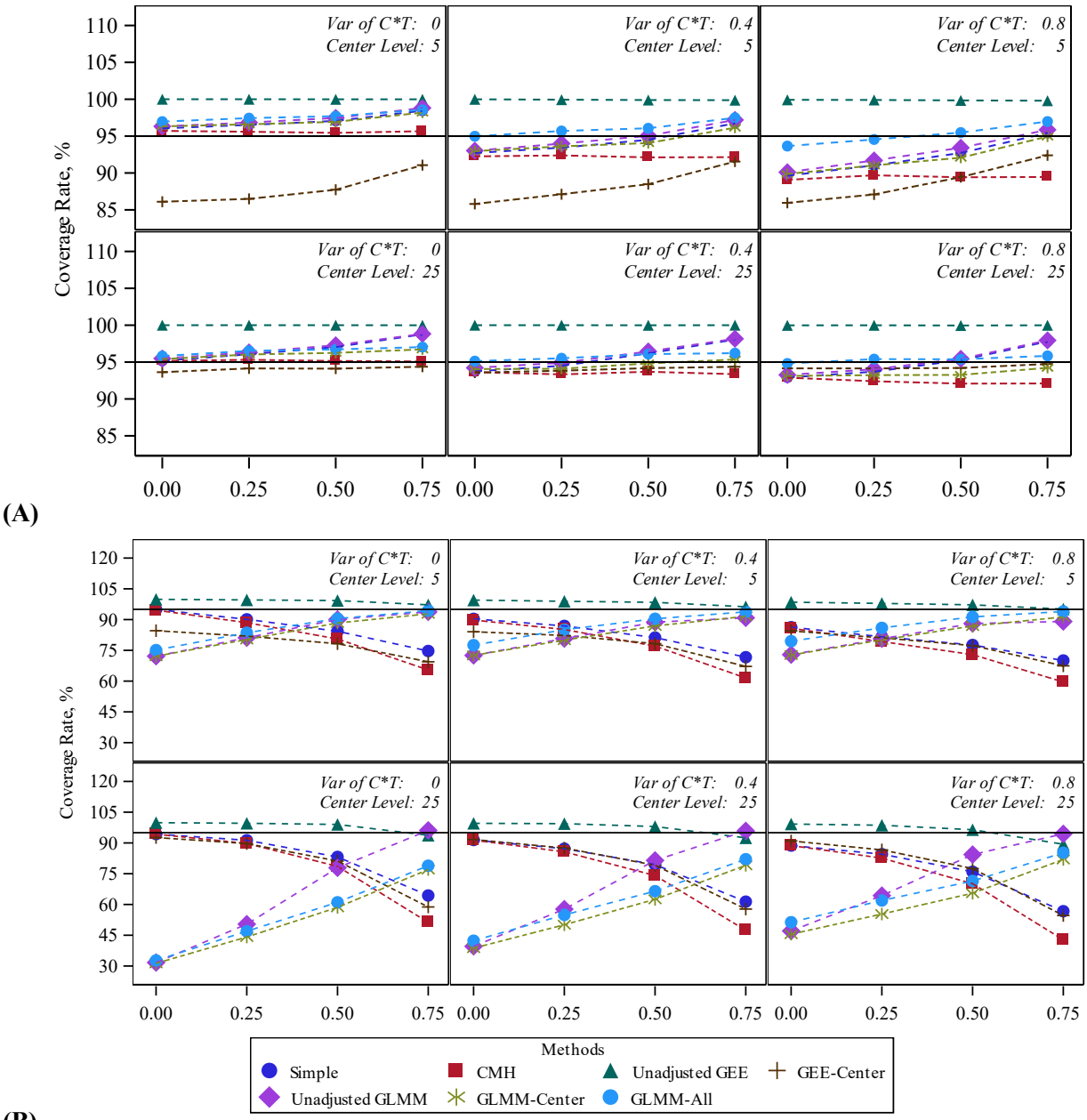


Figure 3.11. Power analysis of hypothesis testing for treatment effect PF using all seven models under both homogeneous or heterogeneous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and the number of centers (5 vs 25). The grey solid reference line represents the nominal power 1 - $\boldsymbol{\beta}$= 0.8.



(A)

(B)

Figure 3.12. Coverage rate of 95% confidence interval for treatment effect PF = 0 using all seven models under both homogeneous or heterogeneous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and center sizes (5 vs 25). (A) Scenarios: true PF was 0. (B) Scenarios: true PF was 0.4.

Chapter 4 - Future Work

To date, there is a lack of agreement on when and how to adjust center variance in multicenter trials. The ICH-E9 guideline for multicenter RCTs focuses on possible treatment effect heterogeneity (ICH, 1999). Mainly, it suggested an exploratory analysis to detect heterogeneity of treatment across centers when multicenter RCTs have positive treatment effects and a significant number of patients per center. It doesn't provide any recommendations for addressing the heterogeneity of treatment when the center size and level are not large.

The work from this dissertation can be summarized as follows. Whatever outcome is continuous or binary, the first step should always be to explore the heterogeneity of treatment effects across stratification factors for the randomization procedure, as this may affect the inference of treatment effect. When using a statistical significance test for the interaction term, once heterogeneity of treatment effect is not found, a reduced model without interaction term should be adopted to detect the main effect of treatment due to the high power.

In future research, simulations could be performed to investigate the ability of analytical procedures to deal with imbalanced data among clusters. Unbalanced data situations not only reflected real-world scenarios but also provided valuable insights into the procedures' efficiency.

References

Agresti, A., & Hartzel, J. (2000). Strategies for comparing treatments on a binary response with multi-centre data. *Statistics in Medicine*, *19*(8), 1115–1139.

Altman, D. G., & Bland, J. M. (1999). How to randomise. *BMJ : British Medical Journal*, *319*(7211), 703–704.

Babor, T. F. (2004). Brief Treatments for Cannabis Dependence: Findings From a Randomized Multisite Trial. *Journal of Consulting and Clinical Psychology*, *72*(3), 455–466.

Callegaro, A., Harsha Shree, B. S., & Karkada, N. (2021). Inference under covariate-adaptive randomization: A simulation study. *Statistical Methods in Medical Research*, *30*(4),

Chu, R., Thabane, L., Ma, J., Holbrook, A., Pullenayegum, E., & Devereaux, P. J. (2011). Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: A simulation study. *BMC Medical Research Methodology*, *11*(1), 21.

Ciolino, J. D., Martin, R. H., Zhao, W., Hill, M. D., Jauch, E. C., & Palesch, Y. Y. (2015). Measuring continuous baseline covariate imbalances in clinical trial data. *Statistical Methods in Medical Research*, *24*(2), 255–272.

Cochran, W. G. (1954). Some Methods for Strengthening the Common χ2 Tests. *Biometrics*, *10*(4), 417–451.

Dickey, D. (2008). *374-2008: PROC MIXED: Underlying Ideas with Examples*.

Edgar, K., Roberts, I., & Sharples, L. (2021). Including random centre effects in design, analysis and presentation of multi-centre trials. *Trials*, *22*(1), 357.

Eisenhart, C. (1947). The Assumptions Underlying the Analysis of Variance. *Biometrics*, *3*(1), 1–21.

Eldridge, S. M., Ukoumunne, O. C., & Carlin, J. B. (2009). The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. *International Statistical Review*, *77*(3), 378–394.

Evans, B. A., Feng, Z., & Peterson, A. V. (2001). A comparison of generalized linear mixed model procedures with estimating equations for variance and covariance parameter estimation in longitudinal studies and group randomized trials. *Statistics in Medicine*, *20*(22), 3353–3373.

Fang, J. (2011). *345-2011: Using SAS® Procedures FREQ, GENMOD, LOGISTIC, and PHREG to Estimate Adjusted Relative Risks: A Case Study*. 11.

Fisher, S. R. A. (1928). *Statistical Methods for Research Workers*. Oliver and Boyd.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical Methods for Rates and Proportions*. John Wiley & Sons.

George Casella. (2006). *Statistical Design*.

Giesbrecht, F. G., & Burns, J. C. (1985). Two-Stage Analysis Based on a Mixed Model: Large-Sample Asymptotic Theory and Small-Sample Simulation Results. *Biometrics*, *41*(2), 477–486.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning Variation in Multilevel Models. *Understanding Statistics*, *1*(4), 223–231.

Graybill, F.A. (1976). Theory and Application of the Linear Model. Duxbury Press, North Scituate, MA.

Greenland, S., & Robins, J. M. (1985). Estimation of a Common Effect Parameter from Sparse Follow-Up Data. *Biometrics*, *41*(1), 55–68.

Gulliford, M. C., Adams, G., Ukoumunne, O. C., Latinovic, R., Chinn, S., & Campbell, M. J. (2005). Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *Journal of Clinical Epidemiology*, *58*(3), 246–251.

Hartley, H. O. (1967). Expectations, Variances and Covariances of Anova Mean Squares by "Synthesis." *Biometrics*, *23*(1), 105–114.

Hauck, W. W., Anderson, S., & Marcus, S. M. (1998). Should We Adjust for Covariates in Nonlinear Regression Analyses of Randomized Trials? *Controlled Clinical Trials*, *19*(3), 249–256.

Hedden, S. L., Woolson, R. F., & Malcolm, R. J. (2006). Randomization in substance abuse clinical trials. *Substance Abuse Treatment, Prevention, and Policy*, *1*(1), 6.

Kackar, R. N., & Harville, D. A. (1984). Approximations for Standard Errors of Estimators of Fixed and Random Effect in Mixed Linear Models. *Journal of the American Statistical Association*, *79*(388), 853–862.

Kahan, B. C. (2014). Accounting for centre-effects in multicentre trials with a binary outcome – when, why, and how? *BMC Medical Research Methodology*, *14*(1), 20.

Kahan, B. C., Forbes, G., Ali, Y., Jairath, V., Bremner, S., Harhay, M. O., Hooper, R., Wright, N., Eldridge, S. M., & Leyrat, C. (2016). Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: A review, reanalysis, and simulation study. *Trials*, *17*(1), 438.

Kahan, B. C., & Morris, T. P. (2012). Improper analysis of trials randomised using stratified blocks or minimisation. *Statistics in Medicine*, *31*(4), 328–340.

Kahan, B. C., & Morris, T. P. (2013a). Analysis of multicentre trials with continuous outcomes: When and how should we account for centre effects? *Statistics in Medicine*, *32*(7), 1136–1149.

Kahan, B. C., & Morris, T. P. (2013b). Assessing potential sources of clustering in individually randomised trials. *BMC Medical Research Methodology*, *13*(1), 58.

Kang, M., Ragan, B. G., & Park, J.-H. (2008). Issues in Outcomes Research: An Overview of Randomization Techniques for Clinical Trials. *Journal of Athletic Training*, *43*(2), 215–221.

Kenward, M. G., & Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics*, *53*(3), 983–997.

Kernan, W. N., Viscoli, C. M., Makuch, R. W., Brass, L. M., & Horwitz, R. I. (1999). Stratified Randomization for Clinical Trials. *Journal of Clinical Epidemiology*, *52*(1), 19–26.

Kim, J., & Shin, W. (2014). How to Do Random Allocation (Randomization). *Clinics in Orthopedic Surgery*, *6*(1), 103–109.

Kim, J., Troxel, A. B., Halpern, S. D., Volpp, K. G., Kahan, B. C., Morris, T. P., & Harhay, M. O. (2020). Analysis of multicenter clinical trials with very low event rates. *Trials*, *21*(1), 917.

Kraemer, H. C. (2000). Pitfalls of Multisite Randomized Clinical Trials of Efficacy and Effectiveness. *Schizophrenia Bulletin*, *26*(3), 533–541.

Lachin, J. M. (2009). *Biostatistical Methods: The Assessment of Relative Risks*. John Wiley & Sons.

Lachin, J. M., Matts, J. P., & Wei, L. J. (1988). Randomization in clinical trials: Conclusions and recommendations. *Controlled Clinical Trials*, *9*(4), 365–374.

Lewis, J. A. (1999). Statistical principles for clinical trials (ICH E9): An introductory note on an international guideline. *Statistics in Medicine*, *18*(15), 1903–1942.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrics*, *71*(1), 13–22.

Lin, Y., Zhu, M., & Su, Z. (2015). The pursuit of balance: An overview of covariate-adaptive randomization techniques in clinical trials. *Contemporary Clinical Trials*, *45*, 21–25.

Lingsma, H. F., Roozenbeek, B., Perel, P., Roberts, I., Maas, A. I., & Steyerberg, E. W. (2011). Between-centre differences and treatment effects in randomized controlled trials: A case study in traumatic brain injury. *Trials*, *12*(1), 201.

Localio, A. R., Berlin, J. A., Ten Have, T. R., & Kimmel, S. E. (2001). Adjustments for Center in Multicenter Studies: An Overview. *Annals of Internal Medicine*, *135*(2), 112.

Localio, A. R., Margolis, D. J., & Berlin, J. A. (2007). Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *Journal of Clinical Epidemiology*, *60*(9), 874–882.

Ma, W., Qin, Y., Li, Y., & Hu, F. (2020). Statistical Inference for Covariate-Adaptive Randomization Procedures. *Journal of the American Statistical Association*, *115*(531), 1488–1497.

Mancl, L. A., & DeRouen, T. A. (2001). A Covariance Estimator for GEE with Improved Small-Sample Properties. *Biometrics*, *57*(1), 126–134.

McGraw, K. O., & Wong, S. P. (n.d.). *Forming Inferences About Some Intraclass Correlation Coefficients*.

McNutt, L.-A. (2003a). Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. *American Journal of Epidemiology*, *157*(10), 940–943.

McNutt, L.-A. (2003b). Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. *American Journal of Epidemiology*, *157*(10), 940–943.

McNutt, L.-A., Wu, C., Xue, X., & Hafner, J. P. (2003). Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. *American Journal of Epidemiology*, *157*(10), 940–943.

McVey, D.-A. (n.d.). *General Licensing Considerations: Efficacy Studies for Prophylactic and Therapeutic Biologics*. *800*, 11.

Parzen, M., Lipsitz, S. R., & Dear, K. B. G. (1998). Does Clustering Affect the Usual Test Statistics of no Treatment Effect in a Randomized Clinical Trial? *Biometrical Journal*, *40*(4), 385–402.

Pedroza, C., & Truong, V. T. T. (2017). Estimating relative risks in multicenter studies with a small number of centers — which methods to use? A simulation study. *Trials*, *18*(1), 512.

Pickering, R. M., & Weatherall, M. (2007). The analysis of continuous outcomes in multi-centre trials with small centre sizes. *Statistics in Medicine*, *26*(30), 5445–5456.

Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practiceand problems. *Statistics in Medicine*, *21*(19), 2917–2930.

Pocock, S. J., & Simon, R. (1975). Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial. *Biometrics*, *31*(1), 103–115.

Pond, G. R., Tang, P. A., Welch, S. A., & Chen, E. X. (2010). Trends in the application of dynamic allocation methods in multi-arm cancer clinical trials. *Clinical Trials*, *7*(3), 227–234.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata, Second Edition*. Stata Press.

Robinson, L. D., & Jewell, N. P. (1991). Some Surprising Results about Covariate Adjustment in Logistic Regression Models. *International Statistical Review / Revue Internationale de Statistique*, *59*(2), 227–240.

Rothman, K. J., & Health (U.S.), N. I. of. (1979). *Epidemiologic Analysis with a Programmable Calculator*. U.S. Department of Health, Education, and Welfare, Public Health Service, National Institutes of Health.

Schulz, K. F., & Grimes, D. A. (2002). Generation of allocation sequences in randomised trials: Chance, not choice. *The Lancet*, *359*(9305), 515–519

Schwemer, G. (2000). General Linear Models for Multicenter Clinical Trials. *Controlled Clinical Trials*, *21*(1), 21–29.

Scott, N. W., McPherson, G. C., Ramsay, C. R., & Campbell, M. K. (2002). The method of minimization for allocation to clinical trials: A review. *Controlled Clinical Trials*, *23*(6), 662–674.

Skove, T., Deddens, J., Petersen, M. R., & Endahl, L. (1998). Prevalence proportion ratios: Estimation and hypothesis testing. *International Journal of Epidemiology*, *27*(1), 91–95.

Tangri, N., Kitsios, G. D., Su, S. H., & Kent, D. M. (2010). Accounting for Center Effects in Multicenter Trials. *Epidemiology*, *21*(6), 912.

Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clinical Pharmacology & Therapeutics*, *15*(5), 443–453.

Tian, L., Jiang, F., Hasegawa, T., Uno, H., Pfeffer, M., & Wei, Lj. (2019). Moving beyond the conventional stratified analysis to estimate an overall treatment efficacy with the data from a comparative randomized clinical study. *Statistics in Medicine*, *38*(6), 917–932.

Turner, R. M., Omar, R. Z., & Thompson, S. G. (2001). Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine*, *20*(3), 453–472.

Verbeke, G., & Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, *59*(2), 254–262.

Wang, W., & Shan, G. (2015). Exact confidence intervals for the relative risk and the odds ratio. *Biometrics*, *71*(4), 985–995.

Weir, C. J., & Lees, K. R. (2003). Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial. *Statistics in Medicine*, *22*(5), 705–726.

Worthington, H. (2004). Methods for Pooling Results from Multi-center Studies. *Journal of Dental Research*, *83*, C119-21.

Zhang, J., & Yu, K. F. (1998). What's the Relative Risk?A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. *JAMA*, *280*(19), 1690–1691.

Zou, G. (2004). A Modified Poisson Regression Approach to Prospective Studies with Binary Data. *American Journal of Epidemiology*, *159*(7), 702–706.

Appendix A - Additional Tables and Figures for Chapter 3.4

Table A. 1. Average point estimated PF and empirical simulation standard deviation (SD) for true PF was 0 using three additional GEE models under homogeneous and heterogenous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and the number of centers (5 *vs* 25).

			Variance of Center-by-Treatment Effect					
			0		0.4		0.8	
n_c	ICC	Method	Avg. $\widehat{PF}$	SD	Avg. $\widehat{PF}$	SD	Avg. $\widehat{PF}$	SD
5	0	GEE* without Robust SEs	-0.005	0.099	-0.005	0.113	-0.011	0.131
		GEE*-Center	-0.005	0.099	-0.005	0.113	-0.011	0.131
		GEE*-Center by Trt	-0.005	0.099	-0.005	0.113	-0.011	0.131
	0.5	GEE* without Robust SEs	-0.007	0.108	-0.006	0.128	-0.012	0.141
		GEE*-Center	-0.007	0.108	-0.006	0.128	-0.011	0.141
		GEE*-Center by Trt	-0.007	0.108	-0.006	0.128	-0.011	0.141
	0.25	GEE* without Robust SEs	-0.007	0.102	-0.007	0.120	-0.008	0.131
		GEE*-Center	-0.007	0.102	-0.007	0.120	-0.008	0.131
		GEE*-Center by Trt	-0.007	0.102	-0.007	0.120	-0.008	0.131
	0.75	GEE* without Robust SEs	-0.005	0.118	-0.010	0.132	-0.009	0.149
		GEE*-Center	-0.003	0.115	-0.011	0.132	-0.010	0.148
		GEE*-Center by Trt	-0.003	0.115	-0.011	0.132	-0.010	0.148
25	0	GEE* without Robust SEs	-0.004	0.060	-0.002	0.066	-0.002	0.071
		GEE*-Center	-0.004	0.060	-0.002	0.066	-0.002	0.071
		GEE*-Center by Trt	-0.004	0.060	-0.002	0.066	-0.002	0.071
	0.5	GEE* without Robust SEs	-0.001	0.064	-0.002	0.071	-0.002	0.075
		GEE*-Center	-0.002	0.064	-0.002	0.071	-0.002	0.074
		GEE*-Center by Trt	-0.002	0.064	-0.002	0.071	-0.002	0.074
	0.25	GEE* without Robust SEs	-0.002	0.061	-0.003	0.068	-0.000	0.072
		GEE*-Center	-0.002	0.061	-0.003	0.068	0.000	0.072
		GEE*-Center by Trt	-0.002	0.061	-0.003	0.068	0.000	0.072
	0.75	GEE* without Robust SEs	-0.002	0.065	-0.001	0.068	-0.003	0.076
		GEE*-Center	-0.000	0.062	0.000	0.066	-0.002	0.071
		GEE*-Center by Trt	-0.000	0.062	0.000	0.066	-0.002	0.071

Table A. 2. Average point estimated PF and empirical simulation standard deviation (SD) for true PF was 0.4 using three additional GEE models under homogeneous and heterogenous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and the number of centers (5 *vs* 25).

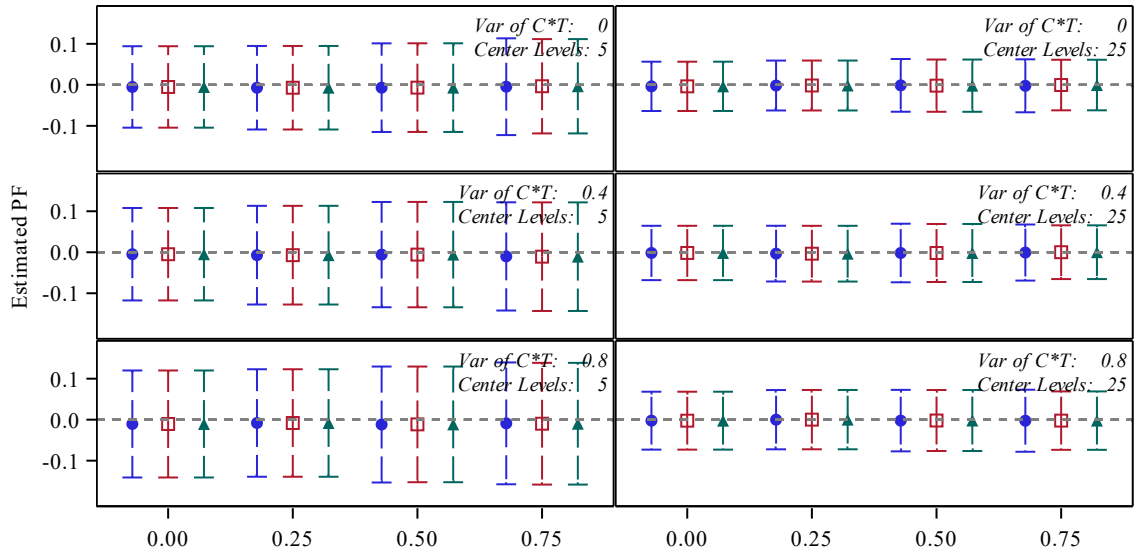
			Variance of Center-by-Treatment Effect					
			0		0.4		0.8	
n_c	ICC	Method	Avg. $\widehat{PF}$	SD	Avg. $\widehat{PF}$	SD	Avg. $\widehat{PF}$	SD
5	0	GEE* without Robust SEs	0.384	0.098	0.380	0.115	0.372	0.134
		GEE*-Center	0.384	0.098	0.380	0.115	0.372	0.134
		GEE*-Center by Trt	0.384	0.098	0.380	0.115	0.372	0.134
	0.5	GEE* without Robust SEs	0.346	0.121	0.342	0.134	0.329	0.148
		GEE*-Center	0.346	0.121	0.342	0.133	0.329	0.147
		GEE*-Center by Trt	0.346	0.121	0.342	0.133	0.329	0.147
	0.25	GEE* without Robust SEs	0.370	0.110	0.363	0.122	0.359	0.139
		GEE*-Center	0.370	0.110	0.363	0.122	0.359	0.139
		GEE*-Center by Trt	0.370	0.110	0.363	0.122	0.359	0.139
	0.75	GEE* without Robust SEs	0.303	0.134	0.294	0.149	0.285	0.157
		GEE*-Center	0.303	0.132	0.296	0.144	0.284	0.150
		GEE*-Center by Trt	0.303	0.132	0.296	0.144	0.284	0.150
25	0	GEE* without Robust SEs	0.385	0.061	0.378	0.067	0.373	0.073
		GEE*-Center	0.385	0.061	0.378	0.067	0.374	0.073
		GEE*-Center by Trt	0.385	0.061	0.378	0.067	0.374	0.073
	0.5	GEE* without Robust SEs	0.344	0.069	0.337	0.072	0.326	0.078
		GEE*-Center	0.346	0.068	0.337	0.072	0.327	0.078
		GEE*-Center by Trt	0.346	0.068	0.337	0.072	0.327	0.078
	0.25	GEE* without Robust SEs	0.370	0.064	0.363	0.068	0.356	0.073
		GEE*-Center	0.370	0.064	0.363	0.068	0.355	0.073
		GEE*-Center by Trt	0.370	0.064	0.363	0.068	0.355	0.073
	0.75	GEE* without Robust SEs	0.294	0.070	0.285	0.073	0.274	0.076
		GEE*-Center	0.289	0.071	0.280	0.073	0.269	0.073
		GEE*-Center by Trt	0.289	0.071	0.280	0.073	0.269	0.073

Table A. 3. Properties of point estimated PF for true PF was 0 using three additional GEE models under homogeneous and heterogenous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and the number of centers (5 *vs* 25).

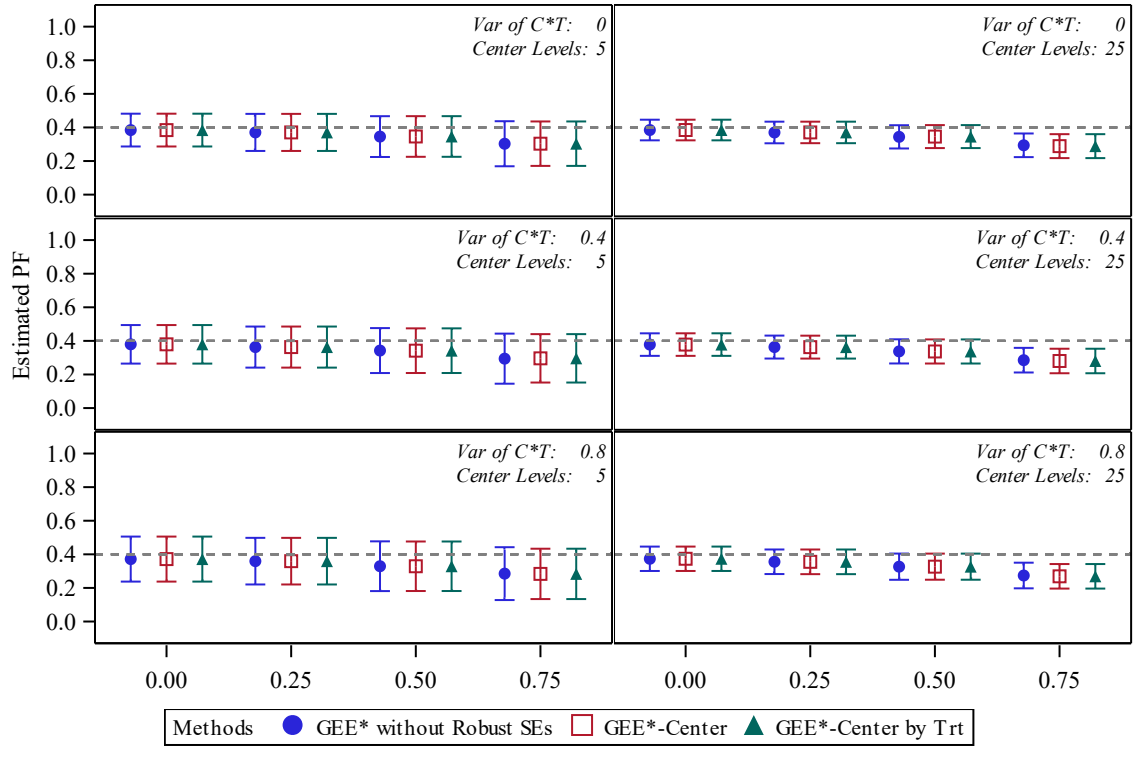
			Variance of Center-by-Treatment Effect								
			0			0.4			0.8		
n_c	ICC	Method	Bias	MSE	Length	Bias	MSE	Length	Bias	MSE	Length
5	0	GEE* without Robust SEs	-0.005	0.010	0.902	-0.005	0.013	0.907	-0.011	0.017	0.918
		GEE*-Center	-0.005	0.010	0.321	-0.005	0.013	0.372	-0.011	0.017	0.428
		GEE*-Center by Trt	-0.005	0.010	0.227	-0.005	0.013	0.263	-0.011	0.017	0.301
	0.5	GEE* without Robust SEs	-0.007	0.012	0.943	-0.006	0.017	0.950	-0.012	0.020	0.962
		GEE*-Center	-0.007	0.012	0.354	-0.006	0.017	0.420	-0.011	0.020	0.469
		GEE*-Center by Trt	-0.007	0.012	0.250	-0.006	0.017	0.296	-0.011	0.020	0.330
	0.25	GEE* without Robust SEs	-0.007	0.010	0.920	-0.007	0.015	0.925	-0.008	0.017	0.930
		GEE*-Center	-0.007	0.010	0.337	-0.007	0.015	0.395	-0.008	0.017	0.442
		GEE*-Center by Trt	-0.007	0.010	0.238	-0.007	0.015	0.278	-0.008	0.017	0.311
	0.75	GEE* without Robust SEs	-0.005	0.014	0.993	-0.010	0.018	1.004	-0.009	0.022	1.008
		GEE*-Center	-0.003	0.013	0.368	-0.011	0.018	0.418	-0.010	0.022	0.473
		GEE*-Center by Trt	-0.003	0.013	0.259	-0.011	0.018	0.294	-0.010	0.022	0.333
25	0	GEE* without Robust SEs	-0.004	0.004	0.559	-0.002	0.004	0.560	-0.002	0.005	0.564
		GEE*-Center	-0.004	0.004	0.231	-0.002	0.004	0.251	-0.002	0.005	0.272
		GEE*-Center by Trt	-0.004	0.004	0.163	-0.002	0.004	0.178	-0.002	0.005	0.192
	0.5	GEE* without Robust SEs	-0.001	0.004	0.577	-0.002	0.005	0.581	-0.002	0.006	0.586
		GEE*-Center	-0.002	0.004	0.241	-0.002	0.005	0.265	-0.002	0.006	0.284
		GEE*-Center by Trt	-0.002	0.004	0.170	-0.002	0.005	0.187	-0.002	0.006	0.200
	0.25	GEE* without Robust SEs	-0.002	0.004	0.565	-0.003	0.005	0.569	-0.000	0.005	0.571
		GEE*-Center	-0.002	0.004	0.237	-0.003	0.005	0.260	0.000	0.005	0.278
		GEE*-Center by Trt	-0.002	0.004	0.167	-0.003	0.005	0.183	0.000	0.005	0.197
	0.75	GEE* without Robust SEs	-0.002	0.004	0.601	-0.001	0.005	0.605	-0.003	0.006	0.609
		GEE*-Center	-0.000	0.004	0.239	0.000	0.004	0.257	-0.002	0.005	0.266
		GEE*-Center by Trt	-0.000	0.004	0.169	0.000	0.004	0.182	-0.002	0.005	0.188

Table A. 4. Properties of point estimated PF for true PF was 0.4 using three additional GEE models under homogeneous and heterogenous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and the number of centers (5 *vs* 25).

			Variance of Center-by-Treatment Effect								
			0			0.4			0.8		
n_c	ICC	Method	Bias	MSE	Length	Bias	MSE	Length	Bias	MSE	Length
5	0	GEE* without Robust SEs	-0.016	0.010	0.639	-0.020	0.014	0.646	-0.028	0.019	0.656
		GEE*-Center	-0.016	0.010	0.327	-0.020	0.014	0.394	-0.028	0.019	0.445
		GEE*-Center by Trt	-0.016	0.010	0.236	-0.020	0.014	0.284	-0.028	0.019	0.320
	0.5	GEE* without Robust SEs	-0.054	0.018	0.691	-0.058	0.021	0.699	-0.071	0.027	0.715
		GEE*-Center	-0.054	0.018	0.413	-0.058	0.021	0.459	-0.071	0.027	0.499
		GEE*-Center by Trt	-0.054	0.018	0.297	-0.058	0.021	0.329	-0.071	0.027	0.356
	0.25	GEE* without Robust SEs	-0.030	0.013	0.659	-0.037	0.016	0.669	-0.041	0.021	0.675
		GEE*-Center	-0.030	0.013	0.372	-0.037	0.016	0.425	-0.041	0.021	0.471
		GEE*-Center by Trt	-0.030	0.013	0.269	-0.037	0.016	0.305	-0.041	0.021	0.338
	0.75	GEE* without Robust SEs	-0.097	0.027	0.754	-0.106	0.033	0.775	-0.115	0.038	0.789
		GEE*-Center	-0.097	0.027	0.451	-0.104	0.032	0.492	-0.116	0.036	0.516
		GEE*-Center by Trt	-0.097	0.027	0.322	-0.104	0.032	0.350	-0.116	0.036	0.367
25	0	GEE* without Robust SEs	-0.015	0.004	0.394	-0.022	0.005	0.399	-0.027	0.006	0.403
		GEE*-Center	-0.015	0.004	0.234	-0.022	0.005	0.255	-0.026	0.006	0.274
		GEE*-Center by Trt	-0.015	0.004	0.170	-0.022	0.005	0.185	-0.026	0.006	0.199
	0.5	GEE* without Robust SEs	-0.056	0.008	0.426	-0.063	0.009	0.431	-0.074	0.011	0.440
		GEE*-Center	-0.054	0.008	0.261	-0.063	0.009	0.275	-0.073	0.011	0.288
		GEE*-Center by Trt	-0.054	0.008	0.188	-0.063	0.009	0.198	-0.073	0.011	0.207
	0.25	GEE* without Robust SEs	-0.030	0.005	0.406	-0.037	0.006	0.411	-0.044	0.007	0.416
		GEE*-Center	-0.030	0.005	0.248	-0.037	0.006	0.268	-0.045	0.007	0.282
		GEE*-Center by Trt	-0.030	0.005	0.179	-0.037	0.006	0.194	-0.045	0.007	0.204
	0.75	GEE* without Robust SEs	-0.106	0.016	0.461	-0.115	0.019	0.466	-0.126	0.022	0.473
		GEE*-Center	-0.111	0.017	0.264	-0.120	0.020	0.279	-0.131	0.022	0.287
		GEE*-Center by Trt	-0.111	0.017	0.189	-0.120	0.020	0.199	-0.131	0.022	0.205



(A)



(B)

Figure A. 1. Average point estimated PF with empirical simulation standard deviation (SD) from three additional GEE models under both homogeneous or heterogeneous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and the number of centers (5 *vs* 25). (A) Scenarios: true PF was 0. (B) Scenarios: true PF was 0.4.

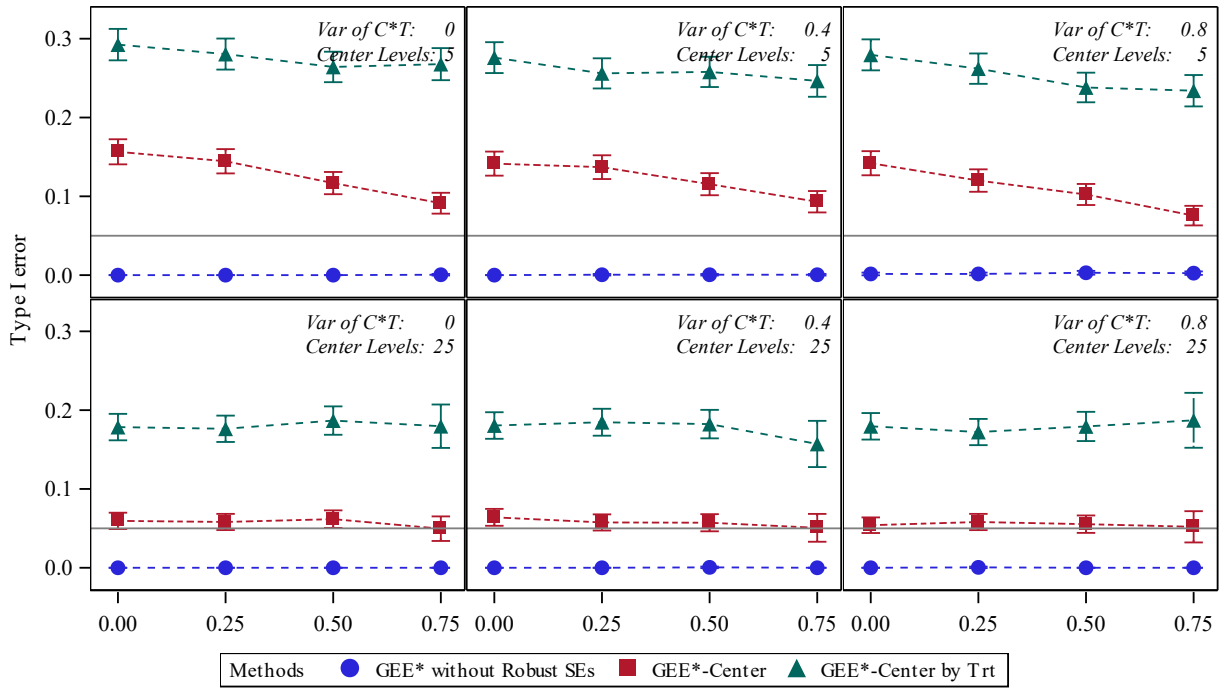


Figure A. 2. Type I error rate from three additional GEE models under both homogeneous or heterogeneous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and the number of centers (5 *vs* 25).

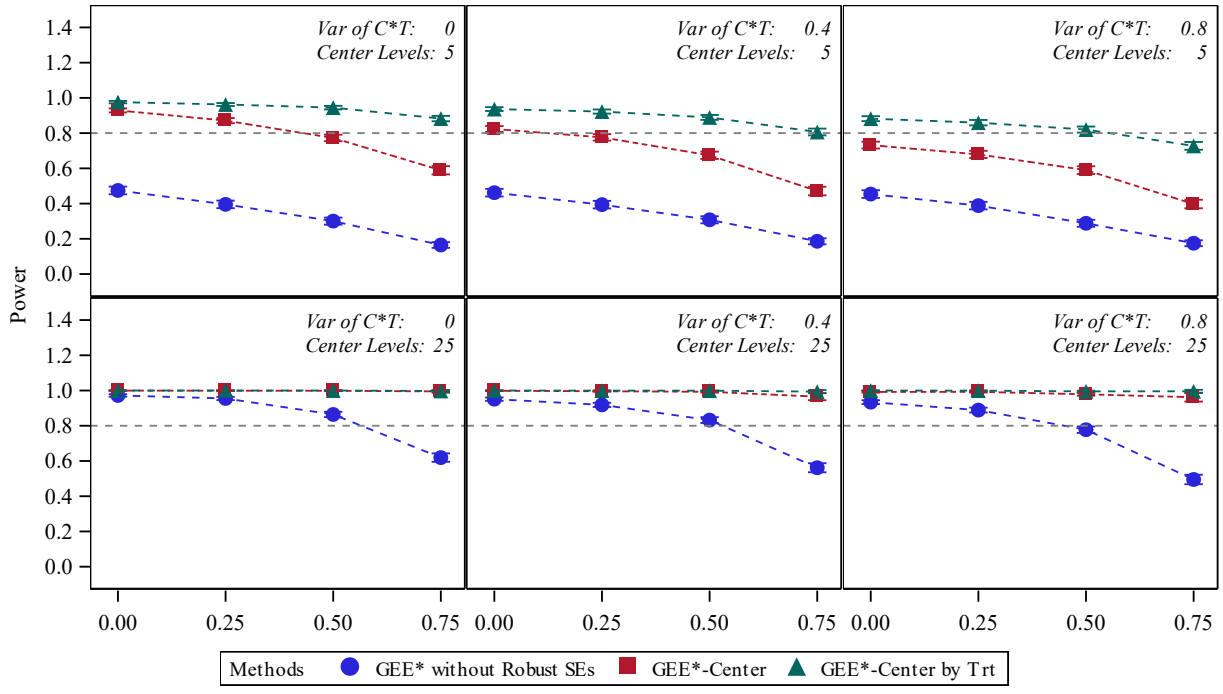


Figure A. 3. Power from three additional GEE models under both homogeneous or heterogeneous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and the number of centers (5 *vs* 25).

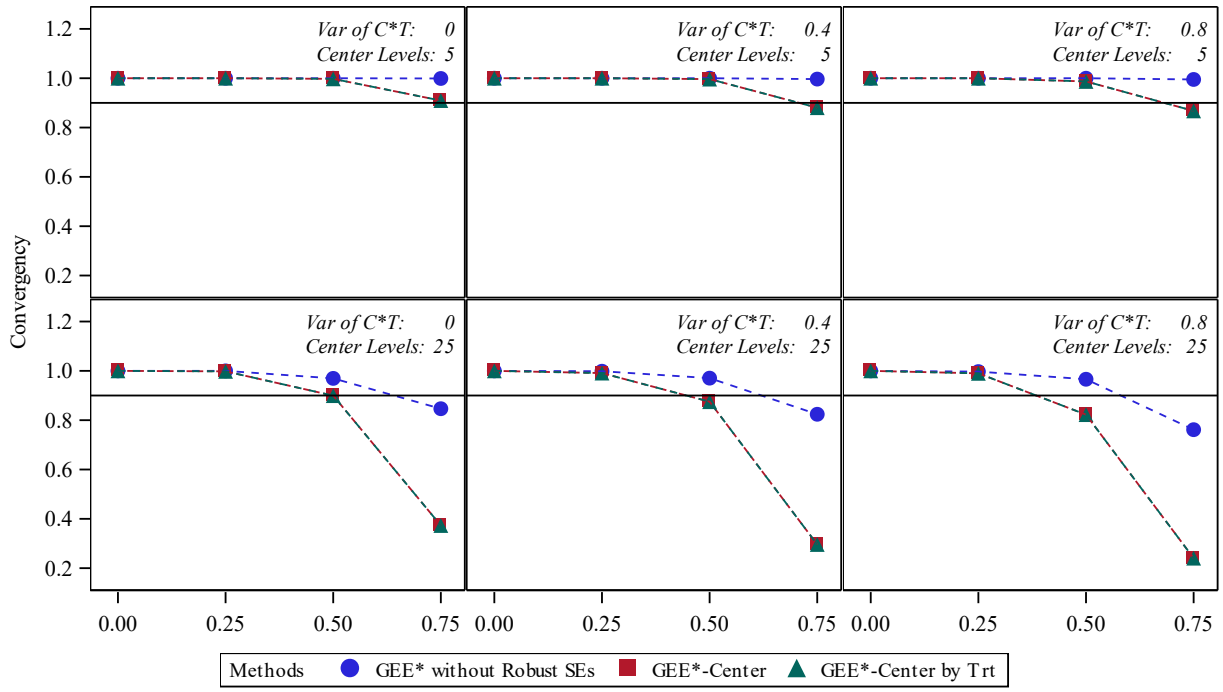


Figure A. 4. The convergency for three additional GEE models under both homogeneous or heterogeneous scenarios with different ICC values (0, 0.25, 0.5, and 0.75) and the number of centers (5 *vs* 25).