Bayesian Latent Models and Causal Inference for Biological and Health Experiments

by

Patrick Gravelle

H.BSc., Queen's University, 2018

S.M., Harvard University, 2019

A Dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Biostatistics at Brown University

Providence, Rhode Island

May 2024

This dissertation by Patrick Gravelle is accepted in its present form
by the Department of Biostatistics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date ______________ ______________________________________

Roee Gutman, Advisor

Recommended to the Graduate Council

Date ______________ ______________________________________

Constantine Gatsonis, Reader

Date ______________ ______________________________________

Vincent Mor, Reader

Approved by the Graduate Council

Date ______________ ______________________________________

Thomas A. Lewis, Dean of the Graduate School

Patrick L. Gravelle

121 South Main St., Providence, RI 02903
Email: patrick_gravelle@brown.edu

EDUCATION

2019 – Present **Brown University** **Providence, RI**
Ph.D. in Biostatistics

2018 – 2019 **Harvard University** **Cambridge, MA**
S.M. in Biostatistics

2014 – 2018 **Queen's University (GPA: 3.9/4.0)** **Kingston, ON, Canada**
H.BSc. in Statistics

PROFESSIONAL EXPERIENCE

Analysis Group — Associate Intern *Montreal — Summer 2023*

◦ Consulting for major US pharmaceutical clients on real-world observational studies using claims data.

◦ Programming in SAS to develop statistical analyses, patient selection, and data cleaning.

◦ Writing of results for the study report to be completed for each client project.

RESEARCH EXPERIENCE

Bayesian Causal Inference & Missing Data *Brown University — July 2021 – Present*

◦ Designed a Bayesian adaptive trial to address noncompliance through a causal inference framework using Markov Chain Monte Carlo (MCMC) methods.

◦ Constructed and coded Metropolis-Hastings algorithms within a Gibbs Sampler in R, unique for each model, to impute missing data and perform posterior inference.

◦ Defined, simulated, and tested 16 distinct causal models for multivariate outcomes with multiple unit-level compliance variables, against conventional methods.

Bayesian Hierarchical Modelling *Brown University — Sept 2020 – Mar 2022*

◦ Analyzed a longitudinal dataset in R using MCMC methods through Stan, a probabilistic programming language that utilizes the No-U-Turn Sampler (NUTS) similar to the Hamiltonian Monte Carlo (HMC).

◦ Developed a zero-inflated hierarchical generalized Dirichlet multinomial regression model with cyclic splines to determine a genotype effect in mice on the time spent performing physical behaviors.

◦ Collaborated and communicated with a team of biologists and computer scientists to incorporate subject-matter expertise and identify significant phenotypes.

Cluster Randomized Trials & Simulation *Brown University — July 2020 – Feb 2022*

◦ Defined a factorial design simulation to consider 17,496 configurations of CRT data generation through a cumulative 8,748,000 iterations in R, analyzing model misspecification for the marginalized treatment effect in cluster randomized trials.

◦ Multilevel/hierarchical models, Generalized Estimating Equation models, and models from CRT literature were considered and compared, with model performance assessed through coverage probability, mean interval width, and mean squared error of the true treatment effect.

PUBLICATIONS

Working

Gravelle, P., Gutman, R., *Methods to adjust for stratification variables in stratified cluster randomized trials.*

Gravelle, P., Gutman, R., *Zero-inflated generalized Dirichlet multinomial Bayesian regression model with cyclic splines for analysis of TDP-43 on the ALS-FTD spectrum.*

Gravelle, P., Gutman, R., *Designing a Bayesian adaptive trial to address noncompliance.*

Madigan, L., **Gravelle, P.**, et al., *Automated Continuous Behavioral Monitoring reveals overlapping and unique early phenotypes in two knock-in mouse models of ALS..*

Computer Software and Language Skills

Advanced R, Stan, LaTeX, SAS, SQL

Intermediate JAGS, Python

Familiar SPSS, GitHub

TEACHING EXPERIENCE

2023
Sep – Dec
Brown University
Principles of Biostatistics and Data Analysis Teaching Assistant
◦ Responsible for grading homework assignments, midterm and final examinations, and hosting weekly office hours.

2021
Jan – May
Brown University
Generalized Linear Models (PhD Level Course) Teaching Assistant
◦ Responsible for grading homework assignments, midterm and final examinations, and hosting weekly office hours.

2020
Sep – Dec
Brown University
Linear Models (PhD Level Course) Teaching Assistant
◦ Responsible for grading homework assignments, midterm and final examinations, and hosting weekly office hours.

2017 – 2018
Sep – Apr
Queen's University
Linear Algebra Teaching Assistant
◦ Responsible for grading undergraduate weekly homework assignments.

AWARDS

2018 Biostatistics Department Fellowship
Harvard T.H. Chan School of Public Health

2017, 2018 Natural Sciences and Engineering Research Council of Canada (NSERC) Undergraduate Student Research Award
Lakehead University

2017 Dr. George L. Edgett Memorial Scholarship in Statistics
Queen's University
◦ Awarded to the 4th Year Honours student with the highest academic achievement in Statistics

PRESENTATIONS

2022 Aug **Joint Statistical Meeting**
Washington, D.C.
◦ Oral Presentation (Accepted) of *Zero-inflated generalized dirichlet multinomial Bayesian regression model with cyclic splines for analysis of TDP-43 on the ALS-FTD spectrum.*

2021, 2022 Jun **World Meeting of the International Society for Bayesian Analysis**
University of Connecticut, McGill University
◦ Oral, Poster Presentations of *Zero-inflated generalized dirichlet multinomial Bayesian regression model with cyclic splines for analysis of TDP-43 on the ALS-FTD spectrum.*

2021 Sep **New England Statistics Symposium**
University of Rhode Island
◦ Poster Presentation of *Adjusting for stratification variables in cluster randomized trials.*

2021 Aug **Joint Statistical Meeting**
Virtual Conference
◦ Oral Presentation of *Adjusting for stratification variables in cluster randomized trials.*

2019 Apr **Canadian Undergraduate Mathematics Conference**
Queen's University
◦ Oral Presentation introducing the subject of Biostatistics to undergraduate students, detailing an application in concussion symptom recovery.

2018 Apr **New England Science Symposium**
Harvard Medical School
◦ Poster Presentation of *Emergency medical services response times to motor vehicles crashes increased over the period 1987 to 2015.*

2017, 2018 Mar **Inquiry at Queen's Undergraduate Research Conference**
Queen's University
◦ Oral Presentation of *Emergency medical services response times to motor vehicles crashes increased over the period 1987 to 2015.*
◦ Oral Presentation of undergraduate summer research project *Comparing Golf Eras.*
◦ Oral Presentation of *The Reality of Olympic Sprinting.*

2017 Sep **New England Symposium on Statistics in Sports**
Harvard University
◦ Poster Presentation of undergraduate summer research project *Comparing Golf Eras.*

Acknowledgements

I would like to thank the faculty, staff, and students in the Biostatistics Department as well as those in the School of Public Health for their support and importance during my time at Brown. It is a community filled with wonderful and caring individuals.

I would like to thank my advisor, Professor Roee Gutman, for his unwavering support and guidance throughout this process. He has encouraged and pushed me to reach beyond the potential and work ethic I originally set out for myself, and it is something of which I am tremendously appreciative. I thank Professor Constantine Gatsonis for his expertise and thoughtfulness in both his teaching of our first year courses and then through his involvement on my dissertation committee. I thank Professor Vincent Mor for taking the time to be on my dissertation committee and offering his valuable insights and perspectives.

Thank you to all of my family and friends, as this degree would not have been possible without your unconditional support and encouragement. I am so fortunate and grateful to be surrounded by each and every one of you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In practice, it is rare that the true generating mechanism of a data source is known. This uncertainty presents a challenge to statisticians, who are commonly tasked with specifying and identifying relationships in a given dataset. However, this challenge is not unique, nor equal, for any particular dataset. Investigators may encounter these challenges when modelling data in a variety of settings, such as those comparing the effects of interventions between groups of individuals. Estimating the effects of interventions is often the primary objective in controlled trials or experiments, but may also be the goal of an observational study. Generally, controlled trials are preferred to observational studies when examining intervention effects as they typically have the benefit of known sampling and randomization mechanisms. These mechanisms aim to provide investigators with groups of trial participants that are similar to one another on average, which generally helps to appropriately and accurately estimate the intervention effects. However, even under controlled settings, there may be latent structures determining the data generation that are not easily discovered or known. The presence and complexity of these structures likely varies for each dataset and the variables within it. Examining the relationships within a dataset in the presence of latent structures may prove challenging. Moreover, specifying statistical models for this type of data requires assumptions and methods that should be tested for their robustness to deviations or violations of these assumptions. Throughout this dissertation, we are interested in methodological developments and comparisons for latent data generating mechanisms across a variety of controlled trial or experiment settings. In particular, we focus on three settings: 1) Stratified Cluster Randomized Trials, 2) Adaptive Randomized Controlled Trials, and 3) Longitudinal Controlled Experiments.

1.1 Aim 1

Randomized controlled trials (RCTs) are the gold-standard for estimating the effects of interventions. While adjustment for covariates in RCTs leads to more powerful statistical procedures, multiple systematic reviews have shown that many RCTs do not adjust for covariates beyond the intervention. In cluster randomized trials (CRTs) groups of subjects, rather than individuals, are randomly assigned to interventions. When there exists a variable believed to be prognostic of the outcome of interest, stratifying the randomization by this variable may lead to improved balance between intervention arms.

In the analysis phase of a stratified CRT (SCRT), generalized linear multilevel models (GLMM) and generalized estimating equations (GEE) have been proposed to adjust for individual-level and cluster-level covariates as well as for the correlations between individuals within the same group. When a SCRT is implemented, the exact specification of GLMM and GEE models that account for stratification variables have not been extensively examined. Aim 1 is to examine the operating characteristics of different specifications of GEE and GLMM models in SCRTs through extensive simulations. The simulations include binary and continuous outcome variables, different outcome error configurations, linear and non-linear relationships between the covariates and outcomes, as well as different number of clusters and individuals within clusters.

1.2 Aim 2

When RCT participants adhere to their assigned intervention, the effects of randomization approximates the effects of the intervention. However, in the presence of noncompliance, assignment to the intervention may not approximate well the receipt of the intervention. Noncompliance to the assigned intervention may become even more significant in pragmatic RCTs, in which researchers have less control over the administration of the intervention. Numerous methods have been developed to address lack of compliance following the collection of data from an RCT as part of the analysis stage. However, limited experimental designs have been developed to address possible noncompliance while the trial is ongoing.

Adaptive designs provide a framework to specify possible modifications during the design stage of a randomized trial. In such designs, investigators can evaluate interventions as data are accrued and apply adaptations to the trial. Aim 2 proposes an adaptive design to address noncompliance with multi-component intervention. We frame the design within the counterfactual causal inference framework, and describe both design and analysis procedures to implement this design.

1.3 Aim 3

Amyotrophic lateral sclerosis (ALS) is a progressive neurological condition impacting nerve cells within the brain and spinal cord. ALS is characterized by motor function impairment, and many ALS patients also experience cognitive and behavioral symptoms that resemble those of frontotemporal dementia. The ALS disease spectrum can be characterized by the RNA-binding protein, TDP-43. Defining the role of TDP-43 in the neurodegeneration process can lead to treatment development of ALS. Animal models enable researchers to manipulate genetics and environmental factors to investigate their roles in development, behavior, and health outcomes of a disease. Mice models are frequently used because of their phylogenetic proximity and physiological resemblance to humans. Many studies that utilize mouse models to examine behavior, motor, and cognitive functions are generally performed using data that are collected over short time periods while the mice undergo a single test. The automated home-cage behavioral phenotyping of mice (ACBM) (Jhuang et al. (2010)) is a computer vision system that is trained with manually annotated behaviors of interest and tracks these behaviors in freely behaving mice in cages. The ACBM enables investigators to record and analyze multiple behaviors over multiple days for multiple mice.

Analysis of Variance (ANOVA) is a commonly used statistical tool to compare two or more types of animals across multiple behavioral tasks. However, The ANOVA model may produce unreliable inference and predictions when the results of a test are correlated over time and within mice, as well as when the outcomes are non-Normally distributed. Following a Bayesian framework, we compare the goodness of fit of multiple models to data annotated by the ACBM algorithm over 5 days for 20 mice. Aim 3 proposes a hierarchical zero-inflated generalized Dirichlet multinomial regression model (abbreviated ZIHGDM) (Tang and Chen (2018)) with cyclic splines to model the time mice spent performing certain behaviors at each circadian hour. This model can integrate periods in which mice do not perform certain behaviors, correlation between the time spent performing the different behaviors, repeated measurement for each mouse and day of the week, and cyclic effects of hours within a day.

Chapter 2

Methods to adjust for stratification variables in Stratified Cluster Randomized Trials

2.1 Introduction

Randomized controlled trials (RCTs) are commonly used study designs for estimating the effects of interventions. Studies that estimate effects of interventions should comprise of two distinct stages: the design stage and the analysis stage (Rubin, 2008). The design stage of a RCT is commonly described in a study protocol which documents the aims, procedures, and policies of the study. This protocol also describes the randomization procedure and the statistical analysis that would be employed at the conclusion of the trial. The analysis stage implements the statistical analysis described in the study protocol (Rosenberger and Lachin, 2015).

One of the objectives of randomization is to balance all observed and unobserved baseline covariates across the arms of the study (Rosenberger and Lachin, 2015). However, randomization only balances covariates on average (Morgan and Rubin, 2012). When there are covariates that are known to be associated with the outcomes, stratified randomization is a design tool that can be used to increase covariates balance (Kernan et al., 1999). Stratified randomization has been shown to increase the precision of treatment effect estimates (Meier, 1981; Ye et al., 2023; Kernan et al., 1999; Lachin, 1988). The increase in precision is expected to be larger in RCTs with small number of participants compared to RCTs with many participants (Kernan et al., 1999; Lachin, 1988).

A different approach to address possible imbalances and increase the efficiency of a study is to adjust for covariates at the analysis stage. Adjusting for covariates in RCTs generally leads to more powerful statistical procedures (Ivers et al., 2012). However, several systematic reviews have shown that only one-quarter to one-third of the RCTs adjust for baseline covariates (Yu et al., 2010; Assmann et al., 2000; Austin et al., 2010; Hernandez et al., 2005). One reason for the lack of covariate adjustment is the possibility of reduced power when these covariates are not highly correlated with the outcomes (Kahan et al., 2014). A simulation study of different RCT settings found that adjustments for covariates that are not correlated with the outcomes resulted in a small decrease in power compared to possible large gains in power when adjusting for covariates that are highly correlated with the outcomes (Kahan et al., 2014). A different reason for the lack of covariate adjustment is that investigators may be concerned with model misspecification when adjusting for covariates beyond the intervention indicator. However, in RCTs with unit-level randomization, the analysis of covariance (ANCOVA) estimator was shown to have a consistent point estimate and consistent standard error even when the model is misspecified (Wang et al., 2019). In addition, if any of the covariates are correlated with the outcome, ANCOVA based estimates are more precise (Yang and Tsiatis, 2001).

A third approach to address possible imbalances combines stratification and covariates adjustment to obtain balanced covariates at the design stage and improved efficiency at the analysis stage (Kahan and Morris, 2011). In simulations, this approach was shown to provide valid and efficient estimates with continuous, binary and time to event outcomes (Kahan and Morris, 2011). Moreover, not adjusting for randomization strata in the analysis stage may result in conservative standard error estimates under stratified randomization procedures (Wang et al., 2021).

In cluster randomized trials (CRTs), groups of subjects, rather than individuals, are randomly assigned to the interventions. Randomization at the cluster level may be preferred because of economic or logistical reasons (Donner and Klar, 2000). CRTs can also reduce experimental contamination, which may occur when individuals assigned to different treatment arms are in frequent contact (Donner and Klar, 2004). Compared to individual-level randomization, CRTs may suffer from loss of efficiency, because outcomes of individuals within a cluster are more similar than outcomes of individuals across clusters.

Because in many CRTs the number of clusters is relatively small, stratification on important prognostic factors has been suggested to reduce imbalances between arms (Donner, 1998). Moreover, because stratification is expected to reduce between-cluster variability, stratified randomization is expected to increase precision (Hayes and Moulton, 2017). Strata in cluster randomized trials are commonly defined using baseline cluster-level characteristics that are expected to be correlated with the outcomes (Raab and Butcher, 2001; Donner and Klar, 2000).

Regression adjustments with CRTs require advanced models, because they commonly include both individual-level and cluster-level covariates as well as accounting for the correlations of individuals within clusters. Two primary methods are commonly used for analysis of CRTs (Turner et al., 2017a,b; Campbell et al., 2006). The first method is the generalized linear multilevel model (GLMM) approach where the first level represents the individual patients, and it is nested within a second level that adjusts for cluster effects (Hayes and Moulton, 2017). The second method is the generalized estimating equations (GEE) which yields population-level inference for the parameters in the model (Liang and Zeger, 1986; Campbell et al., 2006; Turner et al., 2017b). However, interpretation of the parameters from the two models are not equivalent with similar subject-level covariate specification. Therefore, defining a common estimand of interest is important when comparing model performance between the GLMM and GEE approaches.

Although stratification is commonly used in CRTs (Lewsey, 2004; Crespi, 2016), less attention has been given to analysis of stratified CRTs (SCRTs). Turner et al. (2017b) suggested that ignoring stratification in the analysis of CRTs will result in statistically valid procedure that can be more powerful because of the increased degrees of freedom. This advice differs from the ANCOVA result for individual randomized trials if the stratification variables are correlated with the outcomes (Wang et al., 2019). With binary outcomes, multiple χ^2 statistics were proposed to address stratified randomization (Song and Ahn, 2003). Based on simulations, a modified Mantel–Haenszel statistic (Zhang and Boos, 1997) and a GLMM test statistics were shown to be statistically valid and more powerful compared to other test statistics. However, these tests statistics had low power when the number of clusters were small, and they do not provide an easily interpretable effect size estimates. The modified Mantel–Haenszel statistic was also extended to categorical variables (Dobbins and Simpson, 2002), but it may suffer from similar limitations as the statistic with the binary outcome. Lewsey (2004) used simulations analysis to compare the effects of stratification on cluster size in CRTs. With stratified randomization, they relied on GLMMs that included an indicator for strata membership. The analysis model for simple randomization did not adjust for the stratification variables. They describe that stratified randomization resulted in greater power compared to simple randomization. However, the estimands of the simple randomization model and the stratified randomization model differ. Borhan et al. (2023), compared GLMM and GEE models to analyze continuous outcomes in stratified cluster randomized trials. They report that the examined GEE model has larger than nominal type I error for small clusters, while the GLMM model has nominal error for small and large number of clusters. However, the estimands that were used for comparison of GEE model and the GLMM model differed.

We perform an extensive simulations to examine the operating characteristics of different specifications of GEE and GLMM models in stratified CRTs. The simulations include binary and continuous outcome variables, different outcome error configurations, linear and non-linear relationships between the covariates and outcomes, as well as different number of clusters and individuals within clusters.

Our simulations identify four main trends. First, GLMM and GEE models that adjust for strata-defining covariates are generally robust to subject-level model misspecification, and to the cluster-level distribution misspecification. These models are more precise than models that do not adjust for strata-defining covariates. Second, a GLMM that adjusts for within and between association of covariates with the outcome, is generally valid and it has operating characteristics that are generally similar to correctly specified models. Third, GEE models that do not perform sampling variance adjustment may result in invalid statistical procedures. Fourth, GEE models with sampling variance adjustment have similar, although less precise, performance compared to GLMMs with the same mean specification, as the number of clusters increases.

The paper proceeds as follows: Section 2 describes the notations and models. Section

3 describes the simulation configurations. The simulation results are presented in Section 4. Section 5 implements the methods on a study that examined the effects of a video intervention for advanced directive on hospitalizations. Section 6 provides discussion and conclusions.

2.2 Notations and Methods

2.2.1 Notation

Consider a SCRT for estimating the effect of binary intervention T on an outcome Y. Under the Stable Unit Treatment Value Assumption (Imbens and Rubin, 2015), outcome for unit $i \in \{1, \ldots, m_j\}$ from cluster $j \in \{1, \ldots, J\}$, has two potential outcomes: $Y_{ij}(1)$ when cluster j is assigned to the active treatment group, and $Y_{ij}(0)$ when cluster j is assigned to the control group. Because units can only be assigned to one of the interventions at a specific point in time, only one of these two potential outcomes can be observed. The observed outcomes are

$$Y_{ij}^{obs} = (1 - T_j) \times Y_{ij}(0) + T_j \times Y_{ij}(1),$$

where $T_j \in \{0, 1\}$ is the treatment assigned to cluster j. In addition to the observed outcomes, we record P_1 baseline unit-level covariates, $\mathbf{X}_{ij} = \{X_{1ij}, \ldots, X_{P_1 ij}\}$, and P_2 baseline cluster-level covariates, $\mathbf{Z}_{ij} = \{Z_{1ij}, \ldots, Z_{P_2 ij}\}$. The cluster-level covariates are similar for all units in the same cluster $\mathbf{Z}_{ij} = \mathbf{Z}_{i'j}, \forall\ i, i'$. Stratification variables in a SCRT, $\mathbf{Z'}_{ij}$, are generally cluster-level covariates and can be defined as a subset of $\mathbf{Z}_{ij}$, such that $\mathbf{Z'}_{ij} \subseteq \mathbf{Z}_{ij}$. Causal effects in CRTs and SCRTs are summarized using estimands, which are functions of unit-level potential outcomes over a pre-defined set of units (Rubin, 1978). A commonly used estimand is the average treatment effect (ATE), $\gamma = E(Y(1) - Y(0))$, where expectation is taken over the entire population of units. When the number of units within each cluster varies, a different estimand is the difference in average of cluster means in the treated group to the average of cluster means in the control group (Zhai and Gutman, 2021),

$$\xi = \frac{1}{\sum_{j=1}^{J} T_j} \sum_{j=1}^{J} T_j * \frac{1}{m_j} \sum_{i=1}^{m_j} Y_{ij}(T_j) - \frac{1}{\sum_{j=1}^{J}(1-T_j)} \sum_{j=1}^{J}(1-T_j) * \frac{1}{m_j} \sum_{i=1}^{m_j} Y_{ij}(T_j).$$

The methods that are presented here can be used to estimate γ and ξ; however, for simplicity, we will concentrate our derivations and simulations for γ.

2.2.2 Generalized Linear Multilevel Models

A common model for analyzing a CRT is the generalized linear multilevel model (GLMM) (McCulloch, 2003). In GLMMs, the mean of the potential outcomes conditional on a parameter vector $\boldsymbol{\theta}$, $E[Y_{ij}(T_j)|\boldsymbol{\theta}]$, can be related to a set of covariates, $\{\mathbf{X}_{ij}, \mathbf{Z}_{ij}, T_j\}$, through a link function $g(\cdot)$,

$$\begin{aligned} E[Y_{ij}(T_j)|\boldsymbol{\theta}] &= \mu_{ij}(T_j) \\ g(\mu_{ij}(T_j)) &= h\big(\mathbf{Z}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\beta}, T_j\big) + c_j \\ c_j &\sim \mathcal{N}(0, \tau^2), \end{aligned} \tag{2.1}$$

where $\beta \subseteq \theta$ is a set of unknown parameters and c_j represents the effect of cluster j. The identity link function, $g\left(\mu_{ij}(T_j)\right) = \mu_{ij}(T_j)$, is commonly used with continuous outcomes, and the logit link, $g\left(\mu_{ij}(T_j)\right) = \log\left(\frac{\mu_{ij}(T_j)}{1-\mu_{ij}(T_j)}\right)$ with binary outcomes. The form of $h(\cdot)$ is study specific and we describe possible forms for $h(\cdot)$ in the Sections 2.2.2 - 2.2.2. For continuous outcomes we assume $Y_{ij}(T_j) \sim N\big(\mu_{ij}(T_j), \sigma_\epsilon^2\big)$, and $\theta = \{\beta, \sigma_\epsilon^2\}$. To estimate γ, we calculate $\frac{1}{\sum_{j=1}^{J} m_j} \sum_{ij} \int \left(\hat{\mu}_{ij}(1) - \hat{\mu}_{ij}(0)\right) dP_{c_j}$, where $\hat{\mu}_{ij}(T_j)$ are the estimates of $\mu_{ij}(T_j)$ from Equation (2.1) and P_{c_j} is the distribution of c_j. GLMMs can be estimated using the R software packages lme (Bates et al., 2015) or nlme (Pinheiro et al., 2023) as well as PROC GLIMMIX or PROC MIXED in SAS (Turner et al., 2017b).

In CRTs when assuming a multilevel analysis model, the sampling variability of a continuous outcome can be factored into three components, the residual variance after adjustment for the observed covariates $\mathbf{Z}_{ij}$ and $\mathbf{X}_{ij}$, the cluster-level variability, and the predicted variance based on observed covariates,

$$V(Y_{ij}(T_j)) = V\Big(Y_{ij}(T_j) - c_j - h\big(\mathbf{Z}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\beta}, T_j\big)\Big) + V\Big(c_j\Big) + V\Big(h\big(\mathbf{Z}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\beta}, T_j\big)\Big) \quad (2.2)$$

Adjusting for covariates that are correlated with the outcomes can lead to reductions in the residual variance term in Equation (2.2). However, compared to RCTs, this reduction may not be as large in CRTs because it does not influence the cluster-level variability, $V(c_j)$. A derivation of the variance decomposition for a CRT is included in Supplementary Material 2.7.1.

Unadjusted Multilevel Models

One model specification for analyzing RCTs with individual-level randomization does not adjust for covariates beyond the intervention (Yu et al., 2010; Assmann et al., 2000; Austin et al., 2010; Hernandez et al., 2005). In CRTs, a corresponding model assumes that $h(\cdot) = \beta_1 T_j$ and

$$g\left(\mu_{ij}(T_j)\right) = c_j + \beta_1 T_j, \quad (2.3)$$

where β_1 is the conditional treatment effect on the link function scale. Models that do not adjust for covariates are thought to be less prone to type I error because they do not suffer from misspecifications of the relationship between covariates and the outcome. (Kraemer, 2015).

Adjusted Multilevel Models

Under individual-level randomization, linear adjustments for covariates leads to more powerful statistical procedures (Balzer et al., 2016). In CRTs, a corresponding GLMM is,

$$g\left(\mu_{ij}(T_j)\right) = c_j + \beta_1 T_j + \sum_{p=1}^{P_1} \beta_{2p} X_{pij} + \sum_{p=1}^{P_2} \beta_{3p} Z_{pij}, \quad (2.4)$$

where $\boldsymbol{\beta} = (\beta_1, \{\beta_{2p}\}, \{\beta_{3p}\})$. A possible limitation of Equation (2.4) is that models with different sets of $\mathbf{X}_{ij}$ may result in different sampling error estimates (Beach and Meier, 1989). In addition, Equation (2.4) does not consider the variability of the covariates across clusters, which may influence the outcomes. Throughout, we will refer to this model as GLMM.

Neuhaus and Kalbfleisch Model

Neuhaus and Kalbfleisch (1998) suggested a model that adjusts for baseline covariates while considering their between- and within- cluster associations with the outcome. Formally,

$$g\left(\mu_{ij}(T_j)\right) = c_j + \beta_1 T_j + \sum_{p=1}^{P_1} \left(\beta_{2p}\overline{X}_{pj} + \beta_{3p}\left(X_{pij} - \overline{X}_{pj}\right)\right) + \sum_{p=1}^{P_2} \beta_{4p} Z_{pij}, \tag{2.5}$$

where $\boldsymbol{\beta} = (\beta_1, \{\beta_{2p}\}, \{\beta_{3p}\}, \{\beta_{4p}\})$ and $\overline{X}_{pj} = \frac{1}{m_j}\sum_{i=1}^{m_j} X_{pij}\ \forall p, j$. When the associations between the covariates and the outcomes at the individual and cluster level differ, this model can result in more precise estimates of γ. Throughout, we will refer to this model as NKM.

2.2.3 Generalized Estimating Equations

Generalized Estimating Equations (GEE) have been proposed as an alternative to GLMMs for estimating treatment effects in clustered or longitudinal data (Zeger et al., 1988). In GEE, the model for the marginal distribution of $Y_{ij}(T_j)$ does not require specification of the joint distribution of subject's observations (Liang and Zeger, 1986). Define covariance matrix $\mathbf{V}_j(\boldsymbol{\alpha}) = \mathbf{A}_j^{1/2}\mathbf{R}_j(\boldsymbol{\alpha})\mathbf{A}_j^{1/2}$, where $Cov(\mathbf{Y}_{ij}(T_j)) = \mathbf{A}_j \times \phi$, with an unknown scale parameter ϕ, and $\mathbf{A}_j = diag\{g(\mu_{1j}), \ldots, g(\mu_{m_j j})\}$. $\mathbf{R}_j(\boldsymbol{\alpha})$ is a correlation matrix that may depend on a vector of unknown parameters, $\boldsymbol{\alpha}$ (Zeger et al., 1988). We estimate the unknown parameters, $\boldsymbol{\beta}$, by solving the following equations,

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{j=1}^{J} \frac{\partial \boldsymbol{\mu}_{ij}(T_j)}{\partial \boldsymbol{\beta}} \mathbf{V}_j^{-1}(\boldsymbol{\alpha})(\mathbf{Y}_{ij}(T_j) - \boldsymbol{\mu}_{ij}(T_j)) = \mathbf{0}, \tag{2.6}$$

where $\boldsymbol{\mu}_{ij}(T_j) = E(\mathbf{Y}_{ij}(T_j)) = \{g^{-1}(h(\mathbf{Z}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\beta}, T_j)), \ldots, g^{-1}(h(\mathbf{Z}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\beta}, T_j))\}$. GEE requires correct specification of the marginal mean but not the variance-covariance matrix to obtain consistent estimates (Liang and Zeger, 1986). Misspecifying the mean may result in biased and inconsistent estimates (Emond et al., 1997).

The sampling variance of the parameter $\boldsymbol{\beta}$ is usually estimated using the robust "sandwich" estimator (Liang and Zeger, 1986). When the number of clusters is small, this "sandwich" estimator has been shown to be biased (Mancl and DeRouen, 2001; Li and Redden, 2014; Huang et al., 2016) and can lead to inflated type I error rate and below nominal coverage (Kahan et al., 2016; Leyrat et al., 2018). Multiple procedures have been proposed to adjust the sampling covariance matrix of GEE estimates in small samples (Gosho et al., 2021; Ford and Westgate, 2018; Wang et al., 2016). Gosho et al. (2021), identified the bias-adjustment proposed by Mancl and DeRouen (2001) as the best performing method for mixed effects models with repeated measures. This adjustment method reduces the bias of the residual estimator, $\boldsymbol{r}_{ij}\boldsymbol{r}_{ij}^T = (\mathbf{Y}_{ij}(T_j) - \hat{\boldsymbol{\mu}}_{ij}(T_j))(\mathbf{Y}_{ij}(T_j) - \hat{\boldsymbol{\mu}}_{ij}(T_j))^T$, by estimating the expectation of the squared first-order Taylor series expansion of $\boldsymbol{r}_{ij}$ around β. Lu et al. (2007) reported that this adjustment generally overestimates variances, but results in nominal coverage with 10 clusters or more. However, for less than 10 clusters an adjustment by Kauermann and Carroll (2001) may be preferred. In our simulations, we examined the performance of the different GEE models, without any adjustments, with the bias adjustments proposed by Mancl and DeRouen (2001), and with the bias adjustment proposed by Kauermann and Carroll (2001). GEE models can be estimated using the R software package geepack (Halekoh et al., 2006) or in SAS using PROC GENMOD (Turner et al., 2017b). Bias adjustments for the sampling variance can be obtained using the geesmv (Wang, 2015) package.

2.2.4 Estimating $E(Y(1))$ and $E(Y(0))$ Separately

The models in Sections 2.2.2 and 2.2.3 assume that the covariates $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ do not interact with the intervention. This implies that the intervention effect is homogeneous on the link scale. In addition, these models assume that the within- and between cluster variability are similar across intervention groups. To relax these assumptions, γ can be estimated with two separate models, $\mu_{ij}(1)$ and $\mu_{ij}(0)$, in the treated and control group, respectively. For example, under the NKM model structure,

$$g\left(\mu_{ij}(t)\right) = c_j + \sum_{p=1}^{P_1} \left(\beta_{1p}^t \overline{X}_{pj} + \beta_{2p}^t \left(X_{pij} - \overline{X}_{pj}\right)\right) + \sum_{p=1}^{P_2} \beta_{3p}^t Z_{pij}, \tag{2.7}$$

where $\boldsymbol{\beta}^1 = (\{\beta_{1p}^1\}, \{\beta_{2p}^1\}, \{\beta_{3p}^1\})$ may differ from $\boldsymbol{\beta}^0 = (\{\beta_{1p}^0\}, \{\beta_{2p}^0\}, \{\beta_{3p}^0\})$. Other models can be defined similarly. To obtain estimates for γ, we average over the covariates and cluster-specific effects in each model to obtain estimates of $E(Y_{ij}(1))$ and $E(Y_{ij}(0))$, $\hat{E}(Y_{ij}(1))$ and $\hat{E}(Y_{ij}(0))$, respectively. Point estimate of γ is derived as $\hat{\gamma} = \hat{E}(Y_{ij}(1)) - \hat{E}(Y_{ij}(0))$. Sampling variance estimates of $\hat{\gamma}$ are expected to be larger than the sampling variance based on the models in Sections 2.2.2 and 2.2.3, because the covariance between $\hat{E}(Y(1))$ and $\hat{E}(Y(0))$ is assumed to be 0 when these estimates are modeled independently. However, it may lead to more accurate estimates, and better operating characteristics when the effects are heterogeneous.

2.3 Simulations

To examine the performance of the models in Section 2 under varying levels of misspecification, we perform simulation analysis. In the simulations, we generate three covariates for individual $i \in \{1, \ldots, m_j\}$ in cluster $j \in \{1, \ldots, J\}$. The first covariate is a cluster proportion of black individuals $Z_{1j} \sim U(0, 1)$. The second variable is a binary individual-level black indicator, $X_{1ij} \sim Bernoulli(Z_{1j})$, and the third variable is an individual-level standardized cholesterol, $X_{2ij} \sim N(\Delta, 1)$ where $\Delta \in \{0, 1\}$. Given individual- and cluster-level covariates we generate the potential outcomes $Y_{ij}(T_j)$ based on Equation (2.1), with

$$g\big(\mu_{ij}(T_j)\big) = c_j + \beta_1 T_j + (\beta_2 + \beta_3 \nu_j)\, Z_{1j} + \beta_4 X_{1ij} + \beta_5 exp\left(\frac{X_{2ij}}{\beta_6}\right) + \beta_7 \overline{X}_{2j} + \beta_8 X_{2ij} + \beta_9 T_j X_{2ij}, \tag{2.8}$$

where β_k, $k = 1, \ldots, 9$, are pre-defined coefficients, $c_j \sim F$ is the cluster specific mean, $\nu_j \sim N(0,1)$ is the cluster specific change in slope for Z_{1j}, $\overline{X}_{2j} = \frac{1}{m_j}\sum_{i=1}^{m_j} X_{2ij}$ with m_j individuals in cluster j, and $g(\cdot)$ is the identity link function. Each coefficient $\beta_1, \ldots, \beta_9$ represent a relationship of the covariates and treatment assignment with the potential outcomes: β_1 represents the linear relationship with the treatment assignment, β_2 represents the linear relationship between cluster proportion of black individuals, β_3 represents the linear relationship with cluster-specific proportion of black individuals, β_4 represents the linear relationship with individual level race indicator, β_5 represents the non-linear relationship with individual level cholesterol, β_6 represents the degree of non-linear relationship with individual level cholesterol, β_7 represents the linear relationship with cluster average cholesterol, β_8 represents the linear relationship with individual-level cholesterol, and β_9 represents the heterogeneity of the treatment assignment effect for different values of the individual-level cholesterol.

We construct 36 simulation sets such that each follows a factorial design using the three covariates and the model in Equation 2.8. The 36 simulation sets are defined by unique values of 4 elements: the randomization type, the inclusion of treatment effect heterogeneity across clusters, the outcome type (continuous/binary) and model specification. We examine three randomization types: simple randomization (SR0), stratified randomization constructed with one variable (SR1), and stratified randomization with two variables (SR2). The proportion of black individuals in the cluster, Z_{1j}, was used as the first stratifying variable. We stratified clusters based on Z_{1j} to either 2, 3, or 4 using the respective quantiles of Z_{1j}. The average cholesterol level in the strata, $\overline{X}_{2j}$, was used as the second stratification variable. All clusters above the median of average cholesterol were assigned to one strata, and the rest were assigned to the second strata. The randomization strata element was examined with either no stratification, stratification on Z_{1j}, resulting in 2, 3, or 4 strata, or the combination of the two stratification variables, resulting in 4, 6, or 8 strata. Within each strata clusters were randomly assigned with 1:1 ratio to either the intervention group or the control group. In sets that used the two covariates for stratification, β_5 and β_7 were different from 0. Simple randomization was examined with either $\beta_5 = \beta_7 = 0$ or with both different from 0.

The second element that we examined was the introduction of treatment effect heterogeneity across clusters. Formally, we set $\beta_9 = 0$ for a constant homogeneous effect on the link scale (HTE0), and set $\beta_9 \neq 0$ for heterogeneous effect (HTE1). The third element defining the simulation is the outcome type. We examined two types of outcomes: continuous (CO) and binary (BO). Formally, continuous outcomes were generated as $Y_{ij}(T_j) \sim N\big(\mu_{ij}(T_j), \sigma^2_\epsilon = 1\big)$ and binary outcomes were generated as $Y_{ij}(T_j) \sim \text{Ber}\big(\text{logit}^{-1}(\mu_{ij}(T_j))\big)$.

The model specification element consists of 4 levels. The first level comprises a model that does not include varying slope across clusters, $\nu_j = 0$ and the cluster-specific effects follows a Normal distribution $c_j \sim N(0, \sigma^2_c)$ (MS1). The second level of the model specification element assumes that $\nu_j = 0$, and $c_j \sim$Student-t distribution with *df* degrees of freedom (MS2). The third level assumes that $\nu_j = 0$, and $c_j \sim Gamma(\alpha, 1)$ (MS3). The fourth level assumes varying cluster-level covariate effects between clusters, $\nu_j \sim N(0, 1)$ and $c_j \sim N(0, \sigma^2_c)$ (MS4).

Of the 36 simulation sets, 32 are constructed as the unique combinations of the 4 elements, for the following levels: randomization type $\in$ {SR1, SR2}, treatment effect heterogeneity $\in$ {HTE0, HTE1}, outcomes $\in$ {CO, BO}, model specification $\in$ {MS1, MS2, MS3, MS4}. The 4 remaining simulation sets utilize simple randomization of the clusters and generate the data from Equation (2.8) similar to SR1 columns in Table 2.1. Denote these 4 simulations as SR0-A, SR0-B, SR0-C, SR0-D.

Each simulation set comprises of either 1458 or 2916 configurations of unique factor combinations. Every configuration within each simulation set is replicated 100 times. Tables 2.1 and 2.2 summarize the factors that varied within each simulation set. Table 2.1 describes the levels that were examined for coefficients β_1, β_2, β_3, and β_9. In addition, in each set we examine the number of clusters, J, and the number of individuals within each cluster m_j. We examine the performance of each method assuming the the size of the clusters are equal, $m_j = m \in \{20, 50, 200\}, \forall j$, as well as with unequal cluster sizes, such that $log(m_j) \sim N(1, sd = 20)\ \forall j \in \{1, \ldots, J\}$ was rounded to the nearest integer. Table 2.2 summarizes the distributions and values for the cluster specific mean, c_j.

2.3.1 Estimation Procedures

We compared 7 specifications of the models described in Sections 2.2.2 and 2.2.3. Two of the model specifications followed the data-generating specification: a GLMM specification, denoted as True model, and a GEE specification, denoted as GEE True.

The remaining 5 models include Model (2.3), which will be denoted as the Standard model, two models based on Equation (2.4), which will be denoted as TRM and the Strata model, one model based on Equation (2.5), denoted as NKM, and a GEE model with the same mean specification as TRM and denoted as GEE TRM. Table 2.3 summarizes the models that were used for sets with each model specification. Sampling variance estimation with GEE model used the Mancl and DeRouen (2001) adjustment. Comparison to the unadjusted estimates and the Kauermann and Carroll (2001) are provided in the Online Supplement. For MS4, the True Model adds a random slope to its estimation procedure.

2.3.2 Model Evaluation

We consider the estimand of interest to be the marginal difference between the expected values under the active intervention and under the control, γ. To estimate $E(Y_{ij}(0))$ and $E(Y_{ij}(1))$ using multi-level models that adjust for additional covariates and cluster specific intercept, requires integration over these covariates and the distribution of the cluster specific effects. When model parameters are unknown, additional integration over them is required. These integrations commonly do not have analytically tractable solutions when

Table 2.1: The pre-defined variable values for simulating data from Equation (2.8) for MS1, MS2, MS3, and MS4.

	CO				BO			
	HTE0		HTE1		HTE0		HTE1	
Variable	SR1	SR2	SR1	SR2	SR1	SR2	SR1	SR2
β_1	1/4, 1, 4	1/4, 1, 4	1/4, 1, 4	1/4, 1, 4	1, 2, 3	0.25, 0.5, 1	1, 2, 3	0.25, 0.5, 1
β_2	1, 2, 3	1, 2, 3	1, 2, 3	1, 2, 3	-1.5, -1, -0.5	-2, -1.5, -1	-1.5, -1, -0.5	-2, -1.5, -1
β_3	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1
β_4	1	1	1	1	-1	-1	-1	-1
β_5	0	1	0	1	0	1	0	-0.5
β_6	1	1	1	1	1	10	1	10
β_7	0	1	0	1	0	1	0	1
β_8	0	0	1	0	0	0	-1	0
β_9	0	0	1,3	1,3	0	0	1, 2	0.5, 1
Δ	0	0	1	1	0	0	1	1
J	24, 36, 48	48, 72, 96	24, 36, 48	48, 72, 96	24, 36, 48	48, 72, 96	24, 36, 48	48, 72, 96
S	2, 3, 4	4, 6, 8	2, 3, 4	4, 6, 8	2, 3, 4	4, 6, 8	2, 3, 4	4, 6, 8
$m_j \ \forall\ j$	20, 50, 200	20, 50, 200	20, 50, 200	20, 50, 200	20, 50, 200	20, 50, 200	20, 50, 200	20, 50, 200

Table 2.2: Distribution of cluster means for each simulation set and model specification.

Model Specification	Variable	CO	BO
MS1, MS4	c_j	$Normal(0, \sigma_c^2)$	$Normal(0, \sigma_c^2)$
	σ_c^2	1/3, 1, 3	1/3, 1/2, 1
MS2	c_j	t_ν	t_ν
	ν	3, 5, 7	1, 2, 3
MS3	c_j	$Gamma(\alpha, 1)$	$Gamma(\alpha, 1)$
	α	0.5, 8, 16	0.5, 2, 4

the link function is not linear. Moreover, calculating interval estimates for the marginal means requires additional derivations. A possible solution is to rely on Bayesian or empirical Bayes solutions (Seltzer et al., 1996; Skrondal and Rabe-Hesketh, 2009). The empirical Bayes solution does not take into account the uncertainty of the second-level variance component (Seltzer et al., 1996), which could lead to underestimation of the standard error. The Bayesian approach assumes a prior distribution for the second-level variance, and can be sensitive to this assumption (Seltzer et al., 1996). To account for the uncertainty of the second level-variance component, we relied on the Bayesian approach, implemented with the brms (Bürkner, 2017) R package. The default half-Student-t prior distribution was assumed for the second-level variance component. Summary of the estimation procedure is provided

Table 2.3: Mean models for each simulation set presented by randomization type and treatment effect heterogeneity. Here, we suppress outcome type and model specifications because mean models do not change for different levels of these two elements.

Model Type	Model Formula
True	
SR1, HTE0	$g(\mu_{ij}(T_j)) = c_j + \beta_1 T_j + \beta_2 Z_{1j} + \beta_3 X_{1ij}$
SR2, HTE0	$g(\mu_{ij}(T_j)) = c_j + \beta_1 T_j + \beta_2 Z_{1j} + \beta_3 X_{1ij} + \beta_4 exp(X_{2ij}) + \beta_5 \overline{X}_{2j}$
SR1, HTE1	$g(\mu_{ij}(T_j)) = c_j + \beta_1 T_j + \beta_2 Z_{1j} + \beta_3 X_{1ij} + \beta_4 X_{2ij} + \beta_5 T_j X_{2ij}$
SR2, HTE1	$g(\mu_{ij}(T_j)) = c_j + \beta_1 T_j + \beta_2 Z_{1j} + \beta_3 X_{1ij} + \beta_4 exp(X_{2ij}) + \beta_5 \overline{X}_{2j} + \beta_6 T_j X_{2ij}$
Standard	
All Sets	$g(\mu_{ij}(T_j)) = c_j + \beta_1 T_j$
TRM	
SR1,{HTE0, HTE1} & SR2, HTE0	$g(\mu_{ij}(T_j)) = c_j + \beta_1 T_j + \beta_2 X_{1ij}$
SR2, HTE1	$g(\mu_{ij}(T_j)) = c_j + \beta_1 T_j + \beta_2 X_{1ij} + \beta_3 X_{2ij}$
Strata	
SR1, {HTE0, HTE1} & SR2, HTE0	$g(\mu_{ij}(T_j)) = c_j + \beta_1 T_j + \beta_2 X_{1ij} + \beta_3 S_j$
SR2, HTE1	$g(\mu_{ij}(T_j)) = c_j + \beta_1 T_j + \beta_2 X_{1ij} + \beta_3 S_j + \beta_4 X_{2ij}$
NKM	
SR1, HTE0	$g(\mu_{ij}(T_j)) = c_j + \beta_1 T_j + \beta_2 Z_{1j} + \beta_3 (X_{1ij} - Z_{1j})$
SR2, {HTE0, HTE1}	$g(\mu_{ij}(T_j)) = c_j + \beta_1 T_j + \beta_2 Z_{1j} + \beta_3 (X_{1ij} - Z_{1j}) + \beta_4 X_{2ij} + \beta_5 (X_{2ij} - \overline{X}_{2j})$
SR1, HTE1	$g(\mu_{ij}(T_j)) = c_j + \beta_1 T_j + \beta_2 Z_{1j} + \beta_3 (X_{1ij} - Z_{1j}) + \beta_4 X_{2ij}$
GEE True	
SR1, HTE0	$g(\mu_{ij}(T_j)) = \beta_1 T_j + \beta_2 Z_{1j} + \beta_3 X_{1ij}$
SR2, HTE0	$g(\mu_{ij}(T_j)) = \beta_1 T_j + \beta_2 Z_{1j} + \beta_3 X_{1ij} + \beta_4 exp(X_{2ij}) + \beta_5 \overline{X}_{2j}$
SR1, HTE1	$g(\mu_{ij}(T_j)) = \beta_1 T_j + \beta_2 Z_{1j} + \beta_3 X_{1ij} + \beta_4 X_{2ij} + \beta_5 T_j X_{2ij}$
SR2, HTE1	$g(\mu_{ij}(T_j)) = \beta_1 T_j + \beta_2 Z_{1j} + \beta_3 X_{1ij} + \beta_4 exp(X_{2ij}) + \beta_5 \overline{X}_{2j} + \beta_6 T_j X_{2ij}$
GEE TRM	
All Sets	$g(\mu_{ij}(T_j)) = \beta_1 T_j + \beta_2 X_{1ij}$

in Appendix 2.7.2. The brmsmargins (Wiley and Hedeker, 2022) R package was used to obtain marginal estimates for $E(Y_{ij}(1))$ and $E(Y_{ij}(0))$.

GEE models produce marginal parameter estimates, which can be used to estimate γ. The R packages geepack (Halekoh et al., 2006) and geesmv (Wang, 2015) are used to obtain point estimates and bias-adjusted standard errors for the parameters of the GEE models, respectively. Using these parameter estimates, the R package marginaleffects (Arel-Bundock, 2022) is then used to obtain the point estimates and their corresponding sampling variance for γ.

At each replication of each configuration we obtained an estimate for γ and its sampling variances. These estimates are used to compute a binary indicator for whether a 95% confidence interval covers the estimand, the interval width, and the bias of the estimate relative to γ. Using these values, for each configuration within each simulation set, we estimated the coverage probability by calculating the proportion of 95% confidence intervals that cover the estimand, the mean interval width and the mean absolute bias (MAB). To identify the factors that have the largest influence on the coverage probabilities, interval width and MAB we use ANOVA (Gutman and Rubin, 2015). Similar methodology was used by Cangul et al. (2009) and Gutman and Rubin (2015).

2.3.3 Separate Treatment Models

The simulations outlined in Section 2.3 are repeated with separate models for the treatment and control groups (Section 2.2.4). These simulations are similar to the configurations that are described in Section 2.3 (Tables 2.1 and 2.2). An additional simulation setting that allows for σ_c^2 and σ_ϵ^2 to vary across intervention groups is used to assess the performance of the joint and the separate models. Specifically, we examined simulation set with {CO, HTE1, SR2, MS1}, and we assumed that c_j in the intervention and control groups follows $N(0, \sigma_c^2(1 - T_j) + 3T_j)$, where $\sigma_c^2 \in \{1/3, 1, 3\}$. We also assumed that $\sigma_\epsilon^2 = 1$ when $T_j = 0$ and $\sigma_\epsilon^2 = 3$ when $T_j = 1$.

2.4 Results

2.4.1 Continuous Outcome Results

Across the different operating characteristics, the number of clusters and the cluster variance were the most frequent influential factors for continuous outcomes simulation sets. Additionally, the treatment interaction coefficient was influential for coverage probability when the treatment effect is heterogeneous (HTE1). We will summarize the results for each model specification for the number of clusters and the cluster variance for HTE0 simulation sets by averaging across the other factors. For HTE1 simulations we will summarize the results for the number of clusters, the cluster variance and the interaction coefficient while averaging across the other factors. Because the results for the different operating characteristics were similar between varying cluster sizes and equal cluster sizes, we present the results for equal cluster sizes and do not provide the results for the varying cluster sizes.

Models with Normally Distributed Cluster Specific Mean (MS1)

For all of the methods, increasing the number of clusters improves model performance with decreased interval widths and smaller mean absolute bias (MAB). Increasing the variance of the cluster specific mean results in worse model performance with increased interval widths and larger MAB. However, MAB is less than 0.5 in the majority of simulations.

Tables 2.4-2.7 summarizes the coverage rates for {{SR1,SR2}, {HTE0,HTE1}, CO, MS1} simulation sets. In these simulation sets, the True, NKM, and GEE True models coverages are generally at or above nominal. The Standard model, TRM, Strata model, and GEE TRM have the largest above nominal coverage rates.

Table 2.4: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	99 (98,100)	97 (96,100)	95 (92,97)	98 (97,100)	97 (95,100)	95 (93,97)	98 (97,100)	97 (95,98)	95 (93,97)
Strata	96 (94,97)	95 (94,97)	94 (91,96)	96 (94,97)	95 (94,97)	94 (91,97)	94 (91,97)	95 (94,97)	95 (94,97)
True	95 (94,97)	95 (94,97)	94 (91,97)	96 (94,98)	96 (94,98)	94 (91,97)	95 (94,97)	94 (91,97)	94 (91,96)
NKM	95 (94,97)	95 (94,97)	94 (91,97)	96 (94,98)	95 (94,97)	94 (91,97)	95 (94,99)	95 (94,97)	94 (92,97)
Standard	99 (100,100)	99 (97,100)	95 (94,97)	99 (99,100)	98 (97,100)	96 (94,98)	99 (100,100)	98 (97,100)	96 (94,97)
GEE True	95 (94,96)	96 (96,97)	95 (94,96)	95 (94,96)	96 (94,96)	95 (93,97)	95 (94,96)	95 (94,96)	95 (94,96)
GEE TRM	98 (96,99)	98 (96,99)	96 (94,98)	98 (97,100)	97 (96,98)	95 (94,97)	99 (98,100)	97 (96,99)	96 (94,97)

Because all of the methods generally have at or above nominal coverage we summarize their mean interval width in Tables 2.8, 2.9, 2.10, and 2.11. For all models, the mean interval widths increase as σ_c^2 increases and decrease as the number of clusters increases. The largest interval widths and their variability across methods are observed for Set {SR2, HTE1, CO, MS1} for the same levels of σ_c^2 and J. The standard model generally has the largest interval

Table 2.5: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	98 (97,100)	97 (96,98)	96 (95,97)	98 (97,99)	97 (96,100)	95 (94,97)	97 (96,99)	97 (96,98)	97 (96,100)
Strata	96 (94,98)	96 (95,97)	95 (94,96)	96 (94,98)	96 (95,97)	94 (92,96)	95 (94,97)	96 (95,98)	97 (94,100)
True	95 (91,97)	95 (94,96)	94 (94,97)	96 (94,97)	94 (94,97)	93 (91,96)	95 (93,96)	95 (93,97)	96 (94,100)
NKM	94 (92,96)	95 (94,97)	95 (92,97)	95 (94,97)	96 (94,97)	95 (93,96)	95 (94,95)	96 (94,97)	96 (94,97)
Standard	99 (99,100)	98 (97,99)	97 (95,97)	99 (99,100)	98 (97,100)	96 (95,97)	99 (97,100)	98 (97,100)	98 (96,100)
GEE True	96 (94,98)	95 (94,96)	95 (94,96)	96 (94,97)	95 (95,97)	95 (94,97)	95 (94,97)	95 (94,97)	95 (94,96)
GEE TRM	99 (98,100)	98 (96,99)	96 (95,97)	99 (98,100)	98 (97,98)	96 (95,98)	99 (98,100)	98 (96,98)	96 (95,98)

Table 2.6: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE1, CO, MS1}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 1$			$\beta_9 = 3$		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	98 (97,100)	97 (95,100)	96 (94,97)	97 (96,100)	97 (94,100)	95 (94,97)
Strata	95 (94,97)	95 (94,97)	95 (94,97)	95 (92,97)	95 (94,97)	95 (94,97)
True	94 (91,97)	94 (94,97)	94 (93,97)	91 (89,94)	94 (91,97)	94 (92,97)
NKM	94 (92,97)	95 (94,97)	94 (94,97)	95 (93,97)	95 (94,97)	95 (94,97)
Standard	99 (98,100)	98 (97,100)	96 (94,99)	98 (97,100)	97 (97,100)	96 (94,97)
GEE True	95 (94,97)	95 (94,96)	95 (94,97)	92 (91,94)	94 (92,96)	94 (93,96)
GEE TRM	98 (97,99)	97 (96,98)	96 (95,97)	97 (95,98)	96 (95,98)	95 (94,97)

widths across all simulation sets. The Strata model, TRM and GEE TRM have the next larger interval widths. In Set {SR2, HTE1, CO, MS1}, their interval widths can be three times larger than the interval widths of the True model. The True model and the NKM have similar interval widths, except for Set {SR2, HTE1, CO, MS1}, where the NKM has interval widths that are 10% to 67% larger than the interval widths of the True model. As σ_c^2 increases the ratio between the mean interval widths of NKM and the True model is closer to 1. Generally, multilevel models have similar or shorter intervals compared to the corresponding GEE models.

For simulation sets {{SR1,SR2}, {HTE0,HTE1}, CO, MS1}, across all models, the MAB increases as the variance of c_j increases and/or the number of clusters decreases. For sets other than {SR2, HTE1, CO, MS1}, all models generally have similar MAB across configurations; however, the Standard model typically has the largest MAB (Tables 2.12, 2.13,

Table 2.7: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE1, CO, MS1}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 1$			$\beta_9 = 3$		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	99 (98,100)	98 (97,98)	97 (96,98)	99 (98,100)	98 (98,99)	97 (96,99)
Strata	98 (97,100)	97 (96,98)	97 (96,98)	99 (98,100)	98 (97,99)	97 (96,99)
True	95 (93,97)	94 (93,96)	94 (91,96)	92 (90,93)	93 (92,94)	95 (93,96)
NKM	95 (94,97)	94 (93,96)	95 (93,97)	94 (93,96)	95 (94,96)	96 (94,97)
Standard	99 (98,100)	98 (97,99)	97 (97,99)	99 (98,100)	99 (98,100)	98 (97,99)
GEE True	95 (94,97)	95 (94,97)	95 (93,96)	92 (90,94)	94 (93,96)	95 (94,96)
GEE TRM	99 (98,100)	98 (98,99)	97 (97,98)	99 (98,100)	99 (98,100)	98 (97,99)

Table 2.8: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	1.45 (1.16,1.74)	2.02 (1.8,2.26)	3.04 (2.92,3.15)	1.15 (0.92,1.39)	1.59 (1.45,1.75)	2.46 (2.4,2.53)	0.99 (0.8,1.2)	1.37 (1.26,1.5)	2.09 (2.03,2.13)
Strata	1.13 (1.08,1.16)	1.78 (1.72,1.84)	2.91 (2.85,2.96)	0.9 (0.86,0.95)	1.41 (1.38,1.42)	2.35 (2.33,2.38)	0.77 (0.74,0.79)	1.22 (1.19,1.24)	2.01 (1.99,2.03)
True	1.05 (1.03,1.07)	1.73 (1.69,1.76)	2.9 (2.85,2.95)	0.84 (0.83,0.85)	1.38 (1.36,1.4)	2.33 (2.31,2.36)	0.72 (0.71,0.73)	1.18 (1.17,1.21)	1.97 (1.93,2.02)
NKM	1.05 (1.03,1.07)	1.74 (1.71,1.76)	2.9 (2.86,2.97)	0.84 (0.83,0.85)	1.38 (1.36,1.39)	2.34 (2.32,2.38)	0.72 (0.72,0.73)	1.18 (1.17,1.19)	1.98 (1.95,2.02)
Standard	1.82 (1.46,2.17)	2.3 (2,2.57)	3.23 (3.04,3.34)	1.46 (1.17,1.74)	1.82 (1.61,2.01)	2.6 (2.48,2.7)	1.25 (1,1.5)	1.56 (1.38,1.74)	2.2 (2.13,2.27)
GEE True	1.05 (1.04,1.07)	1.74 (1.72,1.75)	2.97 (2.94,3)	0.84 (0.83,0.85)	1.4 (1.38,1.41)	2.37 (2.35,2.39)	0.72 (0.72,0.73)	1.19 (1.18,1.19)	2.03 (2.01,2.05)
GEE TRM	1.43 (1.13,1.75)	1.97 (1.77,2.18)	3.08 (2.95,3.2)	1.14 (0.91,1.39)	1.6 (1.44,1.78)	2.48 (2.4,2.61)	0.99 (0.79,1.2)	1.37 (1.23,1.53)	2.13 (2.06,2.23)

and 2.14). In Set {SR2, HTE1, CO, MS1}, the MAB of the Standard model, GEE TRM, TRM, and Strata model are approximately 50% to 100% greater than the True model, and the MAB for NKM is between 25% and 60% larger (Table 2.15). The GEE True MAB is similar across configurations to the True model MAB.

Simple Randomization of Intervention

Simple randomization of clusters to the intervention resulted in nominal coverage or close to nominal for all models in these simulation sets (data not shown). Mean interval widths are largest for the Standard model ranging from 25-100% increase compared to the True model interval widths. TRM and GEE TRM interval widths are also larger across configurations than the True model but ranging from 2% to 50% increase. NKM and GEE True interval widths are similar to the True model. The Strata model has similar interval widths to the True model, except for heterogeneous treatment effect (Set SR0-B) when $c_j = 1/3$ in which interval widths are approximately 33% larger than the True model interval widths.

Table 2.9: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	1.14 (1.09,1.19)	1.46 (1.43,1.47)	2.2 (2.18,2.2)	0.93 (0.88,0.98)	1.2 (1.17,1.2)	1.78 (1.76,1.79)	0.78 (0.76,0.79)	1.03 (1.01,1.04)	1.53 (1.52,1.56)
Strata	1 (0.98,1.01)	1.37 (1.36,1.39)	2.13 (2.1,2.15)	0.81 (0.8,0.82)	1.12 (1.11,1.12)	1.72 (1.7,1.74)	0.7 (0.69,0.71)	0.96 (0.95,0.97)	1.49 (1.48,1.5)
True	0.72 (0.72,0.73)	1.18 (1.18,1.19)	2.01 (2,2.04)	0.58 (0.58,0.59)	0.96 (0.96,0.97)	1.63 (1.61,1.64)	0.5 (0.5,0.5)	0.83 (0.82,0.83)	1.39 (1.38,1.4)
NKM	0.8 (0.79,0.81)	1.24 (1.22,1.26)	2.05 (2.03,2.08)	0.65 (0.65,0.66)	1.01 (1,1.01)	1.66 (1.65,1.67)	0.56 (0.56,0.56)	0.87 (0.86,0.87)	1.43 (1.42,1.43)
Standard	1.34 (1.26,1.43)	1.6 (1.56,1.62)	2.28 (2.23,2.29)	1.09 (1.02,1.18)	1.31 (1.27,1.32)	1.85 (1.82,1.86)	0.91 (0.87,0.92)	1.13 (1.09,1.13)	1.6 (1.58,1.61)
GEE True	0.73 (0.72,0.73)	1.21 (1.21,1.23)	2.07 (2.05,2.1)	0.59 (0.59,0.59)	0.97 (0.97,0.98)	1.66 (1.65,1.67)	0.51 (0.51,0.51)	0.84 (0.83,0.84)	1.43 (1.42,1.44)
GEE TRM	1.23 (1.1,1.37)	1.56 (1.45,1.69)	2.26 (2.19,2.34)	1 (0.89,1.12)	1.27 (1.18,1.37)	1.84 (1.79,1.91)	0.86 (0.77,0.97)	1.09 (1.02,1.18)	1.59 (1.54,1.64)

Table 2.10: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE1, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	1.48 (1.27,1.77)	2.04 (1.89,2.23)	3.06 (2.98,3.13)	1.18 (1.01,1.41)	1.62 (1.5,1.77)	2.45 (2.39,2.5)	1.01 (0.87,1.2)	1.39 (1.29,1.52)	2.11 (2.06,2.16)
Strata	1.2 (1.13,1.24)	1.84 (1.79,1.88)	2.95 (2.91,2.98)	0.96 (0.9,1)	1.46 (1.42,1.49)	2.37 (2.34,2.4)	0.81 (0.76,0.85)	1.24 (1.22,1.27)	2.02 (1.99,2.05)
True	1.06 (1.04,1.07)	1.74 (1.72,1.77)	2.9 (2.86,2.94)	0.84 (0.83,0.85)	1.38 (1.36,1.41)	2.32 (2.29,2.36)	0.72 (0.71,0.73)	1.19 (1.17,1.2)	1.97 (1.95,2.01)
NKM	1.12 (1.07,1.19)	1.79 (1.74,1.84)	2.93 (2.9,2.98)	0.9 (0.85,0.95)	1.43 (1.4,1.44)	2.35 (2.32,2.39)	0.77 (0.73,0.82)	1.22 (1.2,1.24)	2.01 (1.97,2.04)
Standard	2 (1.79,2.28)	2.43 (2.25,2.66)	3.31 (3.21,3.42)	1.6 (1.43,1.82)	1.95 (1.82,2.1)	2.66 (2.57,2.75)	1.37 (1.24,1.57)	1.67 (1.55,1.83)	2.3 (2.22,2.37)
GEE True	1.05 (1.04,1.07)	1.74 (1.73,1.76)	2.97 (2.94,3)	0.84 (0.84,0.85)	1.39 (1.38,1.41)	2.37 (2.35,2.4)	0.72 (0.72,0.73)	1.19 (1.18,1.2)	2.03 (2.01,2.05)
GEE TRM	1.61 (1.46,1.8)	2.14 (2.02,2.28)	3.19 (3.08,3.27)	1.3 (1.18,1.44)	1.71 (1.62,1.82)	2.56 (2.48,2.64)	1.12 (1.02,1.23)	1.48 (1.39,1.56)	2.21 (2.14,2.28)

For homogeneous treatment effect (Set SR0-A), the MAB of all models are similar, except for the Standard model with MAB that is 10% to 50% larger than the True model. In Set SR0-B, the Standard model has the largest MAB, followed by TRM, GEE TRM, and the Strata model, all of these methods have MABs that are 33% to 60% larger than the True model. The GEE True MAB is similar to the True model, and the NKM MAB is larger by approximately 10% compared to the True model.

Models with Cluster Specific Intercepts Following Non-Normal Distributions (MS2 & MS3)

When the cluster specific mean is generated from a Student's t-distribution, above nominal coverages are observed for the Standard model, TRM, and GEE TRM. The NKM and GEE True have similar coverages to the True model across simulation sets (Supplementary Material 2.7.5). The Strata model has similar coverages to the True model for simulation sets with one stratification variable (SR1), but it has above nominal coverage in sets with two stratification variables (SR2). For mean interval widths and MABs, similar trends are observed when cluster specific means are generated from Student's t-distribution and the

Table 2.11: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE1, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	2.08 (1.97,2.19)	2.26 (2.15,2.37)	2.78 (2.69,2.85)	1.69 (1.6,1.78)	1.86 (1.78,1.94)	2.28 (2.22,2.35)	1.45 (1.37,1.54)	1.6 (1.53,1.68)	1.96 (1.91,2.01)
Strata	1.97 (1.86,2.09)	2.16 (2.05,2.26)	2.68 (2.61,2.75)	1.59 (1.5,1.67)	1.76 (1.69,1.85)	2.2 (2.14,2.25)	1.37 (1.29,1.45)	1.52 (1.45,1.59)	1.9 (1.86,1.93)
True	0.72 (0.72,0.73)	1.19 (1.18,1.19)	2.01 (1.99,2.04)	0.58 (0.58,0.59)	0.96 (0.95,0.97)	1.63 (1.61,1.64)	0.5 (0.5,0.5)	0.83 (0.82,0.84)	1.4 (1.39,1.41)
NKM	1.23 (1.2,1.25)	1.55 (1.52,1.58)	2.24 (2.22,2.26)	0.99 (0.97,1.01)	1.27 (1.25,1.28)	1.83 (1.82,1.85)	0.86 (0.84,0.87)	1.09 (1.07,1.1)	1.58 (1.56,1.6)
Standard	2.19 (2.07,2.28)	2.34 (2.24,2.45)	2.83 (2.75,2.91)	1.79 (1.68,1.85)	1.92 (1.85,2)	2.33 (2.27,2.4)	1.52 (1.44,1.6)	1.66 (1.6,1.74)	2.01 (1.96,2.06)
GEE True	0.73 (0.73,0.74)	1.21 (1.2,1.22)	2.07 (2.05,2.08)	0.59 (0.59,0.59)	0.98 (0.97,0.98)	1.66 (1.65,1.67)	0.51 (0.5,0.51)	0.84 (0.83,0.84)	1.43 (1.42,1.43)
GEE TRM	2.16 (2.04,2.25)	2.36 (2.24,2.46)	2.87 (2.79,2.96)	1.75 (1.64,1.83)	1.92 (1.82,1.99)	2.34 (2.27,2.41)	1.51 (1.41,1.59)	1.65 (1.57,1.72)	2.02 (1.96,2.08)

Table 2.12: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	0.22 (0.2,0.24)	0.35 (0.32,0.38)	0.58 (0.56,0.61)	0.17 (0.16,0.19)	0.28 (0.26,0.29)	0.48 (0.44,0.51)	0.15 (0.14,0.17)	0.25 (0.23,0.27)	0.4 (0.38,0.42)
Strata	0.22 (0.2,0.23)	0.35 (0.32,0.38)	0.58 (0.53,0.62)	0.17 (0.15,0.19)	0.27 (0.26,0.29)	0.48 (0.45,0.51)	0.15 (0.14,0.17)	0.25 (0.23,0.26)	0.4 (0.38,0.43)
True	0.2 (0.19,0.22)	0.34 (0.31,0.37)	0.58 (0.53,0.61)	0.16 (0.15,0.17)	0.26 (0.24,0.28)	0.48 (0.45,0.51)	0.14 (0.13,0.15)	0.24 (0.23,0.26)	0.4 (0.37,0.43)
NKM	0.2 (0.19,0.22)	0.34 (0.31,0.37)	0.59 (0.54,0.62)	0.16 (0.15,0.17)	0.26 (0.24,0.28)	0.47 (0.44,0.52)	0.14 (0.13,0.15)	0.24 (0.23,0.26)	0.4 (0.37,0.43)
Standard	0.24 (0.22,0.27)	0.36 (0.32,0.4)	0.59 (0.53,0.65)	0.19 (0.17,0.2)	0.29 (0.27,0.31)	0.48 (0.45,0.52)	0.17 (0.14,0.19)	0.26 (0.24,0.28)	0.41 (0.37,0.43)
GEE True	0.2 (0.19,0.21)	0.33 (0.31,0.34)	0.58 (0.55,0.6)	0.16 (0.15,0.17)	0.27 (0.25,0.28)	0.48 (0.45,0.5)	0.14 (0.13,0.15)	0.23 (0.21,0.25)	0.4 (0.38,0.43)
GEE TRM	0.22 (0.2,0.24)	0.33 (0.32,0.35)	0.58 (0.55,0.61)	0.18 (0.16,0.19)	0.28 (0.27,0.28)	0.48 (0.46,0.51)	0.15 (0.14,0.16)	0.24 (0.22,0.26)	0.41 (0.39,0.43)

Normal distribution. Specifically, NKM has interval widths that are closest to the True and GEE True models. In addition, GEE models have slightly higher interval widths than the corresponding GLMMs.

When c_j follows a Gamma distribution (MS3), the trends are generally similar to the c_j generated from Normal distribution. However, the differences in model performance are more pronounced when the shape parameter, α, is large. When $\alpha \in \{8, 16\}$, GEE models have interval widths larger than their corresponding GLMM models. The GEE True model can even have interval widths that are larger than the NKM model. When $\alpha = 16$, GEE True model has coverages that are at or above nominal, while the True GLMM model is below nominal. This is a result of the larger interval widths for GEE True. The NKM model is more robust to skewed cluster specific mean, and results in better coverage than the True model in this simulation set.

Models with Cluster Specific Slope (MS4)

When the slope of Z_{1j}, β_3 in Equation (2.8), varies across clusters, coverage probabilities are similar to the ones observed for first level misspecifications (Section 2.4.1). In sets {SR2,

Table 2.13: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	0.18 (0.17,0.19)	0.26 (0.25,0.28)	0.43 (0.39,0.46)	0.15 (0.14,0.17)	0.21 (0.2,0.22)	0.35 (0.32,0.38)	0.13 (0.12,0.14)	0.19 (0.18,0.2)	0.28 (0.26,0.32)
Strata	0.18 (0.17,0.19)	0.26 (0.25,0.28)	0.42 (0.4,0.45)	0.15 (0.14,0.17)	0.21 (0.2,0.22)	0.35 (0.33,0.37)	0.13 (0.12,0.14)	0.19 (0.18,0.2)	0.28 (0.25,0.32)
True	0.14 (0.13,0.15)	0.23 (0.22,0.24)	0.41 (0.38,0.43)	0.12 (0.11,0.13)	0.19 (0.18,0.2)	0.34 (0.32,0.36)	0.1 (0.1,0.11)	0.17 (0.16,0.18)	0.27 (0.24,0.31)
NKM	0.16 (0.15,0.17)	0.24 (0.23,0.25)	0.42 (0.39,0.44)	0.13 (0.13,0.14)	0.2 (0.18,0.2)	0.34 (0.32,0.36)	0.12 (0.11,0.12)	0.17 (0.16,0.19)	0.27 (0.24,0.31)
Standard	0.19 (0.18,0.2)	0.27 (0.25,0.28)	0.43 (0.39,0.45)	0.16 (0.15,0.17)	0.21 (0.2,0.23)	0.35 (0.32,0.38)	0.13 (0.13,0.14)	0.19 (0.18,0.2)	0.28 (0.26,0.32)
GEE True	0.14 (0.14,0.15)	0.24 (0.23,0.25)	0.4 (0.38,0.42)	0.12 (0.11,0.12)	0.19 (0.19,0.2)	0.34 (0.32,0.36)	0.1 (0.09,0.11)	0.17 (0.16,0.18)	0.29 (0.28,0.31)
GEE TRM	0.18 (0.17,0.19)	0.27 (0.26,0.28)	0.42 (0.4,0.43)	0.15 (0.14,0.16)	0.22 (0.21,0.23)	0.35 (0.33,0.38)	0.13 (0.12,0.14)	0.19 (0.18,0.2)	0.3 (0.29,0.32)

Table 2.14: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE1, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	0.23 (0.21,0.25)	0.35 (0.31,0.38)	0.57 (0.52,0.6)	0.19 (0.17,0.2)	0.3 (0.28,0.32)	0.48 (0.45,0.51)	0.17 (0.15,0.18)	0.25 (0.23,0.27)	0.4 (0.37,0.42)
Strata	0.23 (0.21,0.25)	0.35 (0.32,0.38)	0.57 (0.53,0.6)	0.18 (0.17,0.2)	0.29 (0.27,0.32)	0.48 (0.44,0.51)	0.16 (0.14,0.18)	0.25 (0.23,0.27)	0.4 (0.37,0.43)
True	0.22 (0.2,0.23)	0.34 (0.31,0.37)	0.55 (0.51,0.6)	0.18 (0.15,0.19)	0.29 (0.27,0.3)	0.47 (0.44,0.51)	0.16 (0.14,0.17)	0.25 (0.23,0.26)	0.4 (0.37,0.43)
NKM	0.22 (0.2,0.23)	0.35 (0.31,0.37)	0.56 (0.51,0.6)	0.18 (0.16,0.19)	0.29 (0.27,0.31)	0.48 (0.45,0.5)	0.16 (0.14,0.17)	0.25 (0.22,0.27)	0.4 (0.37,0.42)
Standard	0.29 (0.25,0.32)	0.39 (0.36,0.41)	0.59 (0.54,0.63)	0.23 (0.2,0.26)	0.33 (0.31,0.36)	0.51 (0.47,0.54)	0.21 (0.18,0.23)	0.28 (0.25,0.31)	0.42 (0.39,0.45)
GEE True	0.22 (0.21,0.23)	0.35 (0.33,0.37)	0.57 (0.54,0.59)	0.18 (0.16,0.19)	0.28 (0.27,0.3)	0.48 (0.46,0.5)	0.15 (0.14,0.16)	0.24 (0.23,0.25)	0.41 (0.38,0.43)
GEE TRM	0.27 (0.24,0.3)	0.38 (0.36,0.41)	0.59 (0.56,0.61)	0.23 (0.2,0.25)	0.31 (0.29,0.33)	0.5 (0.48,0.52)	0.19 (0.16,0.21)	0.27 (0.25,0.29)	0.43 (0.4,0.45)

{HTE0,HTE1}, CO, MS4} with non-linear effects in the data generating mechanism, the Standard model, TRM, and GEE TRM have above nominal coverage, and the corresponding interval widths range from 25-125% larger than the True model (Supplementary Material 2.7.7). The Strata model has at or above nominal coverages and generally shorter interval widths than the Standard model and TRM. However, it generally has similar or wider interval widths than NKM and the True model. This is more pronounced when two factors are used for stratification (SR2). The NKM, True model, and GEE True model mean interval widths are similar to the results described in Section 2.4.1 for sets {{SR1,SR2}, {HTE0,HTE1}, CO, MS4} (Supplementary Material 2.7.7).

Across sets {{SR1,SR2}, {HTE0,HTE1}, CO, MS4}, the MAB results are similar to those outlined in Section 2.4.1. Changing the value of ν_j has relatively small effects on the coverages, interval widths and MAB (data not shown).

Table 2.15: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE1, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	0.33 (0.31,0.35)	0.39 (0.37,0.41)	0.51 (0.48,0.54)	0.27 (0.26,0.29)	0.3 (0.28,0.32)	0.38 (0.36,0.4)	0.23 (0.21,0.24)	0.26 (0.25,0.28)	0.35 (0.32,0.37)
Strata	0.33 (0.31,0.35)	0.39 (0.37,0.41)	0.51 (0.48,0.53)	0.27 (0.25,0.29)	0.3 (0.29,0.32)	0.38 (0.36,0.4)	0.22 (0.21,0.23)	0.26 (0.25,0.28)	0.35 (0.33,0.37)
True	0.16 (0.14,0.17)	0.25 (0.23,0.25)	0.43 (0.4,0.45)	0.13 (0.12,0.14)	0.2 (0.19,0.21)	0.33 (0.31,0.34)	0.11 (0.1,0.12)	0.17 (0.16,0.18)	0.29 (0.27,0.31)
NKM	0.25 (0.24,0.27)	0.31 (0.3,0.32)	0.46 (0.44,0.48)	0.2 (0.19,0.22)	0.26 (0.25,0.27)	0.35 (0.33,0.37)	0.18 (0.17,0.19)	0.22 (0.21,0.23)	0.32 (0.3,0.34)
Standard	0.33 (0.31,0.35)	0.39 (0.37,0.4)	0.51 (0.48,0.54)	0.27 (0.26,0.29)	0.3 (0.29,0.32)	0.38 (0.36,0.4)	0.23 (0.21,0.24)	0.26 (0.25,0.28)	0.36 (0.33,0.37)
GEE True	0.16 (0.14,0.17)	0.24 (0.23,0.25)	0.42 (0.4,0.44)	0.13 (0.12,0.13)	0.2 (0.19,0.21)	0.33 (0.32,0.35)	0.11 (0.1,0.11)	0.17 (0.17,0.18)	0.29 (0.28,0.31)
GEE TRM	0.33 (0.32,0.35)	0.38 (0.36,0.39)	0.51 (0.48,0.53)	0.27 (0.25,0.29)	0.31 (0.29,0.33)	0.4 (0.38,0.42)	0.23 (0.21,0.24)	0.27 (0.26,0.28)	0.35 (0.33,0.37)

2.4.2 Binary Outcome Results

Models with Normally Distributed Cluster Specific Mean (MS1)

Supplementary Material 2.7.4 summarizes the results for binary outcome simulation sets (BO). The 95% coverage rates are similar to the corresponding sets with continuous outcomes (Table 2.16).

The mean interval width trends are similar to the ones observed for continuous outcomes in Section 2.4.1 for continuous outcomes, although less pronounced because of the smaller interval widths with binary outcomes (Table 2.17). All models have similar MAB across simulation sets {{SR1,SR2}, {HTE0,HTE1}, BO, MS1}, where the bias is smaller when the number of clusters increases and/or cluster variability decreases (Supplementary Material 2.7.4).

The results with simple randomization in sets SR0-C and SR0-D are similar to the ones observed in Section 2.4.1 (data not shown).

Models with Cluster Specific Intercepts Follow Non-Normal Distributions (MS2 & MS3)

When c_j is generated from Student's-t distribution, coverage rates and mean interval width trends are similar to continuous outcomes in Section 2.4.1 (Supplementary Material 2.7.6). Across Sets {{SR1,SR2}, {HTE0,HTE1}, BO, MS2}, GEE interval widths are more similar compared to the interval widths in Section 2.4.1, although both are generally equal or larger than those for NKM and the True model. Across sets {{SR1,SR2}, {HTE0,HTE1}, BO, MS2}, all MAB are similar for all models.

When c_j follows a Gamma distribution, trends in coverage, interval width, and MAB are similar to Sections 2.4.1 and 2.4.1 and to the results when c_j are generated from the Student's t-distribution. Below nominal coverages are observed for GEE TRM and GEE True models in Sets {SR1, HTE1, BO, MS3} when $\alpha = 4$ and number of clusters are 24 and 48. For this set, the corresponding GEE interval widths are smaller than all other models while the bias is similar.

Models with Cluster Specific Slopes (MS4)

GLMMs have performance trends that are similar to the ones observed for continuous outcomes in Section 2.4.1. The percent increases in interval widths for the Standard, TRM, and GEE TRM compared to the True model range between 5% and 30% (Supplementary Material 2.7.8).

Table 2.16: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE0, BO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	97 (96,100)	96 (94,97)	96 (94,97)	95 (94,97)	95 (91,98)	96 (94,98)	97 (94,100)	96 (94,99)	97 (97,97)
Strata	96 (94,97)	95 (93,97)	96 (93,98)	94 (92,95)	95 (94,97)	96 (94,98)	96 (94,97)	94 (91,97)	96 (95,97)
True	96 (94,97)	95 (91,97)	96 (94,97)	94 (93,95)	95 (94,97)	96 (94,97)	96 (94,97)	95 (91,97)	95 (94,97)
NKM	96 (95,97)	94 (91,97)	95 (93,97)	93 (91,95)	95 (91,97)	96 (94,97)	96 (94,97)	95 (94,97)	95 (94,97)
Standard	99 (97,100)	98 (97,100)	97 (96,100)	98 (97,100)	98 (97,100)	98 (97,100)	99 (97,100)	98 (97,100)	98 (97,100)
GEE True	94 (93,96)	95 (92,96)	95 (93,96)	95 (94,96)	95 (93,96)	95 (94,96)	95 (94,96)	95 (94,97)	95 (93,97)
GEE TRM	95 (94,96)	95 (93,98)	95 (93,96)	95 (94,97)	95 (94,96)	95 (94,96)	96 (94,97)	96 (94,97)	95 (93,97)

Table 2.17: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE0, BO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.25 (0.23,0.26)	0.27 (0.25,0.29)	0.33 (0.31,0.34)	0.2 (0.18,0.21)	0.22 (0.2,0.23)	0.27 (0.25,0.28)	0.17 (0.16,0.18)	0.19 (0.17,0.2)	0.23 (0.22,0.24)
Strata	0.23 (0.22,0.25)	0.26 (0.24,0.27)	0.32 (0.31,0.33)	0.19 (0.17,0.2)	0.21 (0.2,0.22)	0.26 (0.24,0.27)	0.16 (0.15,0.17)	0.18 (0.17,0.19)	0.22 (0.21,0.23)
True	0.23 (0.22,0.25)	0.26 (0.24,0.27)	0.32 (0.3,0.33)	0.19 (0.17,0.2)	0.21 (0.19,0.22)	0.26 (0.24,0.27)	0.16 (0.15,0.17)	0.18 (0.16,0.18)	0.22 (0.21,0.23)
NKM	0.23 (0.22,0.25)	0.26 (0.24,0.27)	0.32 (0.3,0.33)	0.19 (0.17,0.2)	0.21 (0.19,0.22)	0.26 (0.24,0.27)	0.16 (0.15,0.17)	0.18 (0.16,0.18)	0.22 (0.21,0.23)
Standard	0.29 (0.27,0.32)	0.31 (0.29,0.34)	0.36 (0.34,0.38)	0.24 (0.22,0.26)	0.25 (0.23,0.27)	0.29 (0.28,0.3)	0.21 (0.19,0.22)	0.22 (0.2,0.23)	0.25 (0.24,0.26)
GEE True	0.24 (0.22,0.25)	0.27 (0.25,0.28)	0.33 (0.31,0.34)	0.19 (0.18,0.2)	0.21 (0.2,0.22)	0.26 (0.24,0.28)	0.17 (0.15,0.17)	0.18 (0.17,0.19)	0.23 (0.21,0.23)
GEE TRM	0.25 (0.23,0.26)	0.27 (0.26,0.29)	0.33 (0.31,0.35)	0.2 (0.18,0.21)	0.22 (0.2,0.24)	0.27 (0.25,0.28)	0.17 (0.16,0.18)	0.19 (0.17,0.2)	0.23 (0.21,0.24)

2.4.3 Separate Treatment Model Results

For both continuous and binary outcomes, simulations using separate models for treatment and control observations show similar trends (data not shown). For GLMMs and GEEs, coverage probability and mean interval widths are larger compared to their corresponding single parallel models. Specifically, coverage probabilities are generally above nominal and mean interval widths are wider by 10% to 25%. However, when σ_c^2 and σ_ϵ^2 differ across intervention groups, coverage probabilities are similar and separate model mean interval widths are wider by less than 10%. MAB for separate models and corresponding parallel model are similar and close to zero.

2.5 Real Data Application

We examined the performance of the different models on data from the PRagmatic trial of Video Education in Nursing Homes (PROVEN) trial (Mor et al., 2017). PROVEN examined the effects of showing an advanced directive video on the propensity of individuals to transfer from a nursing home to a hospital. The trial was implemented in two for-profit nursing home healthcare systems. A total of 360 nursing homes were stratified on two variables: the healthcare system and the hospitalization rate in the year prior to the intervention among patients with advanced illnesses. Within each stratum nursing homes were randomized in a ratio of 1:2 between the intervention and control conditions. The estimand of interest is defined as the marginal difference in the probabilities to be transferred to the hospital in the active intervention group and the control group.

Table 2.18 presents the models used for the analysis, where π_{ij} is the probability of transfer to the hospital for individual i in nursing home j, c_j is nursing home j specific intercept, T_j is the intervention indicator for facility j, $\overline{X}_{1j}$ is the proportion of individuals in nursing home j who identify as Black, X_{1ij} is an indicator representing whether individual i in facility j identifies as Black, Z_{1j} is the hospitalization rate in the year prior to intervention initiation in facility j, and Z_{2j} is facility j's indicator for the healthcare system. A stratification-level indicator, S_j, was defined using stratification variables Z_{1j} and Z_{2j}.

Table 2.18: Models used for the real data application and results for each models estimated average treatment effect and standard error. The GEE model correlation matrix was set as exchangeable, with standard errors computed using the sandwich estimator to obtain robust estimates.

Model Type	Parallel Model Formula	ATE (SE) Parallel	ATE (SE) Separate
Complete	$logit(\pi_{ij}) = c_j + \beta_1 T_j + \beta_2 \overline{X}_{1j} + \beta_3 X_{1ij} + \beta_4 Z_{1j} + \beta_5 Z_{2j}$	-0.0080 (0.0142)	-0.0061 (0.0192)
Standard	$logit(\pi_{ij}) = c_j + \beta_1 T_j$	-0.0073 (0.0155)	-0.0056 (0.0217)
TRM	$logit(\pi_{ij}) = c_j + \beta_1 T_j + \beta_2 X_{1ij}$	-0.0088 (0.0154)	-0.0085 (0.0211)
Strata	$logit(\pi_{ij}) = c_j + \beta_1 T_j + \beta_2 X_{1ij} + \beta_3 S_j$	-0.0080 (0.0143)	-0.0081 (0.0196)
NKM	$logit(\pi_{ij}) = c_j + \beta_1 T_j + \beta_2 \overline{X}_{1j} + \beta_3 (X_{1ij} - \overline{X}_{1j}) + \beta_4 S_j$	-0.0082 (0.0143)	-0.0078 (0.0194)
GEE Complete	$logit(\pi_{ij}) = c_j + \beta_1 T_j + \beta_2 \overline{X}_{1j} + \beta_3 X_{1ij} + \beta_4 Z_{1j} + \beta_5 Z_{2j}$	-0.0074 (0.0144)	-0.0064 (0.0143)
GEE TRM	$logit(\pi_{ij}) = c_j + \beta_1 T_j + \beta_2 X_{1ij}$	-0.0087 (0.0151)	-0.0084 (0.0152)

Table 2.18 also provides estimates of the ATEs and corresponding standard errors for each of the models. All of the models produce similar point estimates and none of these estimates are statistically significant at the 5% nominal level. The separate TRM, Strata, NKM, and GEE TRM models results in similar point estimates compared to the corresponding parallel model. The separate Complete, Standard, and GEE Complete models have between 15% to 25% smaller point estimates compared to their corresponding parallel single models.

The interval widths show similar trends to the simulation study with parallel models. The Standard model and TRM have the largest interval widths and the Complete produces the smallest interval width. The Strata, NKM and the Complete Model have similar point and interval estimates. For separate models, standard errors increase by 35% to 40% for GLMMs compared to the corresponding parallel single models. GEE models do not have substantial differences in standard error between the separate models and the corresponding parallel models.

In the simulations, the large interval widths from the Standard model and TRM typically result in over-coverage of the true treatment effect. In addition, the GEE Complete with bias-adjustment results in increased standard error compared to the GLMM with identical mean specification.

2.6 Discussion

In many individual-level and cluster-level randomized trials investigators do not adjust for predictors beyond the intervention (Yu et al., 2010; Assmann et al., 2000; Austin et al., 2010; Hernandez et al., 2005). We explicitly define the marginal average treatment effect as the estimand of interest, and provide statistical software to estimate it in continuous and binary outcomes. Using simulations, we compare the performance of different adjustment models for the stratification variables when estimating this estimand in SCRTs. Generally, the multilevel model that does not adjust for any additional covariates has above nominal coverage, with the widest interval estimates compared to models that adjust for covariates. This implies that models that do not adjust for covariates are generally less efficient than models that do, resulting in lower statistical power. Similar results were observed for individual level randomization (Ivers et al., 2012).

The performance of the different methods for mean model misspecification and varying slope misspecification are similar for both continuous and binary outcomes. Multilevel and GEE models that adjust for strata-defining covariates result in coverages that are at or above nominal and are more precise than models that do not adjust for strata-defining covariates.

GLMMs that adjust for strata have similar operating characteristics to correctly specified model when the relationships between the covariates and outcomes are linear and the effect of the intervention is additive. When the data generating model include non-linear relationships and interactions, GLMM models that only adjust for the strata has nominal coverage, but with decreased precision compared to NKM and correctly specified models. When the first-level mean is misspecified or the cluster-specific mean follows a non-Normal distribution, an NKM GLMM that adjusts for strata-defining covariates as well as the between and within cluster association is a valid statistical procedure with more precise and accurate estimates, compared to models that do not make these adjustments. These results are observed for both continuous and binary outcomes. Although correctly specified models have the best operating characteristics, they are rarely known in practice. The NKM GLMM operating characteristics are similar to the correctly specified model in many configurations, and provide an efficient alternative when the exact specification of the model is unknown.

For continuous outcomes, a sampling variance bias-adjusted (Mancl and DeRouen, 2001) GEE model with correctly specified mean function results in nominal coverage and similar standard errors to the GLMM with correctly specified mean function when the number of clusters is 24. However, as the number of clusters increases, the standard error of this bias-adjusted GEE is generally larger (ranging from 0.5-13%) than the corresponding GLMM standard errors. Similar results are observed for misspecified GEE models with bias-adjustment when compared to the GLMM with identical mean specification. For binary outcomes, similar trends are observed but are less pronounced (typical increases in interval width ranging from 0.5-5%).

Bias-adjustment is recommended for GEEs when there are few clusters because of underestimated standard errors of the sandwich estimator (Mancl and DeRouen, 2001; Li and Redden, 2014; Huang et al., 2016). However, the improvement in coverage probability may not be substantial compared to GEEs without bias-adjustment when the number of clusters is 36 or larger for continuous outcomes, or more than 48 clusters for binary outcomes (Supplementary Materials 2.7.9). The increase in coverage probability from bias-adjustment stems from increased standard errors.

For both continuous and binary outcomes, modelling the treatment and control observations separately result in higher coverage rates and wider interval widths for both GLMMs and GEEs. However, when the intervention and control groups are generated with differing variances, the performance of separate model and their corresponding parallel models is similar, although still less precise for separate models. These results are limited to our set of simulations, and under more significant differences in response surfaces between the two intervention groups, separate models may result in smaller biases and coverages that are closer nominal. However, these cases may be less plausible in many real applications.

When analyzing the PROVEN data, all methods resulted in interval estimates that have substantial overlap. However, there is a gain in precision when adjusting for stratification variables.

This study is limited by simulation-based data generation, and the selection of the marginal treatment effect as the estimand of interest. The different models may have different operating characteristics in other plausible scenarios. However, because the extensive simulations show similar trends across all configurations, it reinforces the possible gains of adjusting for covariates and the stratification variables, in stratified cluster randomized trials. Moreover, because we do not define the estimand as a parameter in a model, we expect other estimands based on the potential outcomes to have similar performance trends.

In conclusion, when the true response surfaces are unknown, a NKM model that adjusts for within and between association of the covariates and stratification variables is a statistically valid procedure to analyze SCRTs that results in accurate and precise estimates. GEE models with bias-adjusted sampling variance are statistically valid, but have larger interval width compared to their corresponding GLMMs. Lastly, in applications in which the intervention is expected to have a heterogeneous effects, using separate models or a joint model that adjusts for interactions between the covariates and treatment indicator would result in valid statistical inferences.

Chapter 2 References

V. Arel-Bundock. *marginaleffects: Marginal Effects, Marginal Means, Predictions, and Contrasts*, 2022. URL `https://vincentarelbundock.github.io/marginaleffects/`. R package version 0.7.1.9003.

S. F. Assmann, S. J. Pocock, L. E. Enos, and L. E. Kasten. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet*, 355(9209):1064?1069, 2000. doi: 10.1016/s0140-6736(00)02039-0.

P. C. Austin, A. Manca, M. Zwarenstein, D. N. Juurlink, and M. B. Stanbrook. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: A review of trials published in leading medical journals. *Journal of Clinical Epidemiology*, 63(2):142?153, 2010. doi: 10.1016/j.jclinepi.2009.06.002.

L. B. Balzer, M. J. van Laan, M. L. Petersen, and S. Collaboration. Adaptive pre-specification in randomized trials with and without pair-matching. *Statistics in Medicine*, 35(25): 4528–4545, 2016. doi: 10.1002/sim.7023.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

M. L. Beach and P. Meier. Choosing covariates in the analysis of clinical trials. *Controlled Clinical Trials*, 10(4):161?175, 1989. doi: 10.1016/0197-2456(89)90055-x.

S. Borhan, J. Ma, A. Papaioannou, J. Adachi, and L. Thabane. Performance of methods for analyzing continuous data from stratified cluster randomized trials – a simulation study. *Contemporary Clinical Trials Communications*, 33:101115, 2023. doi: 10.1016/j.conctc.2023.101115.

P.-C. Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01.

M. J. Campbell, A. Donner, and N. Klar. Developments in cluster randomized trials and-statistics in medicine. *Statistics in Medicine*, 26(1):2–19, 2006. doi: 10.1002/sim.2731.

M. Z. Cangul, Y. R. Chretien, R. Gutman, and D. B. Rubin. Testing treatment effects in unconfounded studies under model misspecification: Logistic regression, discretization, and their combination. *Statistics in Medicine*, 28(20):2531?2551, 2009. doi: 10.1002/sim.3633.

C. M. Crespi. Improved designs for cluster randomized trials. *Annual Review of Public Health*, 37(1):1–16, 2016. doi: 10.1146/annurev-publhealth-032315-021702.

T. A. Dobbins and J. M. Simpson. Comparison of tests for categorical data from a stratified cluster randomized trial. *Statistics in Medicine*, 21(24):3835–3846, 2002. doi: 10.1002/sim.1256.

A. Donner. Some aspects of the design and analysis of cluster randomization trials. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 47(1):95–113, 1998. doi: 10.1111/1467-9876.00100.

A. Donner and N. Klar. *Design and analysis of Cluster Randomization Trials in health research*. Arnold, 2000.

A. Donner and N. Klar. Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health*, 94(3):416?422, 2004. doi: 10.2105/ajph.94.3.416.

M. J. Emond, J. Ritz, and D. Oakes. Bias in gee estimates from misspecified models for longitudinal data. *Communications in Statistics - Theory and Methods*, 26(1):15?32, 1997. doi: 10.1080/03610929708831899.

W. P. Ford and P. M. Westgate. A comparison of bias-corrected empirical covariance estimators with generalized estimating equations in small-sample longitudinal study settings. *Statistics in Medicine*, 37(28):4318–4329, 2018. doi: 10.1002/sim.7917.

M. Gosho, H. Noma, and K. Maruo. Practical review and comparison of modified covariance estimators for linear mixed models in small-sample longitudinal studies with missing data. *International Statistical Review*, 89(3):550–572, 2021. doi: 10.1111/insr.12447.

R. Gutman and D. B. Rubin. Estimation of causal effects of binary treatments in unconfounded studies. *Statistics in Medicine*, 34(26):3381?3398, 2015. doi: 10.1002/sim.6532.

U. Halekoh, S. Højsgaard, and J. Yan. The r package geepack for generalized estimating equations. *Journal of Statistical Software*, 15/2:1–11, 2006.

R. J. Hayes and L. H. Moulton. *Cluster Randomised Trials*. CRC press, 2017.

A. V. Hernandez, E. W. Steyerberg, G. S. Taylor, A. Marmarou, J. D. Habbema, and A. I. Maas. Subgroup analysis and covariate adjustment in randomized clinical trials of traumatic brain injury: A systematic review. *Neurosurgery*, 57(6):1244?1253, 2005. doi: 10.1227/01.neu.0000186039.57548.96.

S. Huang, M. H. Fiero, and M. L. Bell. Generalized estimating equations in cluster randomized trials with a small number of clusters: Review of practice and simulation study. *Clinical Trials*, 13(4):445–449, 2016. doi: 10.1177/1740774516643498.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and Biomedical Sciences: An introduction*. Cambridge University Press, 2015.

N. M. Ivers, I. J. Halperin, J. Barnsley, J. M. Grimshaw, B. R. Shah, K. Tu, R. Upshur, and M. Zwarenstein. Allocation techniques for balance at baseline in cluster randomized trials: A methodological review. *Trials*, 13(1), 2012. doi: 10.1186/1745-6215-13-120.

B. C. Kahan and T. P. Morris. Improper analysis of trials randomised using stratified blocks or minimisation. *Statistics in Medicine*, 31(4):328?340, 2011. doi: 10.1002/sim.4431.

B. C. Kahan, V. Jairath, C. J. DorÈ, and T. P. Morris. The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies. *Trials*, 15 (1), 2014. doi: 10.1186/1745-6215-15-139.

B. C. Kahan, G. Forbes, Y. Ali, V. Jairath, S. Bremner, M. O. Harhay, R. Hooper, N. Wright, S. M. Eldridge, and C. Leyrat. 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), 2016. doi: 10.1186/s13063-016-1571-2.

G. Kauermann and R. J. Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396, 2001. doi: 10.1198/016214501753382309.

W. Kernan, C. Viscoli, R. Makuch, L. Brass, and R. Horwitz. Stratified randomization for clinical trials. *Journal of Clinical Epidemiology*, 52(1):19?26, 1999. doi: 10.1016/s0895-4356(98)00138-3.

H. C. Kraemer. A source of false findings in published research studies. *JAMA Psychiatry*, 72(10):961, 2015. doi: 10.1001/jamapsychiatry.2015.1178.

J. M. Lachin. Statistical properties of randomization in clinical trials. *Controlled Clinical Trials*, 9(4):289–311, 1988. doi: 10.1016/0197-2456(88)90045-1.

J. D. Lewsey. Comparing completely and stratified randomized designs in cluster randomized trials when the stratifying factor is cluster size: A simulation study. *Statistics in Medicine*, 23(6):897–905, 2004. doi: 10.1002/sim.1665.

C. Leyrat, K. E. Morgan, B. Leurent, and B. C. Kahan. Cluster randomized trials with a small number of clusters: Which analyses should be used? *International Journal of Epidemiology*, 47(3):1012–1012, 2018. doi: 10.1093/ije/dyy057.

P. Li and D. T. Redden. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in Medicine*, 34(2):281–296, 2014. doi: 10.1002/sim.6344.

K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized estimating equations. *Biometrika*, 73(1):13–22, 1986.

B. Lu, J. S. Preisser, B. F. Qaqish, C. Suchindran, S. I. Bangdiwala, and M. Wolfson. A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics*, 63(3):935–941, 2007. doi: 10.1111/j.1541-0420.2007.00764.x.

L. A. Mancl and T. A. DeRouen. A covariance estimator for gee with improved small-sample properties. *Biometrics*, 57(1):126–134, 2001. doi: 10.1111/j.0006-341x.2001.00126.x.

C. E. McCulloch. *Generalized linear mixed models*. Institute of Mathematical Statistics, 2003.

P. Meier. Stratification in the design of a clinical trial. *Controlled Clinical Trials*, 1(4): 355–361, 1981. doi: 10.1016/0197-2456(81)90040-4.

V. Mor, A. E. Volandes, R. Gutman, C. Gatsonis, and S. L. Mitchell. Pragmatic trial of video education in nursing homes: The design and rationale for a pragmatic cluster randomized trial in the nursing home setting. *Clinical Trials*, 14(2):140?151, 2017. doi: 10.1177/1740774516685298.

K. L. Morgan and D. B. Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2), 2012. doi: 10.1214/12-aos1008.

J. M. Neuhaus and J. D. Kalbfleisch. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2):638, 1998. doi: 10.2307/3109770.

J. Pinheiro, D. Bates, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2023. URL `https://CRAN.R-project.org/package=nlme`. R package version 3.1-163.

G. M. Raab and I. Butcher. Balance in cluster randomized trials. *Statistics in Medicine*, 20 (3):351?365, 2001. doi: 10.1002/1097-0258(20010215)20:3⟨351::aid-sim797⟩3.0.co;2-c.

W. F. Rosenberger and J. M. Lachin. *Randomization in clinical trials: Theory and practice.* Wiley, 2015.

D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1), 1978. doi: 10.1214/aos/1176344064.

D. B. Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3), 2008. doi: 10.1214/08-aoas187.

M. H. Seltzer, W. H. Wong, and A. S. Bryk. Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics*, 21(2):131, 1996. doi: 10.2307/1165214.

A. Skrondal and S. Rabe-Hesketh. Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 172(3):659–687, 2009. doi: 10.1111/j.1467-985x.2009.00587.x.

J. X. Song and C. W. Ahn. An evaluation of methods for the stratified analysis of clustered binary data in community intervention trials. *Statistics in Medicine*, 22(13):2205–2216, 2003. doi: 10.1002/sim.1390.

E. L. Turner, F. Li, J. A. Gallis, M. Prague, and D. M. Murray. Review of recent methodological developments in group-randomized trials: Part 1—design. *American Journal of Public Health*, 107(6):907–915, 2017a. doi: 10.2105/ajph.2017.303706.

E. L. Turner, M. Prague, J. A. Gallis, F. Li, and D. M. Murray. Review of recent methodological developments in group-randomized trials: Part 2—analysis. *American Journal of Public Health*, 107(7):1078–1086, 2017b. doi: 10.2105/ajph.2017.303707.

B. Wang, E. L. Ogburn, and M. Rosenblum. Analysis of covariance in randomized trials: More precision and valid confidence intervals, without model assumptions. *Biometrics*, 75 (4):1391?1400, 2019. doi: 10.1111/biom.13062.

B. Wang, R. Susukida, R. Mojtabai, M. Amin-Esmaeili, and M. Rosenblum. Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *Journal of the American Statistical Association*, page 1?12, 2021. doi: 10.1080/01621459.2021.1981338.

M. Wang. *geesmv: Modified Variance Estimators for Generalized Estimating Equations*, 2015. URL `https://CRAN.R-project.org/package=nlme`. R package version ¿= 3.0.2.

M. Wang, L. Kong, Z. Li, and L. Zhang. Covariance estimators for generalized estimating equations (gee) in longitudinal analysis with small samples. *Statistics in Medicine*, 35(28): 5318–5319, 2016. doi: 10.1002/sim.7131.

J. F. Wiley and D. Hedeker. *brmsmargins: Bayesian Marginal Effects for 'brms' Models*, 2022. https://joshuawiley.com/brmsmargins/, https://github.com/JWiley/brmsmargins.

L. Yang and A. A. Tsiatis. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4), 2001.

X. Ye, N. Lu, and Y. Xu. Improving precision and power in randomized trials with a two-stage study design: Stratification using clustering method. *Pharmaceutical Statistics*, 22 (2):396–407, Mar 2023. doi: 10.1002/pst.2279.

L.-M. Yu, A.-W. Chan, S. Hopewell, J. J. Deeks, and D. G. Altman. Reporting on covariate adjustment in randomised controlled trials before and after revision of the 2001 consort statement: A literature review. *Trials*, 11(1), 2010. doi: 10.1186/1745-6215-11-59.

S. L. Zeger, K.-Y. Liang, and P. S. Albert. Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44(4):1049, 1988. doi: 10.2307/2531734.

R. Zhai and R. Gutman. Using the potential outcome framework to estimate optimal sample size for cluster randomized trials: A simulation-based algorithm. *Journal of Statistical Computation and Simulation*, 91(18):3744?3770, 2021. doi: 10.1080/00949655.2021.1946806.

J. Zhang and D. D. Boos. Mantel-haenszel test statistics for correlated binary data. *Biometrics*, 53(4):1185, 1997. doi: 10.2307/2533489.

2.7 CHAPTER 2 Appendix

2.7.1 CRT Variance Derivation

Take the general multilevel model from Neuhaus and Kalbfleisch,

$$Y_{ij} = c_j + \beta X_{ij} + \epsilon_{ij} \tag{2.9}$$

$$E(Y_{ij}|c_j, X_{ij}) = c_j + \beta X_{ij} \tag{2.10}$$

where $E(c_j) = \mu_c$, $V(c_j) = \sigma_c^2$, $E(\epsilon_{ij}) = 0$, $V(\epsilon_{ij}) = \sigma_\epsilon^2$, β is a predefined parameter, and c_j, X_{ij}, ϵ_{ij} are mutually independent. The variance of Y_{ij},

$$V(Y_{ij}) = V(Y_{ij} - c_j - \beta X_{ij} + c_j + \beta X_{ij})$$

$$= V(Y_{ij} - c_j - \beta X_{ij}) + V(c_j + \beta X_{ij})$$

$$+ 2Cov(Y_{ij} - c_j - \beta X_{ij}, c_j + \beta X_{ij})$$

First compute the covariance in the line above,

$$Cov(Y_{ij} - c_j - \beta X_{ij}, c_j + \beta X_{ij}) = E\big((Y_{ij} - c_j - \beta X_{ij})(c_j + \beta X_{ij})\big)$$

$$- E\big(Y_{ij} - c_j - \beta X_{ij}\big)E\big(c_j + \beta X_{ij}\big)$$

$$= E\big(Y_{ij}c_j + Y_{ij}\beta X_{ij} - c_j^2 - 2c_j\beta X_{ij} - \beta^2 X_{ij}^2\big)$$

$$- E\big(Y_{ij} - E(Y_{ij}|c_j, X_{ij})\big)E\big(c_j + \beta X_{ij}\big)$$

$$= E(Y_{ij}c_j) + \beta E(Y_{ij}X_{ij}) - E(c_j^2) - 2\beta E(X_{ij})E(c_j) - \beta^2 E(X_{ij}^2)$$

$$- \Big(E(Y_{ij}) - E\big(E(Y_{ij}|c_j, X_{ij})\big)\Big)E\big(c_j + \beta X_{ij}\big)$$

$$= Cov(Y_{ij}, c_j) + E(Y_{ij})E(c_j) + \beta E((c_j + \beta X_{ij} + \epsilon_{ij})X_{ij}) - \Big(V(c_j) + E(c_j)^2\Big)$$

$$- 2\beta E(X_{ij})\mu_c - \beta^2 E(X_{ij}^2) - \Big(E(Y_{ij}) - E(Y_{ij})\Big)E\big(c_j + \beta X_{ij}\big)$$

$$= Cov(Y_{ij}, c_j) + \mu_c(\mu_c + \beta E(X_{ij})) + \beta\mu_c E(X_{ij}) + \beta^2 E(X_{ij}^2) - V(c_j) - \mu_c^2$$

$$- 2\beta E(X_{ij})\mu_c - \beta^2 E(X_{ij}^2) - \Big(0\Big)E\big(c_j + \beta X_{ij}\big)$$

$$= Cov(Y_{ij}, c_j) - V(c_j)$$

$$= E(Y_{ij}c_j) - E(Y_{ij})E(c_j) - V(c_j)$$

$$= E\big((c_j + \beta X_{ij} + \epsilon_{ij})c_j\big) - E(Y_{ij})E(c_j) - V(c_j)$$

$$= E(c_j^2) + \beta E(X_{ij})\mu_c + \mu_c(0) - \mu_c^2 - \beta E(X_{ij})\mu_c - V(c_j)$$

$$= V(c_j) + E(\mu_c)^2 - \mu_c^2 - V(c_j)$$

$$= 0$$

The variance can be written as,

$$\begin{aligned} V(Y_{ij}) &= V(Y_{ij} - c_j - \beta X_{ij}) + V(c_j + \beta X_{ij}) + 2\big(0\big) \\ &= V(Y_{ij} - c_j - \beta X_{ij}) + V(c_j) + V(\beta X_{ij}) + 2Cov(c_j, \beta X_{ij}) \\ &= V(Y_{ij} - c_j - \beta X_{ij}) + V(c_j) + V(\beta X_{ij}) + 2Cov(c_j, \beta X_{ij}) \end{aligned}$$

Computing the covariance term,

$$\begin{aligned} Cov(c_j, \beta X_{ij}) =& E(c_j \beta X_{ij}) - E(c_j)E(\beta X_{ij}) \\ =& \beta E(c_j)E(X_{ij}) - \beta E(c_j)E(X_{ij}) \\ =& 0 \end{aligned}$$

The variance can be simplified to the following form,

$$V(Y_{ij}) = \ V(Y_{ij} - c_j - \beta X_{ij}) + V(c_j) + V(\beta X_{ij}) \tag{2.11}$$

The variance is broken down into several components, a residual variance, a cluster variance, a predicted variance, and the covariance between the outcomes and clusters.

We can extend this derivation for

$$Y_{ij}(T_j) = c_j + h\big(\mathbf{Z}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\beta}, T_j\big) + \epsilon_{ij}$$

to obtain

$$V(Y_{ij}(T_j)) = \ V\Big(Y_{ij}(T_j) - c_j - h\big(\mathbf{Z}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\beta}, T_j\big)\Big) + V\Big(c_j\Big) + V\Big(h\big(\mathbf{Z}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\beta}, T_j\big)\Big).$$

Additionally we can extend our model from (2.9) and (2.10) to account for between- / within-level cluster effects and a treatment effect,

$$Y_{ij} = c_j + \beta X_{ij} + \beta_T T + \beta_W (X_{ij} - \bar{X}_j) + \beta_B \bar{X}_j + \epsilon_{ij}$$

$$E(Y_{ij}|c_j, X_{ij}, T) = c_j + \beta_T T + \beta_W (X_{ij} - \bar{X}_j) + \beta_B \bar{X}_j$$

Substituting into our derivation we obtain,

$$V(Y_{ij}) = V\Big(Y_{ij} - c_j - \beta_T T + \beta_W (X_{ij} - \bar{X}_j) + \beta_B \bar{X}_j\Big) + V\Big(c_j\Big)$$
$$+ V\Big(\beta_T T + \beta_W (X_{ij} - \bar{X}_j) + \beta_B \bar{X}_j\Big)$$

2.7.2 GLMM Bayesian Estimation Procedures

The R packages brmsmargins (Wiley and Hedeker, 2022) and marginaleffects (Arel-Bundock, 2022) are used to obtain the point estimates for the estimand of interest, γ, and their corresponding sampling variance for each of the corresponding models. These estimates are obtained directly from the posterior distribution of γ as follows,

$$p(\gamma \mid \mathbf{Y}^{obs}) = \int\int\int\int \prod_{ij} p(\gamma \mid Y_{ij}^{obs}, \boldsymbol{\beta}, c_j, \mathbf{X}_{ij}, \mathbf{Z}_j) p(\boldsymbol{\beta}, c_j \mid Y_{ij}^{obs}, \mathbf{X}_{ij}, \mathbf{Z}_j) p(\mathbf{X}_{ij}, \mathbf{Z}_j) \; d\boldsymbol{\beta} \; d\mathbf{c} \; d\mathbf{X} \; d\mathbf{Z}$$

2.7.3 Continuous Outcomes Model Specification 1 Results

Table 2.19: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	99 (98,100)	97 (96,100)	95 (92,97)	98 (97,100)	97 (95,100)	95 (93,97)	98 (97,100)	97 (95,98)	95 (93,97)
Strata	96 (94,97)	95 (94,97)	94 (91,96)	96 (94,97)	95 (94,97)	94 (91,97)	94 (91,97)	95 (94,97)	95 (94,97)
True	95 (94,97)	95 (94,97)	94 (91,97)	96 (94,98)	96 (94,98)	94 (91,97)	95 (94,97)	94 (91,97)	94 (91,96)
NKM	95 (94,97)	95 (94,97)	94 (91,97)	96 (94,98)	95 (94,97)	94 (91,97)	95 (94,99)	95 (94,97)	94 (92,97)
Standard	99 (100,100)	99 (97,100)	95 (94,97)	99 (99,100)	98 (97,100)	96 (94,98)	99 (100,100)	98 (97,100)	96 (94,97)
GEE True	95 (94,96)	96 (96,97)	95 (94,96)	95 (94,96)	96 (94,96)	95 (93,97)	95 (94,96)	95 (94,96)	95 (94,96)
GEE TRM	98 (96,99)	98 (96,99)	96 (94,98)	98 (97,100)	97 (96,98)	95 (94,97)	99 (98,100)	97 (96,99)	96 (94,97)

Table 2.20: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	98 (97,100)	97 (96,98)	96 (95,97)	98 (97,99)	97 (96,100)	95 (94,97)	97 (96,99)	97 (96,98)	97 (96,100)
Strata	96 (94,98)	96 (95,97)	95 (94,96)	96 (94,98)	96 (95,97)	94 (92,96)	95 (94,97)	96 (95,98)	97 (94,100)
True	95 (91,97)	95 (94,96)	94 (94,97)	96 (94,97)	94 (94,97)	93 (91,96)	95 (93,96)	95 (93,97)	96 (94,100)
NKM	94 (92,96)	95 (94,97)	95 (92,97)	95 (94,97)	96 (94,97)	95 (93,96)	95 (94,95)	96 (94,97)	96 (94,97)
Standard	99 (99,100)	98 (97,99)	97 (95,97)	99 (99,100)	98 (97,100)	96 (95,97)	99 (97,100)	98 (97,100)	98 (96,100)
GEE True	96 (94,98)	95 (94,96)	95 (94,96)	96 (94,97)	95 (95,97)	95 (94,97)	95 (94,97)	95 (94,97)	95 (94,96)
GEE TRM	99 (98,100)	98 (96,99)	96 (95,97)	99 (98,100)	98 (97,98)	96 (95,98)	99 (98,100)	98 (96,98)	96 (95,98)

Table 2.21: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE1, CO, MS1}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 1$			$\beta_9 = 3$		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	98 (97,100)	97 (95,100)	96 (94,97)	97 (96,100)	97 (94,100)	95 (94,97)
Strata	95 (94,97)	95 (94,97)	95 (94,97)	95 (92,97)	95 (94,97)	95 (94,97)
True	94 (91,97)	94 (94,97)	94 (93,97)	91 (89,94)	94 (91,97)	94 (92,97)
NKM	94 (92,97)	95 (94,97)	94 (94,97)	95 (93,97)	95 (94,97)	95 (94,97)
Standard	99 (98,100)	98 (97,100)	96 (94,99)	98 (97,100)	97 (97,100)	96 (94,97)
GEE True	95 (94,97)	95 (94,96)	95 (94,97)	92 (91,94)	94 (92,96)	94 (93,96)
GEE TRM	98 (97,99)	97 (96,98)	96 (95,97)	97 (95,98)	96 (95,98)	95 (94,97)

Table 2.22: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE1, CO, MS1}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 1$			$\beta_9 = 3$		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	99 (98,100)	98 (97,98)	97 (96,98)	99 (98,100)	98 (98,99)	97 (96,99)
Strata	98 (97,100)	97 (96,98)	97 (96,98)	99 (98,100)	98 (97,99)	97 (96,99)
True	95 (93,97)	94 (93,96)	94 (91,96)	92 (90,93)	93 (92,94)	95 (93,96)
NKM	95 (94,97)	94 (93,96)	95 (93,97)	94 (93,96)	95 (94,96)	96 (94,97)
Standard	99 (98,100)	98 (97,99)	97 (97,99)	99 (98,100)	99 (98,100)	98 (97,99)
GEE True	95 (94,97)	95 (94,97)	95 (93,96)	92 (90,94)	94 (93,96)	95 (94,96)
GEE TRM	99 (98,100)	98 (98,99)	97 (97,98)	99 (98,100)	99 (98,100)	98 (97,99)

Table 2.23: Mean and interquartile range of interval widths for each model for Simulation {SR1, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	1.45 (1.16,1.74)	2.02 (1.8,2.26)	3.04 (2.92,3.15)	1.15 (0.92,1.39)	1.59 (1.45,1.75)	2.46 (2.4,2.53)	0.99 (0.8,1.2)	1.37 (1.26,1.5)	2.09 (2.03,2.13)
Strata	1.13 (1.08,1.16)	1.78 (1.72,1.84)	2.91 (2.85,2.96)	0.9 (0.86,0.95)	1.41 (1.38,1.42)	2.35 (2.33,2.38)	0.77 (0.74,0.79)	1.22 (1.19,1.24)	2.01 (1.99,2.03)
True	1.05 (1.03,1.07)	1.73 (1.69,1.76)	2.9 (2.85,2.95)	0.84 (0.83,0.85)	1.38 (1.36,1.4)	2.33 (2.31,2.36)	0.72 (0.71,0.73)	1.18 (1.17,1.21)	1.97 (1.93,2.02)
NKM	1.05 (1.03,1.07)	1.74 (1.71,1.76)	2.9 (2.86,2.97)	0.84 (0.83,0.85)	1.38 (1.36,1.39)	2.34 (2.32,2.38)	0.72 (0.72,0.73)	1.18 (1.17,1.19)	1.98 (1.95,2.02)
Standard	1.82 (1.46,2.17)	2.3 (2,2.57)	3.23 (3.04,3.34)	1.46 (1.17,1.74)	1.82 (1.61,2.01)	2.6 (2.48,2.7)	1.25 (1,1.5)	1.56 (1.38,1.74)	2.2 (2.13,2.27)
GEE True	1.05 (1.04,1.07)	1.74 (1.72,1.75)	2.97 (2.94,3)	0.84 (0.83,0.85)	1.4 (1.38,1.41)	2.37 (2.35,2.39)	0.72 (0.72,0.73)	1.19 (1.18,1.19)	2.03 (2.01,2.05)
GEE TRM	1.43 (1.13,1.75)	1.97 (1.77,2.18)	3.08 (2.95,3.2)	1.14 (0.91,1.39)	1.6 (1.44,1.78)	2.48 (2.4,2.61)	0.99 (0.79,1.2)	1.37 (1.23,1.53)	2.13 (2.06,2.23)

Table 2.24: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	1.14 (1.09,1.19)	1.46 (1.43,1.47)	2.2 (2.18,2.2)	0.93 (0.88,0.98)	1.2 (1.17,1.2)	1.78 (1.76,1.79)	0.78 (0.76,0.79)	1.03 (1.01,1.04)	1.53 (1.52,1.56)
Strata	1 (0.98,1.01)	1.37 (1.36,1.39)	2.13 (2.1,2.15)	0.81 (0.8,0.82)	1.12 (1.11,1.12)	1.72 (1.7,1.74)	0.7 (0.69,0.71)	0.96 (0.95,0.97)	1.49 (1.48,1.5)
True	0.72 (0.72,0.73)	1.18 (1.18,1.19)	2.01 (2,2.04)	0.58 (0.58,0.59)	0.96 (0.96,0.97)	1.63 (1.61,1.64)	0.5 (0.5,0.5)	0.83 (0.82,0.83)	1.39 (1.38,1.4)
NKM	0.8 (0.79,0.81)	1.24 (1.22,1.26)	2.05 (2.03,2.08)	0.65 (0.65,0.66)	1.01 (1,1.01)	1.66 (1.65,1.67)	0.56 (0.56,0.56)	0.87 (0.86,0.87)	1.43 (1.42,1.43)
Standard	1.34 (1.26,1.43)	1.6 (1.56,1.62)	2.28 (2.23,2.29)	1.09 (1.02,1.18)	1.31 (1.27,1.32)	1.85 (1.82,1.86)	0.91 (0.87,0.92)	1.13 (1.09,1.13)	1.6 (1.58,1.61)
GEE True	0.73 (0.72,0.73)	1.21 (1.21,1.23)	2.07 (2.05,2.1)	0.59 (0.59,0.59)	0.97 (0.97,0.98)	1.66 (1.65,1.67)	0.51 (0.51,0.51)	0.84 (0.83,0.84)	1.43 (1.42,1.44)
GEE TRM	1.23 (1.1,1.37)	1.56 (1.45,1.69)	2.26 (2.19,2.34)	1 (0.89,1.12)	1.27 (1.18,1.37)	1.84 (1.79,1.91)	0.86 (0.77,0.97)	1.09 (1.02,1.18)	1.59 (1.54,1.64)

Table 2.25: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE1, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	1.48 (1.27,1.77)	2.04 (1.89,2.23)	3.06 (2.98,3.13)	1.18 (1.01,1.41)	1.62 (1.5,1.77)	2.45 (2.39,2.5)	1.01 (0.87,1.2)	1.39 (1.29,1.52)	2.11 (2.06,2.16)
Strata	1.2 (1.13,1.24)	1.84 (1.79,1.88)	2.95 (2.91,2.98)	0.96 (0.9,1)	1.46 (1.42,1.49)	2.37 (2.34,2.4)	0.81 (0.76,0.85)	1.24 (1.22,1.27)	2.02 (1.99,2.05)
True	1.06 (1.04,1.07)	1.74 (1.72,1.77)	2.9 (2.86,2.94)	0.84 (0.83,0.85)	1.38 (1.36,1.41)	2.32 (2.29,2.36)	0.72 (0.71,0.73)	1.19 (1.17,1.2)	1.97 (1.95,2.01)
NKM	1.12 (1.07,1.19)	1.79 (1.74,1.84)	2.93 (2.9,2.98)	0.9 (0.85,0.95)	1.43 (1.4,1.44)	2.35 (2.32,2.39)	0.77 (0.73,0.82)	1.22 (1.2,1.24)	2.01 (1.97,2.04)
Standard	2 (1.79,2.28)	2.43 (2.25,2.66)	3.31 (3.21,3.42)	1.6 (1.43,1.82)	1.95 (1.82,2.1)	2.66 (2.57,2.75)	1.37 (1.24,1.57)	1.67 (1.55,1.83)	2.3 (2.22,2.37)
GEE True	1.05 (1.04,1.07)	1.74 (1.73,1.76)	2.97 (2.94,3)	0.84 (0.84,0.85)	1.39 (1.38,1.41)	2.37 (2.35,2.4)	0.72 (0.72,0.73)	1.19 (1.18,1.2)	2.03 (2.01,2.05)
GEE TRM	1.61 (1.46,1.8)	2.14 (2.02,2.28)	3.19 (3.08,3.27)	1.3 (1.18,1.44)	1.71 (1.62,1.82)	2.56 (2.48,2.64)	1.12 (1.02,1.23)	1.48 (1.39,1.56)	2.21 (2.14,2.28)

Table 2.26: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE1, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	2.08 (1.97,2.19)	2.26 (2.15,2.37)	2.78 (2.69,2.85)	1.69 (1.6,1.78)	1.86 (1.78,1.94)	2.28 (2.22,2.35)	1.45 (1.37,1.54)	1.6 (1.53,1.68)	1.96 (1.91,2.01)
Strata	1.97 (1.86,2.09)	2.16 (2.05,2.26)	2.68 (2.61,2.75)	1.59 (1.5,1.67)	1.76 (1.69,1.85)	2.2 (2.14,2.25)	1.37 (1.29,1.45)	1.52 (1.45,1.59)	1.9 (1.86,1.93)
True	0.72 (0.72,0.73)	1.19 (1.18,1.19)	2.01 (1.99,2.04)	0.58 (0.58,0.59)	0.96 (0.95,0.97)	1.63 (1.61,1.64)	0.5 (0.5,0.5)	0.83 (0.82,0.84)	1.4 (1.39,1.41)
NKM	1.23 (1.2,1.25)	1.55 (1.52,1.58)	2.24 (2.22,2.26)	0.99 (0.97,1.01)	1.27 (1.25,1.28)	1.83 (1.82,1.85)	0.86 (0.84,0.87)	1.09 (1.07,1.1)	1.58 (1.56,1.6)
Standard	2.19 (2.07,2.28)	2.34 (2.24,2.45)	2.83 (2.75,2.91)	1.79 (1.68,1.85)	1.92 (1.85,2)	2.33 (2.27,2.4)	1.52 (1.44,1.6)	1.66 (1.6,1.74)	2.01 (1.96,2.06)
GEE True	0.73 (0.73,0.74)	1.21 (1.2,1.22)	2.07 (2.05,2.08)	0.59 (0.59,0.59)	0.98 (0.97,0.98)	1.66 (1.65,1.67)	0.51 (0.5,0.51)	0.84 (0.83,0.84)	1.43 (1.42,1.43)
GEE TRM	2.16 (2.04,2.25)	2.36 (2.24,2.46)	2.87 (2.79,2.96)	1.75 (1.64,1.83)	1.92 (1.82,1.99)	2.34 (2.27,2.41)	1.51 (1.41,1.59)	1.65 (1.57,1.72)	2.02 (1.96,2.08)

Table 2.27: Mean and interquartile range of absolute biases for each model for Simulation {SR1, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	0.22 (0.2,0.24)	0.35 (0.32,0.38)	0.58 (0.56,0.61)	0.17 (0.16,0.19)	0.28 (0.26,0.29)	0.48 (0.44,0.51)	0.15 (0.14,0.17)	0.25 (0.23,0.27)	0.4 (0.38,0.42)
Strata	0.22 (0.2,0.23)	0.35 (0.32,0.38)	0.58 (0.53,0.62)	0.17 (0.15,0.19)	0.27 (0.26,0.29)	0.48 (0.45,0.51)	0.15 (0.14,0.17)	0.25 (0.23,0.26)	0.4 (0.38,0.43)
True	0.2 (0.19,0.22)	0.34 (0.31,0.37)	0.58 (0.53,0.61)	0.16 (0.15,0.17)	0.26 (0.24,0.28)	0.48 (0.45,0.51)	0.14 (0.13,0.15)	0.24 (0.23,0.26)	0.4 (0.37,0.43)
NKM	0.2 (0.19,0.22)	0.34 (0.31,0.37)	0.59 (0.54,0.62)	0.16 (0.15,0.17)	0.26 (0.24,0.28)	0.47 (0.44,0.52)	0.14 (0.13,0.15)	0.24 (0.23,0.26)	0.4 (0.37,0.43)
Standard	0.24 (0.22,0.27)	0.36 (0.32,0.4)	0.59 (0.53,0.65)	0.19 (0.17,0.2)	0.29 (0.27,0.31)	0.48 (0.45,0.52)	0.17 (0.14,0.19)	0.26 (0.24,0.28)	0.41 (0.37,0.43)
GEE True	0.2 (0.19,0.21)	0.33 (0.31,0.34)	0.58 (0.55,0.6)	0.16 (0.15,0.17)	0.27 (0.25,0.28)	0.48 (0.45,0.5)	0.14 (0.13,0.15)	0.23 (0.21,0.25)	0.4 (0.38,0.43)
GEE TRM	0.22 (0.2,0.24)	0.33 (0.32,0.35)	0.58 (0.55,0.61)	0.18 (0.16,0.19)	0.28 (0.27,0.28)	0.48 (0.46,0.51)	0.15 (0.14,0.16)	0.24 (0.22,0.26)	0.41 (0.39,0.43)

Table 2.28: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	0.18 (0.17,0.19)	0.26 (0.25,0.28)	0.43 (0.39,0.46)	0.15 (0.14,0.17)	0.21 (0.2,0.22)	0.35 (0.32,0.38)	0.13 (0.12,0.14)	0.19 (0.18,0.2)	0.28 (0.26,0.32)
Strata	0.18 (0.17,0.19)	0.26 (0.25,0.28)	0.42 (0.4,0.45)	0.15 (0.14,0.17)	0.21 (0.2,0.22)	0.35 (0.33,0.37)	0.13 (0.12,0.14)	0.19 (0.18,0.2)	0.28 (0.25,0.32)
True	0.14 (0.13,0.15)	0.23 (0.22,0.24)	0.41 (0.38,0.43)	0.12 (0.11,0.13)	0.19 (0.18,0.2)	0.34 (0.32,0.36)	0.1 (0.1,0.11)	0.17 (0.16,0.18)	0.27 (0.24,0.31)
NKM	0.16 (0.15,0.17)	0.24 (0.23,0.25)	0.42 (0.39,0.44)	0.13 (0.13,0.14)	0.2 (0.18,0.2)	0.34 (0.32,0.36)	0.12 (0.11,0.12)	0.17 (0.16,0.19)	0.27 (0.24,0.31)
Standard	0.19 (0.18,0.2)	0.27 (0.25,0.28)	0.43 (0.39,0.45)	0.16 (0.15,0.17)	0.21 (0.2,0.23)	0.35 (0.32,0.38)	0.13 (0.13,0.14)	0.19 (0.18,0.2)	0.28 (0.26,0.32)
GEE True	0.14 (0.14,0.15)	0.24 (0.23,0.25)	0.4 (0.38,0.42)	0.12 (0.11,0.12)	0.19 (0.19,0.2)	0.34 (0.32,0.36)	0.1 (0.09,0.11)	0.17 (0.16,0.18)	0.29 (0.28,0.31)
GEE TRM	0.18 (0.17,0.19)	0.27 (0.26,0.28)	0.42 (0.4,0.43)	0.15 (0.14,0.16)	0.22 (0.21,0.23)	0.35 (0.33,0.38)	0.13 (0.12,0.14)	0.19 (0.18,0.2)	0.3 (0.29,0.32)

Table 2.29: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE1, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	0.23 (0.21,0.25)	0.35 (0.31,0.38)	0.57 (0.52,0.6)	0.19 (0.17,0.2)	0.3 (0.28,0.32)	0.48 (0.45,0.51)	0.17 (0.15,0.18)	0.25 (0.23,0.27)	0.4 (0.37,0.42)
Strata	0.23 (0.21,0.25)	0.35 (0.32,0.38)	0.57 (0.53,0.6)	0.18 (0.17,0.2)	0.29 (0.27,0.32)	0.48 (0.44,0.51)	0.16 (0.14,0.18)	0.25 (0.23,0.27)	0.4 (0.37,0.43)
True	0.22 (0.2,0.23)	0.34 (0.31,0.37)	0.55 (0.51,0.6)	0.18 (0.15,0.19)	0.29 (0.27,0.3)	0.47 (0.44,0.51)	0.16 (0.14,0.17)	0.25 (0.23,0.26)	0.4 (0.37,0.43)
NKM	0.22 (0.2,0.23)	0.35 (0.31,0.37)	0.56 (0.51,0.6)	0.18 (0.16,0.19)	0.29 (0.27,0.31)	0.48 (0.45,0.5)	0.16 (0.14,0.17)	0.25 (0.22,0.27)	0.4 (0.37,0.42)
Standard	0.29 (0.25,0.32)	0.39 (0.36,0.41)	0.59 (0.54,0.63)	0.23 (0.2,0.26)	0.33 (0.31,0.36)	0.51 (0.47,0.54)	0.21 (0.18,0.23)	0.28 (0.25,0.31)	0.42 (0.39,0.45)
GEE True	0.22 (0.21,0.23)	0.35 (0.33,0.37)	0.57 (0.54,0.59)	0.18 (0.16,0.19)	0.28 (0.27,0.3)	0.48 (0.46,0.5)	0.15 (0.14,0.16)	0.24 (0.23,0.25)	0.41 (0.38,0.43)
GEE TRM	0.27 (0.24,0.3)	0.38 (0.36,0.41)	0.59 (0.56,0.61)	0.23 (0.2,0.25)	0.31 (0.29,0.33)	0.5 (0.48,0.52)	0.19 (0.16,0.21)	0.27 (0.25,0.29)	0.43 (0.4,0.45)

Table 2.30: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE1, CO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	0.33 (0.31,0.35)	0.39 (0.37,0.41)	0.51 (0.48,0.54)	0.27 (0.26,0.29)	0.3 (0.28,0.32)	0.38 (0.36,0.4)	0.23 (0.21,0.24)	0.26 (0.25,0.28)	0.35 (0.32,0.37)
Strata	0.33 (0.31,0.35)	0.39 (0.37,0.41)	0.51 (0.48,0.53)	0.27 (0.25,0.29)	0.3 (0.29,0.32)	0.38 (0.36,0.4)	0.22 (0.21,0.23)	0.26 (0.25,0.28)	0.35 (0.33,0.37)
True	0.16 (0.14,0.17)	0.25 (0.23,0.25)	0.43 (0.4,0.45)	0.13 (0.12,0.14)	0.2 (0.19,0.21)	0.33 (0.31,0.34)	0.11 (0.1,0.12)	0.17 (0.16,0.18)	0.29 (0.27,0.31)
NKM	0.25 (0.24,0.27)	0.31 (0.3,0.32)	0.46 (0.44,0.48)	0.2 (0.19,0.22)	0.26 (0.25,0.27)	0.35 (0.33,0.37)	0.18 (0.17,0.19)	0.22 (0.21,0.23)	0.32 (0.3,0.34)
Standard	0.33 (0.31,0.35)	0.39 (0.37,0.4)	0.51 (0.48,0.54)	0.27 (0.26,0.29)	0.3 (0.29,0.32)	0.38 (0.36,0.4)	0.23 (0.21,0.24)	0.26 (0.25,0.28)	0.36 (0.33,0.37)
GEE True	0.16 (0.14,0.17)	0.24 (0.23,0.25)	0.42 (0.4,0.44)	0.13 (0.12,0.13)	0.2 (0.19,0.21)	0.33 (0.32,0.35)	0.11 (0.1,0.11)	0.17 (0.17,0.18)	0.29 (0.28,0.31)
GEE TRM	0.33 (0.32,0.35)	0.38 (0.36,0.39)	0.51 (0.48,0.53)	0.27 (0.25,0.29)	0.31 (0.29,0.33)	0.4 (0.38,0.42)	0.23 (0.21,0.24)	0.27 (0.26,0.28)	0.35 (0.33,0.37)

2.7.4 Binary Outcomes Model Specification 1 Results

Table 2.31: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE0, BO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	97 (96,100)	96 (94,97)	96 (94,97)	95 (94,97)	95 (91,98)	96 (94,98)	97 (94,100)	96 (94,99)	97 (97,97)
Strata	96 (94,97)	95 (93,97)	96 (93,98)	94 (92,95)	95 (94,97)	96 (94,98)	96 (94,97)	94 (91,97)	96 (95,97)
True	96 (94,97)	95 (91,97)	96 (94,97)	94 (93,95)	95 (94,97)	96 (94,97)	96 (94,97)	95 (91,97)	95 (94,97)
NKM	96 (95,97)	94 (91,97)	95 (93,97)	93 (91,95)	95 (91,97)	96 (94,97)	96 (94,97)	95 (94,97)	95 (94,97)
Standard	99 (97,100)	98 (97,100)	97 (96,100)	98 (97,100)	98 (97,100)	98 (97,100)	99 (97,100)	98 (97,100)	98 (97,100)
GEE True	94 (93,96)	95 (92,96)	95 (93,96)	95 (94,96)	95 (93,96)	95 (94,96)	95 (94,96)	95 (94,97)	95 (93,97)
GEE TRM	95 (94,96)	95 (93,98)	95 (93,96)	95 (94,97)	95 (94,96)	95 (94,96)	96 (94,97)	96 (94,97)	95 (93,97)

Table 2.32: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE0, BO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	96 (95,98)	97 (94,100)	96 (95,98)	97 (96,100)	97 (94,100)	96 (94,97)	96 (94,97)	97 (97,98)	95 (94,97)
Strata	96 (94,99)	96 (93,99)	96 (94,97)	97 (96,100)	97 (96,100)	95 (94,97)	95 (94,97)	97 (96,98)	95 (94,97)
True	94 (93,97)	95 (91,97)	95 (93,97)	94 (94,97)	97 (94,100)	95 (94,97)	94 (93,97)	96 (96,97)	95 (91,98)
NKM	95 (92,96)	94 (91,97)	95 (94,97)	95 (94,97)	96 (94,100)	95 (94,97)	94 (93,97)	96 (95,97)	94 (91,97)
Standard	99 (97,100)	98 (97,100)	97 (96,98)	97 (97,100)	99 (97,100)	97 (97,98)	98 (97,100)	99 (97,100)	97 (97,99)
GEE True	95 (94,96)	95 (94,97)	96 (94,97)	95 (94,97)	95 (94,96)	95 (94,96)	95 (94,96)	95 (94,97)	95 (94,96)
GEE TRM	97 (96,98)	97 (96,98)	97 (96,98)	97 (96,99)	96 (95,98)	96 (94,97)	97 (96,98)	97 (95,98)	96 (95,97)

Table 2.33: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE1, BO, MS1}. Results are stratified by β_9 and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	96 (94,97)	95 (94,97)	95 (94,97)	95 (94,97)	95 (94,97)	96 (94,97)	95 (94,98)	96 (94,97)	96 (94,99)
Strata	95 (91,97)	95 (94,97)	94 (94,97)	95 (94,97)	95 (94,97)	94 (93,97)	94 (91,97)	95 (94,97)	96 (94,97)
True	94 (91,97)	95 (94,97)	95 (92,97)	95 (91,97)	94 (92,97)	94 (91,97)	94 (91,97)	95 (93,97)	95 (94,97)
NKM	95 (94,97)	95 (94,97)	95 (93,97)	95 (92,97)	95 (94,97)	95 (93,97)	94 (92,97)	95 (94,97)	96 (94,97)
Standard	97 (95,100)	97 (94,100)	97 (96,98)	98 (97,100)	98 (96,100)	97 (94,100)	98 (97,100)	98 (97,100)	98 (97,100)
GEE True	95 (93,96)	94 (93,96)	94 (92,95)	94 (93,96)	94 (93,96)	94 (92,96)	95 (93,97)	94 (93,96)	94 (93,96)
GEE TRM	96 (94,97)	95 (94,96)	94 (93,95)	95 (94,97)	94 (92,96)	94 (93,96)	95 (94,97)	95 (93,96)	95 (94,96)

Table 2.34: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE1, BO, MS1}. Results are stratified by β_9 and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	98 (97,100)	97 (96,100)	96 (94,97)	97 (96,100)	96 (94,97)	97 (95,100)	97 (97,100)	97 (94,97)	95 (94,97)
Strata	97 (97,98)	97 (96,100)	96 (94,97)	95 (94,97)	97 (96,97)	96 (94,97)	97 (96,98)	96 (94,97)	94 (94,97)
True	96 (94,97)	97 (95,98)	95 (93,97)	94 (92,95)	95 (93,97)	96 (94,97)	94 (91,97)	94 (92,95)	93 (92,95)
NKM	96 (95,97)	97 (95,100)	95 (94,97)	93 (91,94)	95 (94,97)	96 (94,97)	94 (91,97)	94 (92,96)	94 (93,95)
Standard	99 (97,100)	98 (97,100)	97 (95,100)	99 (98,100)	98 (97,100)	98 (96,100)	99 (97,100)	97 (97,100)	95 (94,97)
GEE True	95 (94,97)	95 (93,97)	95 (94,97)	95 (93,96)	95 (94,96)	95 (93,97)	95 (94,96)	95 (94,96)	96 (94,97)
GEE TRM	97 (96,98)	96 (95,98)	96 (95,97)	97 (96,98)	97 (96,98)	96 (94,97)	97 (96,98)	97 (95,98)	96 (95,98)

Table 2.35: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE0, BO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.25 (0.23,0.26)	0.27 (0.25,0.29)	0.33 (0.31,0.34)	0.2 (0.18,0.21)	0.22 (0.2,0.23)	0.27 (0.25,0.28)	0.17 (0.16,0.18)	0.19 (0.17,0.2)	0.23 (0.22,0.24)
Strata	0.23 (0.22,0.25)	0.26 (0.24,0.27)	0.32 (0.31,0.33)	0.19 (0.17,0.2)	0.21 (0.2,0.22)	0.26 (0.24,0.27)	0.16 (0.15,0.17)	0.18 (0.17,0.19)	0.22 (0.21,0.23)
True	0.23 (0.22,0.25)	0.26 (0.24,0.27)	0.32 (0.3,0.33)	0.19 (0.17,0.2)	0.21 (0.19,0.22)	0.26 (0.24,0.27)	0.16 (0.15,0.17)	0.18 (0.16,0.18)	0.22 (0.21,0.23)
NKM	0.23 (0.22,0.25)	0.26 (0.24,0.27)	0.32 (0.3,0.33)	0.19 (0.17,0.2)	0.21 (0.19,0.22)	0.26 (0.24,0.27)	0.16 (0.15,0.17)	0.18 (0.16,0.18)	0.22 (0.21,0.23)
Standard	0.29 (0.27,0.32)	0.31 (0.29,0.34)	0.36 (0.34,0.38)	0.24 (0.22,0.26)	0.25 (0.23,0.27)	0.29 (0.28,0.3)	0.21 (0.19,0.22)	0.22 (0.2,0.23)	0.25 (0.24,0.26)
GEE True	0.24 (0.22,0.25)	0.27 (0.25,0.28)	0.33 (0.31,0.34)	0.19 (0.18,0.2)	0.21 (0.2,0.22)	0.26 (0.24,0.28)	0.17 (0.15,0.17)	0.18 (0.17,0.19)	0.23 (0.21,0.23)
GEE TRM	0.25 (0.23,0.26)	0.27 (0.26,0.29)	0.33 (0.31,0.35)	0.2 (0.18,0.21)	0.22 (0.2,0.24)	0.27 (0.25,0.28)	0.17 (0.16,0.18)	0.19 (0.17,0.2)	0.23 (0.21,0.24)

Table 2.36: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE0, BO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.2 (0.2,0.21)	0.22 (0.22,0.23)	0.26 (0.26,0.26)	0.17 (0.16,0.17)	0.18 (0.18,0.18)	0.21 (0.21,0.21)	0.14 (0.14,0.15)	0.16 (0.15,0.16)	0.18 (0.18,0.19)
Strata	0.19 (0.19,0.2)	0.21 (0.21,0.22)	0.25 (0.25,0.25)	0.16 (0.16,0.16)	0.18 (0.17,0.18)	0.21 (0.2,0.21)	0.14 (0.13,0.14)	0.15 (0.15,0.15)	0.18 (0.18,0.18)
True	0.18 (0.17,0.18)	0.2 (0.2,0.2)	0.24 (0.24,0.24)	0.14 (0.14,0.15)	0.16 (0.16,0.16)	0.2 (0.2,0.2)	0.12 (0.12,0.13)	0.14 (0.14,0.14)	0.17 (0.17,0.17)
NKM	0.18 (0.17,0.18)	0.2 (0.2,0.2)	0.24 (0.24,0.24)	0.14 (0.14,0.15)	0.16 (0.16,0.16)	0.2 (0.2,0.2)	0.12 (0.12,0.13)	0.14 (0.14,0.14)	0.17 (0.17,0.17)
Standard	0.24 (0.23,0.25)	0.25 (0.25,0.26)	0.28 (0.28,0.29)	0.2 (0.19,0.21)	0.21 (0.2,0.21)	0.23 (0.23,0.23)	0.17 (0.17,0.18)	0.18 (0.18,0.18)	0.2 (0.2,0.2)
GEE True	0.18 (0.18,0.18)	0.2 (0.2,0.2)	0.25 (0.24,0.25)	0.15 (0.14,0.15)	0.16 (0.16,0.16)	0.2 (0.2,0.2)	0.12 (0.12,0.13)	0.14 (0.14,0.14)	0.17 (0.17,0.17)
GEE TRM	0.21 (0.2,0.21)	0.22 (0.22,0.23)	0.26 (0.26,0.27)	0.17 (0.16,0.17)	0.18 (0.18,0.19)	0.21 (0.21,0.22)	0.14 (0.14,0.15)	0.16 (0.15,0.16)	0.18 (0.18,0.19)

Table 2.37: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE1, BO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.21 (0.19,0.23)	0.23 (0.21,0.25)	0.28 (0.26,0.3)	0.16 (0.14,0.19)	0.18 (0.16,0.2)	0.22 (0.21,0.25)	0.14 (0.13,0.16)	0.15 (0.14,0.17)	0.19 (0.18,0.21)
Strata	0.2 (0.18,0.22)	0.22 (0.2,0.24)	0.27 (0.26,0.29)	0.16 (0.15,0.17)	0.17 (0.16,0.19)	0.22 (0.21,0.24)	0.14 (0.12,0.15)	0.15 (0.14,0.16)	0.19 (0.18,0.21)
True	0.19 (0.18,0.21)	0.21 (0.2,0.23)	0.27 (0.25,0.29)	0.15 (0.14,0.17)	0.17 (0.16,0.18)	0.22 (0.2,0.24)	0.13 (0.12,0.15)	0.15 (0.14,0.16)	0.19 (0.17,0.2)
NKM	0.2 (0.18,0.22)	0.22 (0.2,0.24)	0.27 (0.26,0.29)	0.16 (0.14,0.17)	0.17 (0.16,0.19)	0.22 (0.21,0.24)	0.14 (0.12,0.15)	0.15 (0.14,0.16)	0.19 (0.18,0.2)
Standard	0.25 (0.22,0.27)	0.26 (0.24,0.29)	0.31 (0.29,0.33)	0.2 (0.17,0.22)	0.21 (0.19,0.23)	0.25 (0.23,0.27)	0.17 (0.15,0.19)	0.18 (0.16,0.2)	0.21 (0.19,0.23)
GEE True	0.2 (0.18,0.22)	0.22 (0.2,0.24)	0.27 (0.25,0.3)	0.16 (0.14,0.18)	0.17 (0.16,0.19)	0.22 (0.2,0.24)	0.13 (0.12,0.15)	0.15 (0.14,0.17)	0.18 (0.17,0.2)
GEE TRM	0.2 (0.18,0.23)	0.22 (0.2,0.25)	0.27 (0.25,0.3)	0.16 (0.15,0.18)	0.18 (0.16,0.2)	0.22 (0.2,0.24)	0.14 (0.13,0.16)	0.15 (0.14,0.17)	0.19 (0.17,0.21)

Table 2.38: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE1, BO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.2 (0.19,0.2)	0.21 (0.21,0.22)	0.25 (0.25,0.26)	0.16 (0.16,0.16)	0.17 (0.17,0.18)	0.21 (0.2,0.21)	0.14 (0.14,0.14)	0.15 (0.15,0.15)	0.18 (0.18,0.18)
Strata	0.19 (0.19,0.19)	0.21 (0.2,0.21)	0.25 (0.24,0.25)	0.16 (0.15,0.16)	0.17 (0.16,0.17)	0.2 (0.2,0.2)	0.13 (0.13,0.14)	0.15 (0.14,0.15)	0.17 (0.17,0.18)
True	0.17 (0.17,0.17)	0.19 (0.18,0.2)	0.23 (0.23,0.24)	0.14 (0.14,0.14)	0.16 (0.15,0.16)	0.19 (0.18,0.19)	0.12 (0.12,0.12)	0.14 (0.13,0.14)	0.16 (0.16,0.17)
NKM	0.17 (0.17,0.17)	0.19 (0.18,0.2)	0.23 (0.23,0.24)	0.14 (0.14,0.14)	0.16 (0.15,0.16)	0.19 (0.19,0.19)	0.12 (0.12,0.12)	0.14 (0.13,0.14)	0.16 (0.16,0.17)
Standard	0.23 (0.23,0.24)	0.24 (0.24,0.25)	0.27 (0.27,0.28)	0.19 (0.19,0.19)	0.19 (0.19,0.2)	0.22 (0.22,0.22)	0.16 (0.16,0.17)	0.17 (0.17,0.17)	0.19 (0.19,0.2)
GEE True	0.17 (0.17,0.17)	0.19 (0.19,0.2)	0.24 (0.23,0.24)	0.14 (0.13,0.14)	0.15 (0.15,0.16)	0.19 (0.18,0.19)	0.12 (0.12,0.12)	0.13 (0.13,0.13)	0.16 (0.16,0.17)
GEE TRM	0.2 (0.19,0.2)	0.21 (0.21,0.22)	0.25 (0.25,0.25)	0.16 (0.16,0.16)	0.17 (0.17,0.18)	0.2 (0.2,0.21)	0.14 (0.13,0.14)	0.15 (0.14,0.15)	0.17 (0.17,0.18)

Table 2.39: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE0, BO, MS1}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.04 (0.04,0.05)	0.05 (0.05,0.05)	0.06 (0.06,0.07)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.03 (0.03,0.03)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
Strata	0.04 (0.04,0.05)	0.05 (0.05,0.05)	0.06 (0.05,0.07)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
True	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.06 (0.06,0.07)	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
NKM	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.06 (0.05,0.07)	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.05)
Standard	0.05 (0.04,0.05)	0.05 (0.05,0.06)	0.06 (0.05,0.07)	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
GEE True	0.05 (0.04,0.05)	0.05 (0.05,0.06)	0.06 (0.06,0.07)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.06)	0.03 (0.03,0.03)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
GEE TRM	0.05 (0.04,0.05)	0.05 (0.05,0.06)	0.06 (0.06,0.07)	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.05)

Table 2.40: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE0, BO, MS1} . Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)
Strata	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)
True	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.04,0.04)	0.02 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.03)
NKM	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.04,0.04)	0.02 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.03)
Standard	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.05)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)
GEE True	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.04,0.04)	0.02 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)
GEE TRM	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)

Table 2.41: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE1, BO, MS1} . Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.06)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.05)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)
Strata	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.06)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.05)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)
True	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.06)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.05)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)
NKM	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.06)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.05)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)
Standard	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.06 (0.05,0.06)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.05 (0.04,0.05)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)
GEE True	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.03 (0.03,0.04)	0.03 (0.03,0.04)	0.04 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)
GEE TRM	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.03 (0.03,0.04)	0.03 (0.03,0.04)	0.04 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)

Table 2.42: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE1, BO, MS1}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)
Strata	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)
True	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.02 (0.02,0.03)	0.03 (0.02,0.03)	0.03 (0.03,0.04)
NKM	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.02 (0.02,0.03)	0.03 (0.02,0.03)	0.03 (0.03,0.04)
Standard	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)
GEE True	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.04,0.04)	0.02 (0.02,0.03)	0.03 (0.02,0.03)	0.03 (0.03,0.03)
GEE TRM	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.04,0.04)	0.02 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.03)

2.7.5 Continuous Outcomes Model Specifications 2 and 3 Results

Table 2.43: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE0, CO, MS2}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$
TRM	97 (94,100)	97 (94,97)	96 (94,97)	97 (97,97)	96 (94,98)	96 (94,97)	96 (94,97)	95 (94,97)	96 (94,97)
Strata	96 (94,98)	95 (94,97)	95 (92,97)	96 (94,97)	95 (92,98)	94 (92,97)	94 (91,97)	94 (92,96)	94 (91,96)
True	96 (94,98)	95 (94,97)	95 (94,97)	96 (94,97)	95 (94,97)	95 (94,97)	94 (91,97)	95 (94,97)	94 (94,97)
NKM	96 (94,98)	95 (94,97)	95 (94,97)	96 (94,98)	95 (94,97)	95 (92,97)	94 (94,97)	94 (91,97)	94 (93,97)
Standard	97 (96,100)	98 (97,100)	97 (96,100)	97 (97,100)	97 (95,100)	97 (95,100)	97 (96,97)	97 (94,100)	97 (96,100)
GEE True	96 (94,97)	96 (95,96)	95 (94,96)	95 (93,96)	95 (94,96)	95 (94,97)	96 (94,97)	96 (94,98)	95 (94,97)
GEE TRM	97 (96,98)	96 (95,98)	96 (95,98)	95 (94,98)	96 (94,98)	97 (96,98)	97 (96,98)	97 (95,98)	96 (95,98)

Table 2.44: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE0, CO, MS2}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$
TRM	97 (94,98)	96 (94,97)	97 (96,99)	96 (94,97)	97 (96,100)	94 (91,97)	96 (95,97)	97 (96,99)	98 (97,100)
Strata	95 (94,97)	96 (94,97)	96 (96,98)	96 (95,97)	96 (94,100)	94 (90,97)	96 (94,97)	96 (94,98)	96 (97,97)
True	95 (93,97)	96 (94,97)	95 (94,96)	95 (94,97)	95 (93,97)	93 (90,95)	93 (89,97)	94 (93,96)	96 (95,97)
NKM	96 (94,97)	95 (94,97)	95 (94,97)	95 (93,97)	95 (93,97)	93 (93,95)	94 (92,95)	95 (94,97)	96 (96,97)
Standard	97 (97,98)	97 (97,99)	99 (98,100)	97 (97,98)	97 (96,100)	96 (94,100)	97 (96,98)	98 (97,99)	98 (97,100)
GEE True	95 (94,97)	95 (94,96)	96 (95,97)	95 (94,97)	95 (94,96)	96 (94,98)	96 (94,97)	95 (94,97)	95 (94,97)
GEE TRM	97 (96,98)	96 (95,98)	97 (96,98)	97 (96,98)	97 (96,98)	97 (96,98)	97 (96,98)	97 (96,98)	97 (97,98)

Table 2.45: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE1, CO, MS2}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 1$			$\beta_9 = 3$		
Model	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$
TRM	96 (94,99)	96 (94,97)	97 (95,100)	96 (94,100)	97 (95,100)	96 (94,97)
Strata	95 (94,97)	95 (94,97)	95 (94,97)	95 (94,97)	95 (94,97)	95 (91,97)
True	95 (94,97)	95 (94,97)	95 (94,97)	95 (93,97)	94 (92,97)	94 (91,97)
NKM	96 (94,97)	95 (94,97)	95 (94,97)	95 (94,97)	96 (94,97)	95 (92,97)
Standard	97 (96,100)	98 (97,100)	98 (97,100)	97 (95,100)	98 (97,100)	97 (96,99)
GEE True	96 (94,97)	95 (94,97)	95 (94,97)	95 (93,97)	95 (93,96)	94 (93,96)
GEE TRM	96 (95,98)	96 (95,98)	96 (95,98)	96 (95,98)	96 (95,98)	96 (95,98)

Table 2.46: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE1, CO, MS2}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 1$			$\beta_9 = 3$		
Model	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$
TRM	97 (97,99)	98 (97,100)	98 (97,99)	98 (97,99)	97 (96,98)	98 (97,99)
Strata	97 (95,99)	98 (97,100)	97 (97,98)	97 (97,98)	97 (96,98)	97 (96,98)
True	95 (93,97)	95 (94,97)	95 (93,97)	94 (91,96)	95 (92,96)	95 (93,96)
NKM	95 (94,97)	96 (94,98)	96 (94,97)	95 (94,98)	95 (94,96)	95 (93,97)
Standard	98 (96,100)	98 (97,100)	98 (97,100)	98 (97,99)	97 (97,99)	98 (97,100)
GEE True	96 (94,97)	95 (94,96)	95 (94,97)	95 (94,96)	95 (93,97)	94 (93,96)
GEE TRM	98 (97,99)	98 (97,99)	98 (97,99)	98 (97,99)	98 (97,99)	98 (98,99)

Table 2.47: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE0, CO, MS2}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$
TRM	2.86 (2.69,2.92)	2.39 (2.23,2.52)	2.24 (2.1,2.41)	2.37 (2.22,2.49)	1.91 (1.79,2.02)	1.79 (1.68,1.94)	1.98 (1.92,2.01)	1.64 (1.55,1.72)	1.55 (1.44,1.68)
Strata	2.72 (2.55,2.82)	2.2 (2.15,2.24)	2.05 (2.02,2.1)	2.26 (2.15,2.33)	1.76 (1.73,1.8)	1.64 (1.61,1.69)	1.89 (1.85,1.93)	1.51 (1.49,1.53)	1.41 (1.38,1.43)
True	2.68 (2.52,2.77)	2.16 (2.1,2.21)	2.01 (1.96,2.05)	2.24 (2.14,2.3)	1.74 (1.71,1.77)	1.62 (1.58,1.64)	1.86 (1.79,1.92)	1.48 (1.46,1.5)	1.38 (1.36,1.4)
NKM	2.69 (2.51,2.78)	2.16 (2.1,2.2)	2.01 (1.96,2.04)	2.24 (2.14,2.3)	1.74 (1.71,1.77)	1.61 (1.58,1.64)	1.86 (1.8,1.91)	1.48 (1.46,1.51)	1.38 (1.36,1.41)
Standard	3.05 (2.84,3.19)	2.63 (2.41,2.84)	2.5 (2.29,2.73)	2.53 (2.34,2.65)	2.1 (1.95,2.27)	2 (1.82,2.18)	2.14 (2.02,2.24)	1.79 (1.66,1.92)	1.71 (1.55,1.87)
GEE True	2.76 (2.68,2.84)	2.18 (2.15,2.21)	2.02 (1.99,2.05)	2.22 (2.18,2.27)	1.76 (1.74,1.78)	1.63 (1.61,1.66)	1.93 (1.9,1.96)	1.52 (1.5,1.53)	1.39 (1.37,1.4)
GEE TRM	2.9 (2.76,3.02)	2.36 (2.19,2.49)	2.22 (2.06,2.41)	2.34 (2.25,2.45)	1.93 (1.81,2.07)	1.81 (1.69,1.95)	2.03 (1.96,2.12)	1.66 (1.55,1.81)	1.54 (1.42,1.68)

Table 2.48: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE0, CO, MS2}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$
TRM	2.1 (2.06,2.15)	1.74 (1.71,1.77)	1.63 (1.61,1.67)	1.72 (1.68,1.75)	1.42 (1.39,1.46)	1.33 (1.31,1.35)	1.47 (1.44,1.49)	1.23 (1.21,1.25)	1.14 (1.12,1.16)
Strata	2.02 (1.97,2.09)	1.64 (1.63,1.67)	1.55 (1.53,1.56)	1.65 (1.62,1.69)	1.34 (1.32,1.35)	1.26 (1.25,1.27)	1.42 (1.39,1.45)	1.17 (1.16,1.17)	1.09 (1.07,1.1)
True	1.89 (1.82,1.95)	1.51 (1.48,1.52)	1.38 (1.37,1.4)	1.54 (1.52,1.56)	1.21 (1.2,1.23)	1.12 (1.12,1.13)	1.32 (1.3,1.35)	1.06 (1.05,1.06)	0.96 (0.95,0.97)
NKM	1.94 (1.89,2)	1.55 (1.53,1.57)	1.43 (1.41,1.46)	1.57 (1.54,1.62)	1.25 (1.24,1.27)	1.16 (1.16,1.17)	1.36 (1.34,1.38)	1.09 (1.08,1.09)	1 (1,1.01)
Standard	2.21 (2.19,2.24)	1.87 (1.8,1.94)	1.77 (1.72,1.81)	1.8 (1.76,1.84)	1.52 (1.47,1.58)	1.43 (1.39,1.46)	1.55 (1.5,1.58)	1.32 (1.28,1.36)	1.23 (1.2,1.24)
GEE True	1.96 (1.91,2.02)	1.55 (1.53,1.56)	1.42 (1.4,1.43)	1.58 (1.55,1.59)	1.23 (1.22,1.24)	1.14 (1.14,1.15)	1.36 (1.34,1.37)	1.06 (1.06,1.07)	0.98 (0.98,0.99)
GEE TRM	2.18 (2.12,2.22)	1.83 (1.73,1.94)	1.72 (1.62,1.84)	1.77 (1.72,1.84)	1.47 (1.4,1.56)	1.4 (1.33,1.5)	1.53 (1.48,1.57)	1.28 (1.22,1.35)	1.21 (1.14,1.29)

Table 2.49: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE1, CO, MS2}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$
TRM	2.91 (2.75,3.04)	2.44 (2.28,2.57)	2.27 (2.12,2.42)	2.35 (2.26,2.42)	1.94 (1.83,2.03)	1.82 (1.7,1.95)	2.02 (1.94,2.09)	1.66 (1.58,1.76)	1.55 (1.46,1.65)
Strata	2.75 (2.64,2.83)	2.26 (2.22,2.32)	2.1 (2.04,2.15)	2.22 (2.17,2.28)	1.8 (1.76,1.83)	1.68 (1.64,1.71)	1.91 (1.85,1.96)	1.54 (1.51,1.57)	1.43 (1.4,1.47)
True	2.7 (2.62,2.8)	2.18 (2.12,2.23)	2.02 (1.97,2.07)	2.18 (2.12,2.23)	1.73 (1.7,1.77)	1.62 (1.59,1.64)	1.86 (1.8,1.92)	1.48 (1.46,1.52)	1.37 (1.35,1.4)
NKM	2.74 (2.61,2.81)	2.22 (2.16,2.26)	2.06 (2.01,2.12)	2.21 (2.16,2.27)	1.78 (1.74,1.81)	1.65 (1.62,1.68)	1.9 (1.84,1.93)	1.52 (1.49,1.56)	1.41 (1.39,1.44)
Standard	3.2 (3.01,3.34)	2.79 (2.59,2.98)	2.64 (2.46,2.83)	2.59 (2.47,2.7)	2.22 (2.09,2.37)	2.11 (1.98,2.25)	2.22 (2.11,2.34)	1.92 (1.81,2.06)	1.81 (1.69,1.94)
GEE True	2.76 (2.68,2.82)	2.19 (2.15,2.23)	2.04 (2.01,2.06)	2.23 (2.19,2.27)	1.77 (1.74,1.79)	1.63 (1.61,1.65)	1.92 (1.87,1.97)	1.51 (1.49,1.52)	1.4 (1.39,1.41)
GEE TRM	3.01 (2.87,3.14)	2.5 (2.4,2.6)	2.38 (2.26,2.51)	2.46 (2.35,2.56)	2.03 (1.94,2.13)	1.91 (1.82,2.01)	2.11 (2.03,2.19)	1.74 (1.67,1.82)	1.65 (1.55,1.74)

Table 2.50: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE1, CO, MS2}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$
TRM	2.75 (2.66,2.81)	2.44 (2.36,2.52)	2.38 (2.29,2.47)	2.24 (2.14,2.31)	2.01 (1.93,2.08)	1.93 (1.85,2.02)	1.95 (1.88,2.02)	1.74 (1.67,1.81)	1.68 (1.61,1.75)
Strata	2.62 (2.54,2.69)	2.33 (2.26,2.4)	2.28 (2.19,2.37)	2.15 (2.09,2.22)	1.92 (1.85,1.99)	1.85 (1.77,1.91)	1.88 (1.82,1.92)	1.66 (1.6,1.73)	1.6 (1.55,1.66)
True	1.92 (1.89,1.95)	1.5 (1.46,1.53)	1.4 (1.39,1.42)	1.55 (1.52,1.59)	1.22 (1.21,1.23)	1.12 (1.11,1.13)	1.34 (1.31,1.35)	1.05 (1.04,1.07)	0.97 (0.97,0.98)
NKM	2.17 (2.13,2.19)	1.8 (1.78,1.85)	1.72 (1.68,1.75)	1.76 (1.74,1.8)	1.48 (1.46,1.5)	1.39 (1.38,1.41)	1.54 (1.51,1.56)	1.27 (1.25,1.29)	1.2 (1.2,1.22)
Standard	2.82 (2.73,2.89)	2.53 (2.46,2.59)	2.46 (2.38,2.54)	2.32 (2.2,2.39)	2.07 (2.01,2.14)	2 (1.93,2.08)	2.01 (1.95,2.06)	1.79 (1.72,1.86)	1.73 (1.67,1.8)
GEE True	1.97 (1.93,1.99)	1.54 (1.52,1.56)	1.42 (1.41,1.44)	1.58 (1.55,1.59)	1.24 (1.23,1.25)	1.14 (1.13,1.15)	1.38 (1.35,1.4)	1.07 (1.06,1.07)	0.98 (0.97,0.99)
GEE TRM	2.84 (2.75,2.92)	2.54 (2.41,2.63)	2.47 (2.37,2.55)	2.29 (2.21,2.37)	2.06 (1.98,2.14)	2.01 (1.92,2.07)	1.99 (1.94,2.06)	1.78 (1.71,1.84)	1.73 (1.65,1.8)

Table 2.51: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE0, CO, MS2}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$
TRM	0.54 (0.5,0.58)	0.44 (0.42,0.48)	0.39 (0.37,0.41)	0.46 (0.41,0.5)	0.34 (0.31,0.37)	0.32 (0.3,0.33)	0.41 (0.37,0.44)	0.32 (0.29,0.33)	0.28 (0.26,0.3)
Strata	0.53 (0.49,0.58)	0.44 (0.42,0.47)	0.39 (0.37,0.4)	0.46 (0.42,0.49)	0.34 (0.32,0.37)	0.32 (0.3,0.34)	0.4 (0.37,0.43)	0.32 (0.29,0.35)	0.28 (0.26,0.29)
True	0.53 (0.49,0.57)	0.43 (0.4,0.46)	0.39 (0.36,0.41)	0.46 (0.42,0.48)	0.34 (0.32,0.36)	0.31 (0.29,0.33)	0.4 (0.36,0.42)	0.31 (0.28,0.33)	0.27 (0.26,0.29)
NKM	0.53 (0.49,0.57)	0.43 (0.4,0.46)	0.39 (0.36,0.41)	0.46 (0.42,0.48)	0.34 (0.32,0.36)	0.31 (0.28,0.33)	0.39 (0.36,0.42)	0.31 (0.29,0.33)	0.27 (0.26,0.29)
Standard	0.55 (0.5,0.6)	0.45 (0.42,0.49)	0.4 (0.38,0.41)	0.47 (0.42,0.5)	0.35 (0.34,0.37)	0.33 (0.31,0.35)	0.42 (0.39,0.46)	0.33 (0.3,0.36)	0.29 (0.27,0.31)
GEE True	0.54 (0.53,0.56)	0.42 (0.4,0.45)	0.39 (0.38,0.41)	0.45 (0.42,0.49)	0.36 (0.34,0.38)	0.32 (0.31,0.33)	0.39 (0.37,0.41)	0.3 (0.29,0.31)	0.28 (0.26,0.29)
GEE TRM	0.54 (0.52,0.55)	0.43 (0.41,0.45)	0.4 (0.39,0.41)	0.46 (0.42,0.49)	0.36 (0.34,0.39)	0.33 (0.32,0.34)	0.39 (0.38,0.4)	0.3 (0.29,0.31)	0.28 (0.27,0.3)

Table 2.52: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE0, CO, MS2}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$
TRM	0.41 (0.4,0.44)	0.32 (0.3,0.33)	0.31 (0.29,0.32)	0.33 (0.3,0.35)	0.26 (0.23,0.28)	0.26 (0.24,0.28)	0.29 (0.27,0.31)	0.23 (0.21,0.24)	0.2 (0.19,0.2)
Strata	0.41 (0.39,0.44)	0.31 (0.3,0.33)	0.31 (0.29,0.32)	0.33 (0.3,0.35)	0.26 (0.23,0.28)	0.26 (0.24,0.28)	0.29 (0.27,0.31)	0.23 (0.21,0.24)	0.2 (0.19,0.21)
True	0.4 (0.38,0.41)	0.29 (0.26,0.31)	0.29 (0.26,0.3)	0.31 (0.3,0.33)	0.25 (0.24,0.26)	0.24 (0.21,0.25)	0.28 (0.26,0.3)	0.22 (0.2,0.23)	0.18 (0.17,0.2)
NKM	0.41 (0.38,0.43)	0.3 (0.27,0.33)	0.29 (0.27,0.31)	0.32 (0.3,0.34)	0.25 (0.24,0.27)	0.25 (0.23,0.26)	0.29 (0.26,0.3)	0.22 (0.21,0.22)	0.19 (0.18,0.2)
Standard	0.41 (0.4,0.44)	0.32 (0.3,0.34)	0.31 (0.29,0.32)	0.33 (0.3,0.36)	0.26 (0.24,0.28)	0.26 (0.24,0.28)	0.3 (0.27,0.32)	0.23 (0.22,0.24)	0.2 (0.19,0.22)
GEE True	0.4 (0.38,0.42)	0.3 (0.28,0.32)	0.28 (0.27,0.29)	0.32 (0.31,0.34)	0.25 (0.24,0.26)	0.23 (0.21,0.24)	0.27 (0.26,0.29)	0.21 (0.2,0.22)	0.2 (0.19,0.21)
GEE TRM	0.42 (0.4,0.44)	0.33 (0.31,0.35)	0.3 (0.29,0.32)	0.34 (0.32,0.36)	0.27 (0.25,0.28)	0.25 (0.24,0.26)	0.29 (0.27,0.3)	0.23 (0.22,0.24)	0.21 (0.2,0.22)

Table 2.53: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE1, CO, MS2}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$
TRM	0.55 (0.51,0.6)	0.43 (0.4,0.47)	0.4 (0.37,0.43)	0.45 (0.41,0.48)	0.36 (0.32,0.4)	0.33 (0.31,0.36)	0.39 (0.36,0.43)	0.31 (0.28,0.32)	0.29 (0.27,0.31)
Strata	0.55 (0.51,0.59)	0.43 (0.41,0.47)	0.41 (0.37,0.45)	0.45 (0.4,0.49)	0.36 (0.32,0.39)	0.33 (0.3,0.35)	0.39 (0.36,0.43)	0.3 (0.28,0.32)	0.29 (0.27,0.31)
True	0.54 (0.51,0.58)	0.43 (0.39,0.46)	0.4 (0.36,0.44)	0.44 (0.41,0.48)	0.35 (0.33,0.38)	0.32 (0.3,0.35)	0.39 (0.35,0.42)	0.3 (0.28,0.33)	0.28 (0.26,0.3)
NKM	0.54 (0.51,0.58)	0.43 (0.4,0.45)	0.4 (0.36,0.43)	0.44 (0.4,0.47)	0.36 (0.32,0.39)	0.32 (0.3,0.34)	0.39 (0.34,0.41)	0.3 (0.28,0.32)	0.28 (0.26,0.31)
Standard	0.57 (0.54,0.61)	0.47 (0.43,0.51)	0.44 (0.41,0.48)	0.47 (0.44,0.5)	0.39 (0.35,0.43)	0.36 (0.33,0.39)	0.41 (0.37,0.44)	0.32 (0.29,0.35)	0.31 (0.28,0.34)
GEE True	0.56 (0.53,0.59)	0.43 (0.41,0.44)	0.4 (0.38,0.42)	0.45 (0.43,0.47)	0.35 (0.34,0.37)	0.32 (0.31,0.34)	0.39 (0.37,0.41)	0.3 (0.28,0.32)	0.29 (0.26,0.3)
GEE TRM	0.58 (0.55,0.61)	0.46 (0.44,0.49)	0.43 (0.41,0.45)	0.47 (0.45,0.49)	0.38 (0.36,0.39)	0.35 (0.33,0.37)	0.41 (0.39,0.44)	0.32 (0.31,0.34)	0.31 (0.29,0.32)

Table 2.54: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE1, CO, MS2}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$	$\nu = 3$	$\nu = 5$	$\nu = 7$
TRM	0.49 (0.46,0.52)	0.4 (0.36,0.43)	0.4 (0.37,0.43)	0.39 (0.37,0.41)	0.35 (0.33,0.37)	0.33 (0.31,0.34)	0.35 (0.32,0.38)	0.3 (0.28,0.32)	0.28 (0.26,0.3)
Strata	0.49 (0.47,0.52)	0.4 (0.36,0.43)	0.4 (0.37,0.43)	0.39 (0.37,0.41)	0.35 (0.33,0.37)	0.33 (0.31,0.34)	0.35 (0.32,0.38)	0.3 (0.28,0.32)	0.28 (0.27,0.3)
True	0.4 (0.38,0.43)	0.3 (0.28,0.33)	0.27 (0.26,0.29)	0.32 (0.31,0.34)	0.24 (0.22,0.26)	0.23 (0.22,0.24)	0.29 (0.27,0.3)	0.22 (0.21,0.23)	0.2 (0.2,0.22)
NKM	0.44 (0.41,0.49)	0.35 (0.34,0.38)	0.34 (0.31,0.36)	0.35 (0.34,0.37)	0.3 (0.29,0.31)	0.28 (0.27,0.3)	0.32 (0.29,0.33)	0.25 (0.24,0.26)	0.24 (0.23,0.25)
Standard	0.5 (0.47,0.52)	0.4 (0.37,0.43)	0.4 (0.37,0.43)	0.39 (0.37,0.42)	0.35 (0.33,0.37)	0.33 (0.32,0.34)	0.36 (0.33,0.38)	0.3 (0.28,0.32)	0.29 (0.27,0.3)
GEE True	0.4 (0.38,0.42)	0.31 (0.29,0.33)	0.29 (0.27,0.3)	0.32 (0.31,0.34)	0.25 (0.23,0.27)	0.23 (0.22,0.25)	0.28 (0.26,0.29)	0.21 (0.2,0.23)	0.2 (0.19,0.21)
GEE TRM	0.5 (0.48,0.53)	0.43 (0.4,0.45)	0.41 (0.39,0.44)	0.4 (0.38,0.42)	0.35 (0.33,0.37)	0.34 (0.32,0.36)	0.35 (0.33,0.37)	0.3 (0.28,0.32)	0.29 (0.27,0.3)

Table 2.55: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE0, CO, MS3}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$
TRM	99 (97,100)	95 (92,97)	93 (91,97)	98 (97,100)	95 (94,97)	95 (91,97)	98 (97,100)	94 (91,97)	92 (91,97)
Strata	96 (94,98)	94 (91,97)	94 (91,97)	95 (94,97)	94 (92,96)	96 (94,97)	96 (94,97)	94 (91,97)	92 (88,95)
True	96 (94,99)	94 (94,97)	95 (92,97)	95 (92,97)	94 (92,97)	95 (94,98)	95 (94,97)	94 (92,97)	92 (89,97)
NKM	96 (94,99)	94 (91,97)	94 (91,97)	94 (92,97)	95 (93,97)	96 (94,98)	96 (94,97)	94 (92,97)	91 (88,97)
Standard	99 (99,100)	94 (91,97)	95 (92,97)	99 (98,100)	94 (93,97)	95 (94,97)	99 (100,100)	94 (91,97)	92 (91,94)
GEE True	96 (95,98)	95 (94,96)	95 (94,97)	95 (94,96)	96 (94,98)	95 (94,96)	96 (94,96)	95 (94,96)	96 (94,97)
GEE TRM	98 (98,100)	95 (93,97)	95 (93,97)	98 (97,99)	96 (94,97)	95 (93,96)	98 (96,99)	95 (93,97)	96 (94,97)

Table 2.56: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE0, CO, MS3}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$
TRM	98 (97,99)	94 (92,97)	93 (91,94)	98 (97,100)	94 (91,96)	93 (88,96)	98 (96,100)	95 (94,97)	93 (91,95)
Strata	97 (96,97)	95 (93,97)	94 (94,94)	96 (95,98)	93 (91,95)	93 (91,95)	96 (94,100)	93 (91,95)	94 (91,96)
True	95 (94,97)	93 (91,96)	93 (93,95)	96 (94,97)	92 (89,94)	90 (87,93)	94 (91,98)	94 (94,96)	90 (88,93)
NKM	95 (94,97)	94 (92,96)	93 (91,95)	95 (93,97)	93 (91,95)	90 (87,95)	95 (93,97)	93 (91,97)	92 (88,95)
Standard	100 (99,100)	95 (94,97)	93 (91,94)	99 (98,100)	95 (91,97)	93 (92,97)	99 (98,100)	95 (94,97)	94 (94,95)
GEE True	95 (94,97)	95 (94,96)	96 (95,97)	95 (94,96)	95 (94,97)	96 (94,98)	95 (94,97)	95 (94,97)	95 (94,96)
GEE TRM	99 (98,100)	95 (95,96)	95 (94,96)	98 (98,99)	95 (94,97)	96 (95,97)	98 (97,100)	95 (94,97)	95 (94,96)

Table 2.57: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE1, CO, MS3}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 1$			$\beta_9 = 3$		
Model	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$
TRM	98 (97,100)	94 (91,97)	92 (91,96)	98 (96,100)	95 (93,97)	94 (92,97)
Strata	95 (94,97)	94 (94,97)	93 (91,97)	95 (92,97)	94 (91,97)	94 (94,97)
True	95 (94,97)	94 (91,97)	94 (91,97)	92 (90,97)	94 (93,97)	94 (93,97)
NKM	96 (94,97)	94 (91,97)	93 (91,97)	95 (94,97)	94 (91,97)	94 (92,97)
Standard	99 (98,100)	95 (94,97)	94 (91,97)	98 (97,100)	95 (94,97)	94 (94,97)
GEE True	95 (94,97)	95 (94,96)	95 (93,97)	93 (91,95)	95 (94,97)	95 (94,96)
GEE TRM	98 (96,99)	95 (94,97)	95 (94,97)	97 (96,99)	95 (94,96)	95 (93,96)

Table 2.58: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE1, CO, MS3}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 1$			$\beta_9 = 3$		
Model	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$
TRM	98 (97,100)	96 (94,97)	96 (94,97)	99 (98,100)	97 (96,98)	96 (95,97)
Strata	97 (96,99)	96 (94,97)	96 (95,97)	98 (97,99)	96 (96,98)	95 (94,97)
True	95 (94,97)	94 (92,96)	93 (91,96)	93 (92,96)	93 (91,96)	93 (91,97)
NKM	95 (93,97)	95 (94,97)	95 (94,97)	96 (94,97)	96 (94,98)	95 (94,97)
Standard	99 (97,100)	96 (94,98)	96 (94,97)	99 (99,100)	97 (96,99)	95 (94,97)
GEE True	95 (94,97)	96 (94,97)	96 (95,97)	93 (92,94)	95 (93,97)	95 (94,96)
GEE TRM	99 (98,100)	96 (95,98)	96 (95,97)	99 (98,100)	97 (96,98)	96 (95,97)

Table 2.59: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE0, CO, MS3}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$
TRM	1.61 (1.45,1.9)	4.64 (4.53,4.75)	6.26 (6.13,6.38)	1.31 (1.18,1.5)	3.77 (3.68,3.85)	5.05 (4.98,5.17)	1.09 (0.92,1.27)	3.18 (3.1,3.27)	4.27 (4.11,4.48)
Strata	1.29 (1.23,1.33)	4.6 (4.5,4.65)	6.27 (6.15,6.38)	1.06 (1,1.11)	3.72 (3.66,3.81)	5.07 (4.97,5.18)	0.88 (0.85,0.91)	3.13 (3.05,3.24)	4.26 (4.11,4.45)
True	1.21 (1.2,1.23)	4.57 (4.47,4.66)	6.3 (6.22,6.4)	1 (0.97,1.03)	3.72 (3.62,3.79)	5.06 (4.95,5.18)	0.84 (0.82,0.86)	3.14 (3.06,3.21)	4.31 (4.14,4.48)
NKM	1.21 (1.19,1.25)	4.56 (4.46,4.65)	6.28 (6.17,6.41)	0.99 (0.95,1.03)	3.69 (3.61,3.74)	5.04 (4.94,5.11)	0.84 (0.82,0.86)	3.14 (3.05,3.23)	4.27 (4.14,4.44)
Standard	1.99 (1.82,2.29)	4.74 (4.61,4.9)	6.32 (6.26,6.43)	1.6 (1.45,1.84)	3.8 (3.73,3.85)	5.06 (4.95,5.19)	1.33 (1.1,1.55)	3.22 (3.17,3.3)	4.26 (4.1,4.47)
GEE True	1.23 (1.2,1.26)	4.79 (4.73,4.84)	6.8 (6.74,6.89)	0.98 (0.96,1)	3.83 (3.78,3.87)	5.44 (5.39,5.5)	0.84 (0.83,0.85)	3.3 (3.27,3.34)	4.67 (4.61,4.71)
GEE TRM	1.57 (1.3,1.85)	4.78 (4.68,4.88)	6.7 (6.61,6.79)	1.26 (1.06,1.48)	3.85 (3.77,3.92)	5.41 (5.37,5.46)	1.08 (0.9,1.28)	3.34 (3.31,3.39)	4.66 (4.63,4.69)

Table 2.60: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE0, CO, MS3}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$
TRM	1.25 (1.19,1.3)	3.31 (3.28,3.35)	4.4 (4.32,4.5)	1.01 (0.95,1.06)	2.69 (2.66,2.71)	3.61 (3.54,3.67)	0.86 (0.83,0.9)	2.34 (2.31,2.35)	3.13 (3.11,3.19)
Strata	1.1 (1.09,1.12)	3.27 (3.24,3.31)	4.38 (4.26,4.52)	0.89 (0.88,0.9)	2.66 (2.63,2.68)	3.61 (3.54,3.68)	0.77 (0.75,0.78)	2.3 (2.28,2.32)	3.17 (3.15,3.19)
True	0.85 (0.84,0.86)	3.16 (3.11,3.22)	4.35 (4.26,4.49)	0.69 (0.67,0.7)	2.52 (2.45,2.6)	3.36 (3.18,3.53)	0.59 (0.59,0.6)	2.21 (2.18,2.25)	3.03 (2.93,3.1)
NKM	0.92 (0.91,0.93)	3.21 (3.17,3.25)	4.36 (4.26,4.45)	0.75 (0.73,0.76)	2.62 (2.58,2.67)	3.49 (3.36,3.62)	0.65 (0.64,0.65)	2.25 (2.22,2.32)	3.11 (3.05,3.16)
Standard	1.44 (1.34,1.55)	3.37 (3.32,3.41)	4.45 (4.38,4.52)	1.16 (1.07,1.25)	2.74 (2.7,2.76)	3.65 (3.58,3.73)	0.99 (0.93,1.06)	2.38 (2.36,2.41)	3.17 (3.13,3.22)
GEE True	0.86 (0.84,0.87)	3.34 (3.31,3.37)	4.73 (4.69,4.77)	0.7 (0.69,0.7)	2.69 (2.68,2.7)	3.81 (3.79,3.83)	0.6 (0.59,0.61)	2.32 (2.31,2.33)	3.28 (3.26,3.29)
GEE TRM	1.32 (1.19,1.47)	3.41 (3.36,3.46)	4.72 (4.68,4.77)	1.07 (0.97,1.18)	2.77 (2.73,2.81)	3.84 (3.81,3.85)	0.92 (0.84,1.02)	2.4 (2.38,2.42)	3.31 (3.28,3.34)

Table 2.61: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE1, CO, MS3}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$
TRM	1.64 (1.42,1.89)	4.68 (4.58,4.76)	6.24 (6.17,6.36)	1.32 (1.15,1.49)	3.75 (3.69,3.82)	5.13 (5.08,5.22)	1.12 (0.99,1.29)	3.2 (3.15,3.27)	4.28 (4.17,4.45)
Strata	1.35 (1.3,1.41)	4.61 (4.54,4.68)	6.26 (6.15,6.35)	1.09 (1.04,1.15)	3.72 (3.69,3.76)	5.11 (5.02,5.19)	0.94 (0.89,0.96)	3.16 (3.09,3.23)	4.31 (4.23,4.43)
True	1.21 (1.18,1.25)	4.58 (4.48,4.67)	6.2 (6.11,6.31)	0.98 (0.94,1.01)	3.7 (3.64,3.75)	5.11 (5,5.21)	0.85 (0.83,0.87)	3.11 (3.02,3.22)	4.26 (4.16,4.38)
NKM	1.28 (1.24,1.32)	4.6 (4.52,4.68)	6.25 (6.16,6.35)	1.04 (0.99,1.09)	3.72 (3.65,3.77)	5.12 (5.02,5.22)	0.9 (0.86,0.93)	3.16 (3.08,3.24)	4.32 (4.18,4.46)
Standard	2.15 (1.91,2.42)	4.83 (4.71,4.93)	6.38 (6.25,6.51)	1.73 (1.52,1.9)	3.9 (3.83,3.96)	5.24 (5.13,5.33)	1.47 (1.33,1.64)	3.38 (3.32,3.43)	4.46 (4.35,4.55)
GEE True	1.21 (1.18,1.23)	4.81 (4.76,4.86)	6.8 (6.74,6.86)	0.98 (0.96,1)	3.84 (3.8,3.87)	5.44 (5.41,5.49)	0.84 (0.83,0.86)	3.3 (3.28,3.32)	4.65 (4.62,4.69)
GEE TRM	1.74 (1.58,1.9)	4.88 (4.8,4.96)	6.75 (6.69,6.83)	1.4 (1.29,1.55)	3.92 (3.87,3.98)	5.46 (5.41,5.53)	1.21 (1.11,1.33)	3.39 (3.35,3.42)	4.68 (4.65,4.71)

Table 2.62: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE1, CO, MS3}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$
TRM	2.14 (2.01,2.26)	3.77 (3.72,3.83)	4.79 (4.71,4.85)	1.74 (1.65,1.83)	3.06 (3.02,3.1)	3.99 (3.93,4.07)	1.49 (1.41,1.57)	2.65 (2.59,2.7)	3.44 (3.42,3.47)
Strata	2.01 (1.9,2.14)	3.69 (3.61,3.74)	4.77 (4.72,4.8)	1.63 (1.54,1.72)	3.01 (2.95,3.05)	3.94 (3.87,3.98)	1.4 (1.34,1.48)	2.6 (2.55,2.64)	3.41 (3.39,3.43)
True	0.84 (0.83,0.86)	3.22 (3.18,3.29)	4.43 (4.42,4.48)	0.69 (0.68,0.7)	2.57 (2.52,2.64)	3.56 (3.47,3.65)	0.59 (0.58,0.6)	2.2 (2.17,2.24)	2.94 (2.69,3.13)
NKM	1.31 (1.27,1.36)	3.41 (3.36,3.44)	4.55 (4.53,4.59)	1.07 (1.06,1.09)	2.76 (2.73,2.78)	3.77 (3.74,3.78)	0.92 (0.9,0.94)	2.37 (2.36,2.39)	3.21 (3.16,3.25)
Standard	2.24 (2.11,2.34)	3.83 (3.77,3.88)	4.83 (4.79,4.88)	1.84 (1.73,1.9)	3.11 (3.07,3.16)	4.02 (3.96,4.07)	1.57 (1.49,1.64)	2.68 (2.64,2.74)	3.45 (3.44,3.48)
GEE True	0.86 (0.85,0.87)	3.35 (3.32,3.39)	4.73 (4.67,4.76)	0.7 (0.69,0.7)	2.69 (2.67,2.7)	3.81 (3.79,3.83)	0.6 (0.59,0.6)	2.32 (2.3,2.34)	3.27 (3.25,3.29)
GEE TRM	2.2 (2.1,2.29)	3.86 (3.8,3.92)	5.04 (4.99,5.1)	1.8 (1.69,1.88)	3.13 (3.07,3.17)	4.11 (4.07,4.14)	1.55 (1.46,1.61)	2.71 (2.66,2.76)	3.54 (3.51,3.56)

Table 2.63: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE0, CO, MS3}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$
TRM	0.24 (0.22,0.26)	0.9 (0.84,0.95)	1.17 (1.06,1.26)	0.21 (0.19,0.23)	0.76 (0.69,0.82)	1.06 (1,1.13)	0.18 (0.17,0.19)	0.66 (0.6,0.7)	0.93 (0.84,1.03)
Strata	0.24 (0.22,0.26)	0.89 (0.83,0.94)	1.17 (1.06,1.28)	0.21 (0.19,0.23)	0.75 (0.68,0.82)	1.05 (0.95,1.12)	0.18 (0.16,0.19)	0.64 (0.6,0.66)	0.88 (0.81,0.97)
True	0.23 (0.21,0.25)	0.9 (0.84,0.94)	1.17 (1.09,1.3)	0.2 (0.19,0.21)	0.75 (0.67,0.82)	1.06 (0.98,1.13)	0.17 (0.16,0.18)	0.64 (0.61,0.68)	0.91 (0.85,0.97)
NKM	0.23 (0.21,0.24)	0.89 (0.83,0.94)	1.16 (1.07,1.3)	0.2 (0.19,0.21)	0.76 (0.7,0.81)	1.04 (0.97,1.1)	0.17 (0.16,0.18)	0.65 (0.61,0.67)	0.89 (0.8,0.97)
Standard	0.26 (0.23,0.29)	0.9 (0.83,0.96)	1.17 (1.05,1.28)	0.23 (0.2,0.25)	0.76 (0.66,0.84)	1.07 (0.99,1.15)	0.19 (0.17,0.2)	0.65 (0.6,0.68)	0.9 (0.85,0.98)
GEE True	0.24 (0.23,0.25)	0.94 (0.88,0.99)	1.31 (1.26,1.35)	0.2 (0.19,0.22)	0.75 (0.7,0.78)	1.05 (0.98,1.08)	0.17 (0.16,0.18)	0.67 (0.63,0.7)	0.96 (0.92,1.01)
GEE TRM	0.26 (0.23,0.27)	0.93 (0.88,0.98)	1.3 (1.25,1.35)	0.21 (0.2,0.23)	0.75 (0.71,0.77)	1.05 (0.99,1.08)	0.18 (0.17,0.19)	0.67 (0.62,0.7)	0.96 (0.92,1.02)

Table 2.64: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE0, CO, MS3}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$
TRM	0.21 (0.19,0.22)	0.67 (0.63,0.72)	0.91 (0.87,0.99)	0.17 (0.16,0.18)	0.55 (0.52,0.61)	0.76 (0.72,0.82)	0.15 (0.14,0.16)	0.49 (0.47,0.51)	0.66 (0.63,0.69)
Strata	0.2 (0.19,0.22)	0.67 (0.64,0.71)	0.9 (0.87,0.96)	0.17 (0.15,0.18)	0.56 (0.52,0.6)	0.76 (0.73,0.8)	0.15 (0.14,0.16)	0.49 (0.47,0.51)	0.64 (0.58,0.69)
True	0.17 (0.16,0.17)	0.67 (0.63,0.72)	0.9 (0.83,0.98)	0.14 (0.13,0.14)	0.57 (0.53,0.61)	0.78 (0.71,0.84)	0.12 (0.11,0.13)	0.49 (0.44,0.53)	0.66 (0.58,0.7)
NKM	0.19 (0.18,0.19)	0.67 (0.64,0.71)	0.9 (0.83,0.96)	0.15 (0.14,0.17)	0.55 (0.51,0.59)	0.8 (0.74,0.88)	0.13 (0.12,0.15)	0.48 (0.44,0.52)	0.65 (0.58,0.73)
Standard	0.21 (0.2,0.23)	0.67 (0.65,0.71)	0.92 (0.87,0.97)	0.18 (0.17,0.18)	0.54 (0.52,0.6)	0.77 (0.74,0.8)	0.16 (0.14,0.17)	0.48 (0.46,0.52)	0.65 (0.6,0.71)
GEE True	0.17 (0.16,0.18)	0.66 (0.62,0.7)	0.93 (0.86,0.99)	0.14 (0.14,0.15)	0.54 (0.52,0.57)	0.75 (0.71,0.78)	0.12 (0.12,0.13)	0.47 (0.45,0.5)	0.66 (0.63,0.68)
GEE TRM	0.22 (0.21,0.23)	0.67 (0.63,0.71)	0.93 (0.87,1.01)	0.17 (0.16,0.18)	0.55 (0.53,0.58)	0.75 (0.72,0.79)	0.15 (0.14,0.16)	0.47 (0.46,0.5)	0.66 (0.63,0.68)

Table 2.65: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE1, CO, MS3}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$
TRM	0.27 (0.25,0.29)	0.92 (0.84,1)	1.27 (1.17,1.35)	0.22 (0.2,0.23)	0.77 (0.72,0.79)	1.03 (0.94,1.11)	0.18 (0.17,0.2)	0.68 (0.63,0.71)	0.96 (0.88,1.04)
Strata	0.27 (0.25,0.29)	0.91 (0.82,0.98)	1.26 (1.15,1.37)	0.22 (0.2,0.23)	0.75 (0.7,0.77)	1 (0.91,1.08)	0.18 (0.17,0.2)	0.68 (0.64,0.71)	0.94 (0.84,1.01)
True	0.26 (0.24,0.27)	0.9 (0.81,0.99)	1.22 (1.12,1.31)	0.21 (0.19,0.22)	0.74 (0.68,0.78)	0.99 (0.9,1.08)	0.18 (0.16,0.19)	0.67 (0.62,0.71)	0.95 (0.87,1.03)
NKM	0.26 (0.24,0.28)	0.91 (0.81,0.96)	1.27 (1.15,1.35)	0.21 (0.19,0.22)	0.75 (0.69,0.79)	1.01 (0.93,1.06)	0.18 (0.16,0.19)	0.68 (0.63,0.71)	0.95 (0.87,1.05)
Standard	0.32 (0.29,0.34)	0.93 (0.84,1.01)	1.28 (1.18,1.36)	0.26 (0.24,0.29)	0.77 (0.72,0.8)	1.02 (0.95,1.09)	0.22 (0.19,0.25)	0.68 (0.63,0.73)	0.95 (0.86,1.02)
GEE True	0.26 (0.24,0.27)	0.93 (0.89,0.97)	1.31 (1.23,1.38)	0.21 (0.19,0.22)	0.76 (0.73,0.79)	1.08 (1,1.14)	0.18 (0.17,0.19)	0.65 (0.62,0.68)	0.94 (0.9,0.99)
GEE TRM	0.3 (0.28,0.33)	0.94 (0.91,0.98)	1.32 (1.24,1.39)	0.25 (0.23,0.26)	0.77 (0.73,0.81)	1.09 (1.01,1.15)	0.21 (0.2,0.23)	0.66 (0.63,0.68)	0.95 (0.91,0.99)

Table 2.66: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE1, CO, MS3}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$	$\alpha = 0.5$	$\alpha = 8$	$\alpha = 16$
TRM	0.35 (0.33,0.37)	0.68 (0.63,0.73)	0.95 (0.91,1.01)	0.28 (0.27,0.3)	0.58 (0.55,0.6)	0.82 (0.76,0.87)	0.25 (0.23,0.26)	0.51 (0.47,0.53)	0.66 (0.63,0.67)
Strata	0.35 (0.33,0.37)	0.68 (0.63,0.72)	0.95 (0.91,1)	0.28 (0.26,0.3)	0.58 (0.54,0.6)	0.82 (0.75,0.87)	0.25 (0.23,0.26)	0.5 (0.47,0.52)	0.66 (0.63,0.67)
True	0.18 (0.16,0.18)	0.63 (0.59,0.69)	0.89 (0.83,0.93)	0.15 (0.15,0.16)	0.54 (0.5,0.58)	0.78 (0.73,0.82)	0.12 (0.11,0.13)	0.47 (0.44,0.48)	0.67 (0.59,0.76)
NKM	0.27 (0.25,0.28)	0.65 (0.6,0.7)	0.93 (0.87,0.96)	0.22 (0.21,0.24)	0.55 (0.52,0.57)	0.8 (0.72,0.86)	0.19 (0.18,0.2)	0.49 (0.47,0.51)	0.66 (0.62,0.69)
Standard	0.35 (0.33,0.38)	0.68 (0.64,0.73)	0.95 (0.9,0.99)	0.29 (0.27,0.31)	0.58 (0.54,0.6)	0.82 (0.75,0.93)	0.25 (0.23,0.27)	0.51 (0.47,0.53)	0.66 (0.64,0.69)
GEE True	0.18 (0.17,0.19)	0.65 (0.62,0.69)	0.93 (0.88,0.98)	0.15 (0.14,0.16)	0.53 (0.5,0.57)	0.77 (0.74,0.82)	0.13 (0.12,0.13)	0.46 (0.44,0.48)	0.65 (0.61,0.68)
GEE TRM	0.35 (0.32,0.37)	0.72 (0.69,0.76)	0.97 (0.92,1.02)	0.28 (0.27,0.3)	0.58 (0.55,0.62)	0.8 (0.76,0.84)	0.24 (0.23,0.26)	0.51 (0.48,0.54)	0.67 (0.64,0.71)

2.7.6 Binary Outcomes Model Specifications 2 and 3 Results

Table 2.67: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE0, BO, MS2}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$
TRM	96 (95,97)	95 (91,99)	96 (94,100)	96 (94,97)	96 (94,100)	95 (94,97)	95 (91,97)	96 (94,100)	96 (95,97)
Strata	95 (93,97)	95 (94,97)	95 (94,97)	95 (94,97)	96 (94,97)	94 (94,97)	94 (92,97)	96 (94,97)	96 (94,97)
True	94 (93,97)	95 (94,97)	96 (94,97)	95 (94,97)	96 (94,100)	95 (94,97)	95 (94,97)	96 (93,100)	95 (94,97)
NKM	95 (93,97)	95 (94,97)	95 (93,97)	95 (94,97)	95 (94,97)	95 (94,97)	95 (91,97)	96 (93,100)	96 (94,97)
Standard	96 (97,97)	97 (94,100)	97 (94,100)	96 (94,100)	97 (95,100)	96 (94,97)	97 (94,100)	97 (97,100)	97 (97,100)
GEE True	95 (94,96)	95 (94,96)	95 (94,96)	95 (93,96)	95 (94,96)	95 (94,97)	95 (95,97)	96 (94,97)	95 (94,96)
GEE TRM	95 (93,96)	95 (94,96)	95 (94,96)	95 (93,96)	95 (94,96)	95 (94,97)	96 (95,97)	96 (94,97)	95 (94,96)

Table 2.68: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE0, BO, MS2}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$
TRM	97 (97,98)	97 (95,100)	96 (94,97)	93 (91,94)	96 (97,97)	97 (94,100)	96 (94,97)	95 (91,97)	96 (94,100)
Strata	96 (94,97)	96 (94,97)	95 (94,97)	94 (94,94)	96 (94,97)	96 (97,97)	95 (93,97)	95 (91,97)	97 (97,97)
True	96 (94,97)	96 (94,100)	93 (91,97)	92 (91,94)	95 (93,97)	96 (94,100)	95 (92,98)	95 (94,97)	95 (94,97)
NKM	96 (96,97)	96 (95,98)	94 (91,97)	91 (89,94)	96 (94,97)	96 (94,100)	96 (94,97)	95 (91,100)	96 (97,97)
Standard	98 (97,100)	97 (96,100)	96 (97,97)	95 (94,97)	97 (96,100)	98 (97,100)	97 (96,98)	97 (94,100)	98 (97,100)
GEE True	95 (94,96)	96 (95,97)	95 (94,96)	94 (92,96)	96 (96,98)	95 (94,96)	95 (94,96)	95 (94,96)	95 (94,97)
GEE TRM	95 (94,97)	96 (96,98)	96 (95,97)	95 (94,97)	97 (96,98)	96 (95,97)	96 (94,97)	96 (95,97)	96 (95,98)

Table 2.69: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE1, BO, MS2}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 1$			$\beta_9 = 3$		
Model	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$
TRM	95 (93,97)	95 (93,97)	93 (91,97)	95 (93,97)	95 (94,97)	96 (94,97)
Strata	94 (91,97)	94 (93,97)	94 (91,97)	95 (94,97)	94 (92,97)	96 (94,97)
True	95 (94,97)	95 (94,97)	94 (91,97)	94 (91,97)	94 (91,97)	95 (94,97)
NKM	94 (91,97)	94 (91,97)	93 (91,97)	95 (94,97)	94 (93,97)	95 (94,97)
Standard	96 (94,97)	96 (94,97)	96 (94,97)	96 (94,97)	96 (94,100)	97 (94,100)
GEE True	95 (94,97)	95 (93,97)	95 (94,96)	94 (92,95)	94 (92,96)	94 (93,96)
GEE TRM	95 (94,97)	95 (93,96)	95 (94,96)	94 (92,96)	94 (93,96)	95 (93,96)

Table 2.70: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE1, BO, MS2}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 1$			$\beta_9 = 3$		
Model	$J = 48$	$J = 72$	$J = 96$	$J = 48$	$J = 72$	$J = 96$
TRM	95 (94,97)	96 (93,100)	96 (94,97)	95 (94,97)	96 (94,97)	96 (94,97)
Strata	94 (93,97)	95 (92,100)	96 (94,97)	95 (94,97)	96 (94,97)	95 (94,97)
True	94 (92,97)	94 (91,100)	95 (94,97)	95 (94,97)	95 (94,97)	96 (94,97)
NKM	95 (92,97)	95 (91,98)	95 (94,97)	95 (92,97)	95 (94,97)	95 (94,97)
Standard	96 (94,97)	96 (94,99)	97 (95,100)	97 (97,97)	97 (97,97)	97 (95,99)
GEE True	95 (94,97)	95 (94,96)	95 (93,96)	95 (94,96)	95 (94,97)	95 (94,97)
GEE TRM	96 (95,97)	96 (95,97)	96 (95,97)	96 (95,97)	96 (95,98)	96 (95,97)

Table 2.71: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE0, BO, MS2}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$
TRM	0.39 (0.37,0.4)	0.36 (0.34,0.37)	0.35 (0.33,0.37)	0.32 (0.31,0.33)	0.29 (0.28,0.31)	0.29 (0.27,0.3)	0.27 (0.26,0.28)	0.26 (0.24,0.27)	0.25 (0.23,0.26)
Strata	0.38 (0.36,0.39)	0.35 (0.34,0.36)	0.34 (0.33,0.36)	0.31 (0.3,0.32)	0.29 (0.27,0.3)	0.28 (0.26,0.29)	0.27 (0.26,0.27)	0.25 (0.23,0.26)	0.24 (0.23,0.25)
True	0.38 (0.37,0.38)	0.35 (0.34,0.36)	0.34 (0.32,0.35)	0.31 (0.3,0.31)	0.29 (0.27,0.3)	0.28 (0.26,0.29)	0.27 (0.26,0.27)	0.25 (0.24,0.26)	0.24 (0.23,0.25)
NKM	0.38 (0.37,0.39)	0.35 (0.34,0.36)	0.34 (0.32,0.36)	0.31 (0.3,0.32)	0.29 (0.27,0.3)	0.28 (0.26,0.29)	0.26 (0.26,0.27)	0.25 (0.24,0.26)	0.24 (0.23,0.25)
Standard	0.41 (0.39,0.42)	0.38 (0.36,0.4)	0.38 (0.35,0.4)	0.33 (0.32,0.34)	0.31 (0.3,0.33)	0.31 (0.29,0.32)	0.29 (0.28,0.3)	0.27 (0.26,0.29)	0.27 (0.25,0.28)
GEE True	0.4 (0.38,0.41)	0.37 (0.34,0.38)	0.36 (0.34,0.37)	0.32 (0.3,0.33)	0.3 (0.28,0.3)	0.29 (0.27,0.3)	0.27 (0.26,0.28)	0.25 (0.24,0.26)	0.25 (0.23,0.25)
GEE TRM	0.4 (0.37,0.41)	0.37 (0.34,0.39)	0.36 (0.34,0.38)	0.32 (0.3,0.33)	0.3 (0.28,0.31)	0.29 (0.27,0.31)	0.28 (0.26,0.28)	0.26 (0.24,0.27)	0.25 (0.23,0.26)

Table 2.72: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE0, BO, MS2}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$
TRM	0.3 (0.29,0.3)	0.28 (0.28,0.28)	0.27 (0.27,0.28)	0.24 (0.24,0.25)	0.23 (0.23,0.23)	0.23 (0.22,0.23)	0.21 (0.21,0.21)	0.2 (0.2,0.2)	0.19 (0.19,0.2)
Strata	0.29 (0.28,0.29)	0.27 (0.27,0.27)	0.26 (0.26,0.27)	0.24 (0.24,0.24)	0.22 (0.22,0.23)	0.22 (0.21,0.22)	0.21 (0.2,0.21)	0.19 (0.19,0.19)	0.19 (0.19,0.19)
True	0.28 (0.27,0.28)	0.26 (0.26,0.27)	0.25 (0.25,0.26)	0.23 (0.23,0.23)	0.21 (0.21,0.22)	0.21 (0.21,0.21)	0.2 (0.2,0.2)	0.19 (0.19,0.19)	0.18 (0.18,0.18)
NKM	0.28 (0.27,0.28)	0.26 (0.26,0.27)	0.25 (0.25,0.25)	0.23 (0.23,0.23)	0.21 (0.21,0.22)	0.21 (0.2,0.21)	0.2 (0.2,0.2)	0.19 (0.18,0.19)	0.18 (0.18,0.18)
Standard	0.31 (0.31,0.31)	0.3 (0.29,0.3)	0.29 (0.29,0.3)	0.26 (0.26,0.26)	0.24 (0.24,0.25)	0.24 (0.24,0.25)	0.22 (0.22,0.22)	0.21 (0.21,0.22)	0.21 (0.21,0.21)
GEE True	0.29 (0.29,0.29)	0.27 (0.27,0.28)	0.27 (0.26,0.27)	0.23 (0.23,0.24)	0.22 (0.22,0.22)	0.21 (0.21,0.22)	0.2 (0.2,0.2)	0.19 (0.19,0.19)	0.18 (0.18,0.19)
GEE TRM	0.3 (0.3,0.3)	0.28 (0.28,0.29)	0.28 (0.28,0.28)	0.24 (0.24,0.25)	0.23 (0.23,0.23)	0.22 (0.22,0.23)	0.21 (0.21,0.21)	0.2 (0.2,0.2)	0.19 (0.19,0.2)

Table 2.73: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE1, BO, MS2}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$
TRM	0.32 (0.31,0.34)	0.3 (0.28,0.31)	0.29 (0.27,0.3)	0.26 (0.25,0.28)	0.24 (0.22,0.25)	0.23 (0.22,0.24)	0.23 (0.22,0.24)	0.21 (0.2,0.22)	0.2 (0.19,0.21)
Strata	0.32 (0.31,0.33)	0.29 (0.28,0.31)	0.28 (0.27,0.3)	0.26 (0.25,0.27)	0.24 (0.22,0.25)	0.23 (0.22,0.24)	0.23 (0.22,0.23)	0.2 (0.19,0.21)	0.2 (0.19,0.2)
True	0.32 (0.3,0.33)	0.29 (0.28,0.3)	0.28 (0.27,0.29)	0.26 (0.25,0.27)	0.23 (0.22,0.25)	0.22 (0.22,0.24)	0.22 (0.22,0.23)	0.2 (0.2,0.21)	0.19 (0.19,0.2)
NKM	0.32 (0.31,0.33)	0.29 (0.28,0.31)	0.28 (0.27,0.3)	0.26 (0.25,0.27)	0.24 (0.22,0.25)	0.23 (0.22,0.24)	0.22 (0.22,0.23)	0.2 (0.19,0.21)	0.2 (0.19,0.2)
Standard	0.35 (0.33,0.36)	0.32 (0.31,0.34)	0.31 (0.3,0.33)	0.28 (0.27,0.29)	0.26 (0.24,0.28)	0.25 (0.24,0.27)	0.25 (0.24,0.26)	0.23 (0.21,0.24)	0.22 (0.21,0.23)
GEE True	0.33 (0.32,0.35)	0.3 (0.29,0.32)	0.29 (0.28,0.31)	0.26 (0.25,0.28)	0.24 (0.23,0.26)	0.23 (0.22,0.25)	0.23 (0.22,0.24)	0.2 (0.2,0.22)	0.2 (0.19,0.21)
GEE TRM	0.33 (0.32,0.35)	0.3 (0.29,0.32)	0.29 (0.28,0.31)	0.27 (0.26,0.28)	0.24 (0.23,0.26)	0.23 (0.22,0.25)	0.23 (0.22,0.24)	0.21 (0.2,0.22)	0.2 (0.19,0.22)

Table 2.74: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE1, BO, MS2}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$
TRM	0.29 (0.29,0.29)	0.27 (0.27,0.28)	0.27 (0.26,0.27)	0.24 (0.23,0.24)	0.22 (0.22,0.23)	0.22 (0.22,0.22)	0.21 (0.2,0.21)	0.19 (0.19,0.2)	0.19 (0.19,0.19)
Strata	0.28 (0.28,0.29)	0.27 (0.26,0.27)	0.26 (0.26,0.26)	0.23 (0.23,0.23)	0.22 (0.22,0.22)	0.21 (0.21,0.21)	0.2 (0.2,0.21)	0.19 (0.19,0.19)	0.18 (0.18,0.19)
True	0.27 (0.27,0.28)	0.25 (0.25,0.26)	0.25 (0.24,0.25)	0.22 (0.22,0.22)	0.21 (0.21,0.21)	0.2 (0.2,0.21)	0.19 (0.19,0.2)	0.18 (0.18,0.18)	0.17 (0.17,0.18)
NKM	0.27 (0.27,0.28)	0.26 (0.25,0.26)	0.25 (0.24,0.25)	0.22 (0.22,0.22)	0.21 (0.21,0.21)	0.2 (0.2,0.21)	0.19 (0.19,0.2)	0.18 (0.18,0.18)	0.17 (0.17,0.18)
Standard	0.3 (0.3,0.31)	0.29 (0.29,0.3)	0.28 (0.28,0.29)	0.25 (0.24,0.25)	0.24 (0.23,0.24)	0.23 (0.23,0.24)	0.22 (0.21,0.22)	0.21 (0.2,0.21)	0.2 (0.2,0.2)
GEE True	0.28 (0.27,0.29)	0.26 (0.25,0.27)	0.25 (0.25,0.26)	0.23 (0.22,0.23)	0.21 (0.2,0.21)	0.2 (0.2,0.21)	0.19 (0.19,0.2)	0.18 (0.18,0.18)	0.17 (0.17,0.18)
GEE TRM	0.29 (0.29,0.29)	0.27 (0.27,0.28)	0.27 (0.26,0.27)	0.23 (0.23,0.24)	0.22 (0.22,0.23)	0.22 (0.21,0.22)	0.2 (0.2,0.21)	0.19 (0.19,0.2)	0.19 (0.18,0.19)

Table 2.75: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE0, BO, MS2}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$
TRM	0.08 (0.07,0.08)	0.07 (0.06,0.07)	0.07 (0.06,0.07)	0.06 (0.06,0.07)	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.04,0.05)	0.05 (0.04,0.05)
Strata	0.08 (0.07,0.08)	0.07 (0.06,0.07)	0.07 (0.06,0.07)	0.06 (0.06,0.07)	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.04,0.05)
True	0.08 (0.07,0.08)	0.07 (0.06,0.07)	0.07 (0.06,0.07)	0.07 (0.06,0.07)	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.04,0.05)
NKM	0.08 (0.07,0.08)	0.07 (0.06,0.07)	0.07 (0.06,0.07)	0.07 (0.06,0.07)	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.04,0.05)
Standard	0.08 (0.07,0.08)	0.07 (0.06,0.07)	0.07 (0.06,0.08)	0.07 (0.06,0.07)	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.04,0.05)
GEE True	0.08 (0.07,0.08)	0.07 (0.07,0.08)	0.07 (0.06,0.07)	0.06 (0.06,0.06)	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.05,0.05)
GEE TRM	0.08 (0.07,0.08)	0.07 (0.07,0.08)	0.07 (0.06,0.07)	0.06 (0.06,0.06)	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.05,0.05)

Table 2.76: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE0, BO, MS2}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$
TRM	0.06 (0.06,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.05,0.06)	0.05 (0.04,0.05)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.03,0.04)	0.04 (0.03,0.04)
Strata	0.06 (0.06,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.05,0.06)	0.05 (0.04,0.05)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.03,0.04)	0.04 (0.03,0.04)
True	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.05,0.06)	0.04 (0.04,0.05)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.03,0.04)	0.04 (0.03,0.04)
NKM	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.05,0.06)	0.04 (0.04,0.05)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.03,0.04)	0.04 (0.03,0.04)
Standard	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.05,0.06)	0.05 (0.04,0.05)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.03,0.04)	0.04 (0.03,0.04)
GEE True	0.06 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.05,0.06)	0.05 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.04 (0.03,0.04)
GEE TRM	0.06 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.05,0.06)	0.05 (0.04,0.05)	0.04 (0.04,0.05)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.04 (0.03,0.04)

Table 2.77: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE1, BO, MS2}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$
TRM	0.07 (0.06,0.07)	0.06 (0.05,0.07)	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.04)
Strata	0.07 (0.06,0.08)	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.04)
True	0.07 (0.06,0.07)	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.04,0.05)	0.04 (0.04,0.05)	0.05 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.04)
NKM	0.07 (0.06,0.08)	0.06 (0.05,0.07)	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.04 (0.04,0.05)	0.04 (0.04,0.04)
Standard	0.07 (0.06,0.08)	0.06 (0.06,0.07)	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.04 (0.04,0.05)	0.04 (0.04,0.05)
GEE True	0.06 (0.06,0.07)	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.04)
GEE TRM	0.06 (0.06,0.07)	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.04)

Table 2.78: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE1, BO, MS2}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 1$	$\nu = 2$	$\nu = 3$
TRM	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.03,0.04)	0.04 (0.03,0.04)
Strata	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.03,0.04)	0.04 (0.03,0.04)
True	0.05 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.03,0.05)	0.04 (0.04,0.04)	0.04 (0.03,0.04)	0.03 (0.03,0.04)
NKM	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.05)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.03,0.04)	0.03 (0.03,0.04)
Standard	0.06 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.05,0.06)	0.05 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.04 (0.03,0.04)
GEE True	0.06 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.05,0.05)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.04 (0.03,0.04)	0.04 (0.03,0.04)
GEE TRM	0.06 (0.05,0.06)	0.05 (0.05,0.05)	0.05 (0.05,0.05)	0.05 (0.04,0.05)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.04 (0.03,0.04)	0.04 (0.03,0.04)

Table 2.79: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE0, BO, MS3}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
TRM	94 (91,97)	95 (93,97)	95 (94,97)	96 (94,97)	95 (94,97)	97 (97,97)	96 (94,98)	96 (94,97)	95 (92,97)
Strata	94 (91,96)	94 (91,96)	96 (94,97)	94 (93,97)	94 (92,97)	96 (94,97)	94 (91,95)	96 (94,99)	96 (94,99)
True	94 (91,97)	94 (91,96)	96 (94,98)	94 (91,97)	95 (94,97)	96 (97,97)	95 (93,98)	95 (92,99)	96 (94,97)
NKM	94 (91,97)	94 (91,97)	96 (94,97)	94 (91,97)	95 (94,97)	97 (97,97)	95 (94,97)	96 (94,99)	96 (94,97)
Standard	97 (97,99)	96 (94,98)	97 (94,99)	98 (97,100)	96 (96,99)	98 (97,100)	98 (97,100)	96 (94,100)	97 (95,100)
GEE True	94 (92,97)	94 (92,95)	92 (90,94)	95 (93,97)	95 (94,97)	94 (92,95)	95 (94,96)	95 (93,96)	92 (91,94)
GEE TRM	94 (93,96)	94 (92,95)	91 (90,93)	96 (95,98)	95 (94,97)	94 (92,96)	95 (94,97)	95 (93,97)	92 (90,94)

Table 2.80: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE0, BO, MS3}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
TRM	97 (95,97)	97 (96,98)	97 (96,97)	97 (96,97)	97 (96,97)	97 (97,99)	96 (97,97)	96 (94,99)	96 (94,97)
Strata	96 (95,97)	97 (95,100)	97 (96,100)	97 (97,98)	97 (96,97)	97 (97,99)	96 (96,98)	97 (94,100)	95 (93,97)
True	96 (94,98)	95 (94,96)	96 (95,97)	96 (94,97)	96 (96,97)	97 (96,99)	96 (94,98)	96 (94,97)	95 (94,97)
NKM	96 (94,97)	96 (94,97)	96 (95,97)	96 (94,97)	97 (95,97)	97 (96,99)	96 (96,98)	96 (94,98)	94 (93,97)
Standard	97 (97,100)	99 (98,100)	99 (98,100)	97 (95,100)	98 (97,100)	98 (99,100)	98 (97,100)	98 (94,100)	98 (97,100)
GEE True	95 (94,97)	95 (94,97)	95 (94,96)	95 (94,96)	96 (94,97)	95 (94,96)	95 (94,98)	96 (94,97)	95 (94,97)
GEE TRM	96 (94,98)	96 (94,98)	95 (94,97)	96 (94,97)	96 (95,98)	96 (95,98)	96 (96,98)	96 (95,98)	96 (95,98)

Table 2.81: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE1, BO, MS3}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 1$			$\beta_9 = 3$		
Model	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
TRM	95 (94,97)	94 (92,97)	94 (92,97)	94 (93,97)	94 (91,97)	96 (94,97)
Strata	94 (91,97)	94 (91,97)	94 (91,97)	94 (91,97)	93 (91,95)	96 (94,97)
True	94 (92,97)	93 (91,97)	95 (91,97)	93 (91,97)	93 (91,97)	96 (94,97)
NKM	94 (92,97)	94 (91,97)	95 (91,97)	93 (91,97)	92 (91,94)	96 (94,97)
Standard	98 (97,100)	96 (94,97)	96 (94,97)	97 (96,100)	96 (94,97)	97 (96,100)
GEE True	94 (93,96)	94 (93,96)	94 (92,95)	94 (92,95)	94 (93,96)	94 (92,95)
GEE TRM	95 (93,97)	94 (93,96)	93 (92,95)	95 (93,97)	94 (92,96)	93 (92,95)

Table 2.82: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE1, BO, MS3}. Results are stratified by β_9 and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
TRM	95 (94,97)	94 (92,95)	97 (96,98)	97 (96,98)	94 (92,97)	97 (95,97)	97 (95,98)	96 (94,97)	94 (92,96)
Strata	95 (93,96)	94 (91,95)	97 (96,100)	97 (95,98)	94 (91,97)	97 (97,98)	96 (94,98)	95 (94,97)	95 (92,98)
True	93 (92,94)	94 (91,95)	97 (96,98)	94 (93,97)	95 (93,97)	96 (94,97)	96 (94,97)	94 (94,97)	96 (94,99)
NKM	93 (92,96)	93 (91,95)	96 (96,98)	95 (93,96)	94 (94,97)	96 (95,97)	96 (94,98)	94 (94,97)	96 (94,97)
Standard	98 (98,100)	94 (92,97)	97 (97,100)	99 (97,100)	96 (94,97)	97 (97,98)	98 (97,100)	97 (97,98)	95 (94,96)
GEE True	95 (93,97)	95 (94,97)	94 (93,96)	95 (94,96)	95 (94,96)	95 (94,96)	95 (93,97)	95 (94,96)	95 (94,97)
GEE TRM	97 (95,98)	96 (94,97)	94 (93,96)	97 (96,98)	95 (94,97)	95 (94,97)	97 (96,98)	96 (94,97)	95 (94,97)

Table 2.83: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE0, BO, MS3}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
TRM	0.25 (0.23,0.28)	0.3 (0.29,0.32)	0.23 (0.22,0.25)	0.21 (0.18,0.22)	0.25 (0.23,0.26)	0.18 (0.17,0.2)	0.18 (0.16,0.19)	0.21 (0.2,0.22)	0.16 (0.15,0.16)
Strata	0.24 (0.22,0.26)	0.3 (0.28,0.32)	0.23 (0.22,0.25)	0.19 (0.18,0.21)	0.24 (0.22,0.26)	0.19 (0.17,0.2)	0.17 (0.15,0.18)	0.2 (0.19,0.21)	0.16 (0.15,0.16)
True	0.24 (0.22,0.26)	0.29 (0.28,0.31)	0.23 (0.22,0.24)	0.19 (0.17,0.21)	0.24 (0.22,0.26)	0.18 (0.17,0.2)	0.17 (0.15,0.18)	0.2 (0.19,0.21)	0.16 (0.15,0.16)
NKM	0.24 (0.22,0.26)	0.29 (0.28,0.31)	0.23 (0.22,0.24)	0.19 (0.18,0.21)	0.24 (0.22,0.26)	0.18 (0.17,0.2)	0.17 (0.15,0.18)	0.2 (0.19,0.21)	0.16 (0.15,0.16)
Standard	0.3 (0.27,0.33)	0.32 (0.31,0.35)	0.25 (0.23,0.27)	0.25 (0.22,0.26)	0.27 (0.25,0.29)	0.2 (0.18,0.21)	0.21 (0.2,0.23)	0.23 (0.21,0.24)	0.17 (0.15,0.18)
GEE True	0.25 (0.23,0.27)	0.3 (0.28,0.32)	0.2 (0.19,0.22)	0.2 (0.19,0.21)	0.24 (0.22,0.26)	0.16 (0.15,0.18)	0.17 (0.16,0.18)	0.2 (0.19,0.21)	0.14 (0.13,0.15)
GEE TRM	0.25 (0.24,0.27)	0.3 (0.28,0.32)	0.2 (0.18,0.22)	0.21 (0.19,0.22)	0.24 (0.22,0.26)	0.16 (0.14,0.18)	0.18 (0.16,0.19)	0.21 (0.19,0.22)	0.14 (0.12,0.15)

Table 2.84: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE0, BO, MS3}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
TRM	0.2 (0.19,0.21)	0.19 (0.18,0.21)	0.12 (0.12,0.13)	0.16 (0.16,0.17)	0.15 (0.15,0.17)	0.1 (0.09,0.11)	0.14 (0.13,0.15)	0.13 (0.13,0.14)	0.08 (0.07,0.09)
Strata	0.19 (0.18,0.19)	0.19 (0.18,0.2)	0.13 (0.12,0.13)	0.15 (0.15,0.16)	0.15 (0.15,0.17)	0.1 (0.09,0.11)	0.13 (0.13,0.14)	0.13 (0.13,0.14)	0.08 (0.08,0.09)
True	0.17 (0.17,0.17)	0.19 (0.18,0.2)	0.13 (0.12,0.13)	0.14 (0.14,0.14)	0.15 (0.14,0.16)	0.1 (0.09,0.11)	0.12 (0.12,0.12)	0.13 (0.12,0.13)	0.08 (0.08,0.09)
NKM	0.17 (0.17,0.18)	0.19 (0.18,0.2)	0.13 (0.12,0.13)	0.14 (0.13,0.14)	0.15 (0.14,0.16)	0.1 (0.09,0.11)	0.12 (0.12,0.12)	0.13 (0.12,0.13)	0.08 (0.08,0.09)
Standard	0.24 (0.23,0.25)	0.21 (0.2,0.23)	0.13 (0.13,0.14)	0.19 (0.19,0.2)	0.17 (0.16,0.19)	0.1 (0.1,0.11)	0.17 (0.16,0.18)	0.15 (0.14,0.16)	0.09 (0.08,0.1)
GEE True	0.18 (0.17,0.18)	0.19 (0.18,0.21)	0.12 (0.11,0.13)	0.14 (0.14,0.15)	0.15 (0.14,0.16)	0.1 (0.09,0.11)	0.12 (0.12,0.12)	0.13 (0.12,0.14)	0.08 (0.07,0.09)
GEE TRM	0.2 (0.19,0.21)	0.2 (0.18,0.21)	0.12 (0.1,0.13)	0.16 (0.15,0.17)	0.16 (0.15,0.17)	0.1 (0.09,0.11)	0.14 (0.13,0.15)	0.14 (0.13,0.15)	0.08 (0.07,0.09)

Table 2.85: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE1, BO, MS3}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
TRM	0.22 (0.2,0.24)	0.3 (0.28,0.32)	0.28 (0.27,0.29)	0.18 (0.16,0.2)	0.25 (0.23,0.26)	0.23 (0.22,0.24)	0.15 (0.14,0.16)	0.22 (0.21,0.22)	0.2 (0.19,0.21)
Strata	0.21 (0.19,0.23)	0.3 (0.28,0.31)	0.27 (0.26,0.29)	0.17 (0.15,0.19)	0.24 (0.23,0.25)	0.23 (0.22,0.24)	0.15 (0.13,0.16)	0.21 (0.2,0.21)	0.2 (0.19,0.2)
True	0.21 (0.19,0.22)	0.29 (0.27,0.31)	0.27 (0.26,0.28)	0.17 (0.15,0.18)	0.24 (0.23,0.25)	0.23 (0.22,0.23)	0.14 (0.13,0.15)	0.21 (0.2,0.21)	0.2 (0.19,0.2)
NKM	0.21 (0.19,0.23)	0.29 (0.27,0.31)	0.27 (0.26,0.28)	0.17 (0.15,0.18)	0.24 (0.23,0.25)	0.23 (0.22,0.23)	0.14 (0.13,0.16)	0.21 (0.2,0.21)	0.2 (0.19,0.2)
Standard	0.26 (0.23,0.29)	0.32 (0.3,0.34)	0.29 (0.28,0.31)	0.21 (0.18,0.23)	0.27 (0.25,0.28)	0.24 (0.23,0.26)	0.18 (0.16,0.19)	0.23 (0.22,0.24)	0.21 (0.2,0.22)
GEE True	0.21 (0.2,0.22)	0.3 (0.29,0.31)	0.27 (0.25,0.29)	0.17 (0.15,0.18)	0.24 (0.23,0.25)	0.22 (0.2,0.23)	0.15 (0.13,0.15)	0.21 (0.2,0.22)	0.19 (0.17,0.2)
GEE TRM	0.22 (0.2,0.24)	0.3 (0.29,0.31)	0.27 (0.25,0.29)	0.18 (0.16,0.19)	0.25 (0.24,0.26)	0.22 (0.2,0.23)	0.15 (0.14,0.16)	0.21 (0.2,0.22)	0.19 (0.17,0.2)

Table 2.86: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE1, BO, MS3}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
TRM	0.2 (0.2,0.2)	0.21 (0.2,0.22)	0.14 (0.13,0.14)	0.16 (0.16,0.17)	0.17 (0.16,0.18)	0.11 (0.11,0.11)	0.14 (0.13,0.14)	0.15 (0.14,0.15)	0.09 (0.09,0.09)
Strata	0.19 (0.19,0.2)	0.21 (0.2,0.22)	0.14 (0.14,0.14)	0.16 (0.16,0.16)	0.17 (0.16,0.18)	0.11 (0.11,0.11)	0.14 (0.13,0.14)	0.15 (0.14,0.15)	0.09 (0.09,0.09)
True	0.17 (0.17,0.18)	0.2 (0.19,0.21)	0.14 (0.14,0.14)	0.14 (0.14,0.14)	0.16 (0.16,0.17)	0.11 (0.11,0.11)	0.12 (0.12,0.12)	0.14 (0.14,0.14)	0.09 (0.09,0.09)
NKM	0.17 (0.17,0.18)	0.2 (0.19,0.21)	0.14 (0.14,0.14)	0.14 (0.14,0.14)	0.16 (0.16,0.17)	0.11 (0.11,0.11)	0.12 (0.12,0.12)	0.14 (0.14,0.14)	0.09 (0.09,0.09)
Standard	0.24 (0.23,0.24)	0.23 (0.22,0.24)	0.15 (0.14,0.15)	0.19 (0.19,0.2)	0.19 (0.18,0.2)	0.12 (0.11,0.12)	0.16 (0.16,0.17)	0.16 (0.15,0.17)	0.1 (0.1,0.1)
GEE True	0.18 (0.17,0.18)	0.21 (0.2,0.22)	0.14 (0.13,0.16)	0.14 (0.14,0.15)	0.17 (0.16,0.18)	0.12 (0.1,0.13)	0.12 (0.12,0.12)	0.14 (0.14,0.15)	0.1 (0.09,0.11)
GEE TRM	0.2 (0.19,0.21)	0.22 (0.2,0.23)	0.14 (0.13,0.16)	0.16 (0.15,0.17)	0.18 (0.16,0.19)	0.12 (0.1,0.13)	0.14 (0.13,0.15)	0.15 (0.14,0.16)	0.1 (0.09,0.11)

Table 2.87: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE0, BO, MS3}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
TRM	0.05 (0.04,0.05)	0.06 (0.05,0.07)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)
Strata	0.05 (0.04,0.05)	0.06 (0.05,0.06)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.04)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.03,0.03)
True	0.05 (0.04,0.05)	0.06 (0.05,0.06)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.04)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.03,0.03)
NKM	0.05 (0.04,0.05)	0.06 (0.05,0.06)	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.03,0.03)
Standard	0.05 (0.05,0.05)	0.06 (0.05,0.06)	0.04 (0.04,0.05)	0.04 (0.04,0.05)	0.05 (0.05,0.05)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.03 (0.02,0.03)
GEE True	0.05 (0.04,0.05)	0.06 (0.05,0.06)	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.04)	0.03 (0.03,0.03)	0.04 (0.04,0.05)	0.03 (0.02,0.03)
GEE TRM	0.05 (0.04,0.05)	0.06 (0.05,0.06)	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.04)	0.03 (0.03,0.04)	0.04 (0.04,0.05)	0.03 (0.02,0.03)

Table 2.88: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE0, BO, MS3}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
TRM	0.03 (0.03,0.03)	0.04 (0.03,0.04)	0.02 (0.02,0.02)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)	0.02 (0.02,0.03)	0.03 (0.02,0.03)	0.02 (0.01,0.02)
Strata	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.02 (0.02,0.02)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)	0.02 (0.02,0.03)	0.03 (0.02,0.03)	0.02 (0.01,0.02)
True	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.02 (0.02,0.02)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)	0.02 (0.02,0.03)	0.03 (0.02,0.03)	0.02 (0.02,0.02)
NKM	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.02 (0.02,0.02)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)	0.02 (0.02,0.03)	0.03 (0.02,0.03)	0.02 (0.02,0.02)
Standard	0.04 (0.03,0.04)	0.04 (0.03,0.04)	0.02 (0.02,0.02)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)	0.03 (0.02,0.03)	0.03 (0.02,0.03)	0.02 (0.02,0.02)
GEE True	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.02 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)	0.02 (0.02,0.03)	0.03 (0.02,0.03)	0.02 (0.01,0.02)
GEE TRM	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.02 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)	0.03 (0.02,0.03)	0.03 (0.02,0.03)	0.02 (0.01,0.02)

Table 2.89: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE1, BO, MS3}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
TRM	0.04 (0.04,0.05)	0.06 (0.06,0.07)	0.05 (0.05,0.06)	0.03 (0.03,0.04)	0.05 (0.05,0.05)	0.04 (0.04,0.05)	0.03 (0.03,0.03)	0.05 (0.04,0.05)	0.04 (0.03,0.04)
Strata	0.04 (0.04,0.05)	0.06 (0.06,0.07)	0.05 (0.05,0.06)	0.04 (0.03,0.04)	0.05 (0.05,0.05)	0.04 (0.04,0.05)	0.03 (0.03,0.03)	0.05 (0.04,0.05)	0.04 (0.03,0.04)
True	0.04 (0.04,0.05)	0.06 (0.06,0.07)	0.05 (0.05,0.06)	0.04 (0.03,0.04)	0.05 (0.05,0.05)	0.04 (0.04,0.05)	0.03 (0.03,0.03)	0.05 (0.04,0.05)	0.04 (0.03,0.04)
NKM	0.04 (0.04,0.05)	0.06 (0.06,0.07)	0.05 (0.05,0.06)	0.04 (0.03,0.04)	0.05 (0.05,0.05)	0.04 (0.04,0.05)	0.03 (0.03,0.03)	0.05 (0.04,0.05)	0.04 (0.03,0.04)
Standard	0.05 (0.04,0.05)	0.06 (0.06,0.07)	0.05 (0.05,0.06)	0.04 (0.03,0.04)	0.05 (0.05,0.06)	0.04 (0.04,0.05)	0.03 (0.03,0.03)	0.05 (0.04,0.05)	0.04 (0.03,0.04)
GEE True	0.04 (0.04,0.04)	0.06 (0.06,0.06)	0.05 (0.05,0.06)	0.03 (0.03,0.04)	0.05 (0.05,0.05)	0.04 (0.04,0.05)	0.03 (0.03,0.03)	0.04 (0.04,0.04)	0.04 (0.04,0.04)
GEE TRM	0.04 (0.04,0.05)	0.06 (0.06,0.06)	0.05 (0.05,0.06)	0.03 (0.03,0.04)	0.05 (0.05,0.05)	0.04 (0.04,0.05)	0.03 (0.03,0.03)	0.04 (0.04,0.04)	0.04 (0.04,0.04)

Table 2.90: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE1, BO, MS3}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$	$\alpha = 0.5$	$\alpha = 2$	$\alpha = 4$
TRM	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.02 (0.02,0.02)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)
Strata	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.02 (0.02,0.02)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)
True	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.02 (0.02,0.02)	0.02 (0.02,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)
NKM	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.02 (0.02,0.02)	0.02 (0.02,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)
Standard	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)	0.02 (0.02,0.02)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)
GEE True	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.02 (0.02,0.02)	0.02 (0.02,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)
GEE TRM	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.02 (0.02,0.02)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.02 (0.02,0.02)

2.7.7 Continuous Outcomes Model Specification 4 Results

Table 2.91: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE0, CO, MS4}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	98 (97,100)	97 (96,100)	94 (92,97)	96 (94,98)	95 (94,97)	95 (94,97)	97 (94,100)	95 (94,97)	95 (94,97)
Strata	96 (95,97)	96 (94,100)	93 (91,97)	94 (94,96)	93 (91,96)	94 (92,97)	94 (94,96)	94 (91,97)	95 (94,97)
True	96 (94,98)	97 (94,100)	94 (91,97)	94 (94,96)	93 (91,97)	94 (94,96)	96 (94,100)	94 (93,97)	95 (93,99)
NKM	96 (95,98)	96 (94,97)	93 (91,96)	94 (93,95)	93 (91,97)	93 (91,96)	96 (94,97)	94 (94,97)	94 (93,97)
Standard	99 (99,100)	98 (97,100)	95 (94,97)	98 (97,100)	95 (94,97)	95 (91,97)	98 (97,100)	97 (96,99)	95 (94,97)
GEE True	95 (94,96)	96 (95,97)	96 (95,97)	95 (94,96)	95 (94,96)	96 (94,97)	96 (94,97)	95 (94,97)	95 (94,97)
GEE TRM	97 (97,99)	97 (96,99)	96 (94,97)	98 (96,99)	97 (95,98)	96 (94,98)	97 (96,99)	97 (96,98)	96 (95,97)

Table 2.92: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE0, CO, MS4}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	98 (97,99)	98 (97,100)	96 (94,99)	97 (97,99)	96 (94,97)	96 (95,97)	96 (94,100)	96 (96,98)	94 (93,95)
Strata	97 (95,98)	97 (95,100)	97 (95,99)	97 (96,97)	95 (94,97)	95 (95,97)	95 (93,100)	96 (95,98)	95 (94,97)
True	94 (94,95)	96 (94,96)	96 (95,97)	94 (92,97)	92 (91,93)	96 (94,100)	95 (91,97)	95 (94,97)	94 (93,97)
NKM	95 (94,96)	96 (94,98)	96 (95,99)	95 (94,97)	94 (94,94)	95 (94,99)	94 (91,97)	96 (94,99)	94 (94,96)
Standard	98 (97,100)	99 (97,100)	97 (95,99)	99 (98,100)	97 (97,98)	96 (95,98)	98 (97,100)	97 (97,98)	95 (94,97)
GEE True	95 (94,97)	95 (94,96)	95 (94,96)	96 (94,97)	96 (94,97)	95 (94,96)	95 (94,96)	95 (94,96)	95 (94,96)
GEE TRM	98 (97,99)	97 (96,98)	96 (95,98)	98 (97,99)	97 (97,98)	97 (96,98)	98 (97,99)	97 (96,98)	96 (94,98)

Table 2.93: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE1, CO, MS4}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 1$			$\beta_9 = 3$		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	96 (94,97)	96 (94,97)	94 (92,97)	97 (95,100)	95 (94,97)	96 (94,97)
Strata	94 (94,97)	95 (94,97)	94 (92,97)	95 (94,97)	95 (92,97)	95 (94,97)
True	94 (91,97)	95 (94,97)	95 (94,97)	93 (91,95)	94 (91,97)	95 (92,97)
NKM	94 (94,97)	95 (94,97)	94 (91,97)	95 (94,97)	94 (92,97)	95 (94,97)
Standard	97 (96,100)	97 (97,100)	96 (94,97)	97 (97,99)	97 (96,100)	96 (96,98)
GEE True	95 (94,97)	95 (94,97)	95 (94,97)	93 (92,95)	94 (93,96)	94 (93,96)
GEE TRM	97 (96,99)	96 (95,98)	95 (94,97)	96 (95,98)	96 (95,97)	95 (93,97)

Table 2.94: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE1, CO, MS4}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 1$			$\beta_9 = 3$		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	98 (97,99)	98 (97,100)	97 (96,98)	99 (98,100)	98 (97,100)	98 (97,99)
Strata	97 (97,98)	97 (96,99)	97 (95,98)	98 (97,99)	98 (97,99)	98 (97,99)
True	94 (94,97)	95 (93,97)	94 (93,95)	94 (93,96)	94 (92,96)	95 (94,96)
NKM	95 (93,98)	96 (94,97)	95 (93,95)	95 (93,98)	95 (92,97)	95 (93,97)
Standard	99 (98,100)	98 (97,100)	98 (96,99)	99 (98,100)	99 (97,100)	98 (97,99)
GEE True	95 (94,97)	95 (93,96)	95 (94,97)	93 (91,95)	94 (93,96)	95 (93,97)
GEE TRM	99 (98,99)	98 (98,99)	97 (96,98)	99 (98,100)	99 (98,99)	98 (97,99)

Table 2.95: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE0, CO, MS4}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	1.71 (1.54,1.78)	2.14 (2.02,2.2)	3.1 (3.05,3.13)	1.35 (1.21,1.39)	1.72 (1.63,1.74)	2.51 (2.45,2.52)	1.13 (1.03,1.18)	1.44 (1.4,1.47)	2.14 (2.13,2.17)
Strata	1.5 (1.46,1.53)	1.99 (1.97,2.01)	3.04 (2.99,3.08)	1.19 (1.16,1.2)	1.6 (1.57,1.61)	2.45 (2.4,2.48)	1 (0.98,1.01)	1.37 (1.36,1.38)	2.09 (2.07,2.13)
True	1.46 (1.45,1.48)	2 (1.97,2.02)	3.08 (3.02,3.13)	1.15 (1.14,1.16)	1.6 (1.58,1.63)	2.49 (2.44,2.54)	0.98 (0.97,0.99)	1.37 (1.35,1.39)	2.14 (2.12,2.15)
NKM	1.46 (1.44,1.47)	1.96 (1.93,2.01)	3.03 (3.01,3.09)	1.14 (1.14,1.16)	1.57 (1.54,1.6)	2.44 (2.42,2.47)	0.98 (0.97,0.99)	1.36 (1.34,1.37)	2.09 (2.05,2.13)
Standard	2.01 (1.77,2.12)	2.38 (2.18,2.46)	3.26 (3.21,3.3)	1.61 (1.41,1.68)	1.9 (1.77,1.97)	2.61 (2.55,2.67)	1.33 (1.19,1.42)	1.58 (1.51,1.66)	2.24 (2.21,2.29)
GEE True	1.44 (1.42,1.46)	1.99 (1.97,2.01)	3.13 (3.1,3.17)	1.15 (1.14,1.16)	1.6 (1.58,1.61)	2.49 (2.48,2.51)	0.98 (0.98,0.99)	1.37 (1.36,1.38)	2.15 (2.13,2.16)
GEE TRM	1.72 (1.5,1.99)	2.19 (2.03,2.38)	3.22 (3.11,3.31)	1.39 (1.2,1.59)	1.77 (1.63,1.93)	2.59 (2.5,2.71)	1.19 (1.03,1.37)	1.52 (1.41,1.65)	2.23 (2.17,2.32)

Table 2.96: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE0, CO, MS4}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	1.31 (1.27,1.35)	1.61 (1.6,1.63)	2.26 (2.25,2.26)	1.08 (1.04,1.13)	1.3 (1.29,1.3)	1.86 (1.86,1.87)	0.93 (0.89,0.97)	1.11 (1.1,1.11)	1.58 (1.58,1.6)
Strata	1.2 (1.19,1.2)	1.55 (1.53,1.57)	2.23 (2.21,2.23)	0.98 (0.98,0.99)	1.24 (1.24,1.24)	1.82 (1.81,1.83)	0.84 (0.84,0.85)	1.06 (1.06,1.07)	1.55 (1.54,1.57)
True	0.98 (0.97,1)	1.38 (1.37,1.4)	2.13 (2.12,2.14)	0.8 (0.79,0.8)	1.12 (1.11,1.12)	1.75 (1.73,1.76)	0.68 (0.68,0.69)	0.95 (0.94,0.95)	1.47 (1.46,1.48)
NKM	1.04 (1.03,1.06)	1.42 (1.41,1.44)	2.13 (2.12,2.15)	0.86 (0.85,0.86)	1.15 (1.14,1.15)	1.77 (1.75,1.78)	0.73 (0.73,0.74)	0.99 (0.98,0.99)	1.5 (1.49,1.51)
Standard	1.47 (1.4,1.57)	1.72 (1.71,1.72)	2.33 (2.31,2.34)	1.21 (1.16,1.31)	1.38 (1.38,1.39)	1.93 (1.92,1.93)	1.04 (0.99,1.12)	1.18 (1.18,1.19)	1.63 (1.62,1.64)
GEE True	1 (0.99,1.01)	1.39 (1.38,1.4)	2.18 (2.16,2.19)	0.81 (0.8,0.81)	1.12 (1.12,1.12)	1.75 (1.74,1.76)	0.69 (0.69,0.69)	0.96 (0.96,0.97)	1.5 (1.5,1.51)
GEE TRM	1.41 (1.29,1.54)	1.69 (1.61,1.8)	2.36 (2.28,2.44)	1.14 (1.05,1.26)	1.38 (1.29,1.48)	1.92 (1.86,1.99)	0.98 (0.9,1.08)	1.19 (1.12,1.27)	1.65 (1.6,1.72)

Table 2.97: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE1, CO, MS4}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	1.69 (1.58,1.79)	2.15 (2.07,2.22)	3.17 (3.09,3.23)	1.35 (1.26,1.42)	1.72 (1.65,1.79)	2.51 (2.48,2.54)	1.13 (1.07,1.2)	1.47 (1.41,1.54)	2.17 (2.13,2.21)
Strata	1.53 (1.48,1.57)	2.04 (2.01,2.1)	3.11 (3.07,3.16)	1.22 (1.18,1.25)	1.64 (1.61,1.66)	2.47 (2.45,2.5)	1.04 (1,1.07)	1.41 (1.38,1.43)	2.12 (2.09,2.16)
True	1.45 (1.43,1.47)	2.01 (1.97,2.05)	3.11 (3.07,3.16)	1.14 (1.13,1.16)	1.6 (1.59,1.61)	2.49 (2.46,2.52)	0.98 (0.97,1)	1.37 (1.36,1.39)	2.14 (2.11,2.17)
NKM	1.49 (1.43,1.55)	2.02 (1.95,2.06)	3.1 (3.07,3.14)	1.18 (1.14,1.21)	1.62 (1.59,1.64)	2.46 (2.44,2.49)	1.02 (0.99,1.05)	1.39 (1.37,1.41)	2.12 (2.09,2.16)
Standard	2.11 (2,2.3)	2.48 (2.37,2.63)	3.39 (3.31,3.46)	1.7 (1.61,1.83)	1.99 (1.88,2.09)	2.7 (2.65,2.73)	1.43 (1.37,1.5)	1.71 (1.64,1.82)	2.34 (2.26,2.41)
GEE True	1.44 (1.42,1.46)	2 (1.98,2.02)	3.13 (3.1,3.15)	1.15 (1.14,1.16)	1.59 (1.58,1.6)	2.5 (2.48,2.52)	0.99 (0.98,0.99)	1.37 (1.36,1.38)	2.14 (2.12,2.16)
GEE TRM	1.89 (1.77,2.05)	2.34 (2.25,2.47)	3.32 (3.22,3.43)	1.52 (1.43,1.65)	1.88 (1.79,1.99)	2.68 (2.6,2.75)	1.31 (1.23,1.41)	1.62 (1.54,1.72)	2.3 (2.25,2.37)

Table 2.98: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE1, CO, MS4}. Results are stratified by number of clusters and cluster variance.

Model	48 Clusters			72 Clusters			96 Clusters		
	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	2.14 (2.05,2.28)	2.33 (2.23,2.44)	2.86 (2.8,2.94)	1.77 (1.68,1.86)	1.93 (1.86,1.99)	2.33 (2.28,2.4)	1.52 (1.45,1.61)	1.67 (1.61,1.73)	2.03 (1.95,2.1)
Strata	2.03 (1.94,2.18)	2.22 (2.11,2.34)	2.77 (2.69,2.84)	1.68 (1.59,1.78)	1.84 (1.78,1.91)	2.26 (2.22,2.32)	1.44 (1.38,1.51)	1.59 (1.54,1.64)	1.97 (1.9,2.04)
True	0.99 (0.98,0.99)	1.38 (1.37,1.39)	2.14 (2.12,2.16)	0.79 (0.79,0.8)	1.1 (1.09,1.11)	1.73 (1.72,1.73)	0.68 (0.68,0.69)	0.95 (0.95,0.96)	1.49 (1.49,1.49)
NKM	1.39 (1.37,1.4)	1.68 (1.64,1.71)	2.35 (2.33,2.36)	1.14 (1.12,1.16)	1.37 (1.36,1.39)	1.91 (1.9,1.92)	0.98 (0.96,1)	1.19 (1.18,1.21)	1.65 (1.64,1.66)
Standard	2.22 (2.13,2.37)	2.4 (2.3,2.52)	2.93 (2.86,3.01)	1.84 (1.76,1.93)	2 (1.94,2.05)	2.38 (2.32,2.45)	1.59 (1.51,1.67)	1.73 (1.66,1.8)	2.07 (2,2.15)
GEE True	1 (0.99,1.01)	1.39 (1.38,1.4)	2.18 (2.16,2.19)	0.8 (0.8,0.81)	1.12 (1.11,1.12)	1.75 (1.74,1.76)	0.69 (0.69,0.7)	0.96 (0.95,0.96)	1.51 (1.5,1.51)
GEE TRM	2.26 (2.13,2.37)	2.44 (2.34,2.55)	2.96 (2.86,3.04)	1.83 (1.73,1.91)	2 (1.9,2.06)	2.4 (2.33,2.46)	1.58 (1.5,1.65)	1.72 (1.65,1.78)	2.07 (2.01,2.12)

Table 2.99: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE0, CO, MS4}. Results are stratified by number of clusters and cluster variance.

Model	24 Clusters			36 Clusters			48 Clusters		
	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	0.28 (0.26,0.29)	0.38 (0.33,0.42)	0.61 (0.53,0.66)	0.24 (0.22,0.26)	0.33 (0.31,0.35)	0.48 (0.44,0.5)	0.2 (0.19,0.22)	0.28 (0.26,0.3)	0.43 (0.4,0.46)
Strata	0.28 (0.26,0.29)	0.38 (0.33,0.4)	0.61 (0.55,0.65)	0.24 (0.21,0.25)	0.33 (0.3,0.35)	0.48 (0.44,0.51)	0.2 (0.18,0.22)	0.28 (0.26,0.3)	0.43 (0.39,0.46)
True	0.28 (0.26,0.29)	0.37 (0.34,0.41)	0.61 (0.56,0.65)	0.23 (0.21,0.23)	0.32 (0.3,0.35)	0.49 (0.45,0.52)	0.19 (0.18,0.21)	0.28 (0.26,0.3)	0.42 (0.38,0.46)
NKM	0.27 (0.26,0.28)	0.37 (0.34,0.4)	0.6 (0.55,0.65)	0.23 (0.21,0.23)	0.33 (0.3,0.34)	0.49 (0.44,0.5)	0.19 (0.18,0.21)	0.28 (0.25,0.3)	0.43 (0.4,0.45)
Standard	0.29 (0.26,0.31)	0.39 (0.34,0.42)	0.61 (0.57,0.65)	0.25 (0.23,0.28)	0.34 (0.3,0.37)	0.5 (0.45,0.52)	0.21 (0.2,0.23)	0.29 (0.27,0.31)	0.43 (0.39,0.46)
GEE True	0.27 (0.26,0.28)	0.37 (0.35,0.39)	0.6 (0.57,0.63)	0.22 (0.21,0.23)	0.32 (0.3,0.32)	0.48 (0.45,0.49)	0.2 (0.18,0.21)	0.27 (0.26,0.29)	0.42 (0.4,0.45)
GEE TRM	0.28 (0.27,0.3)	0.38 (0.36,0.39)	0.6 (0.56,0.63)	0.23 (0.22,0.24)	0.32 (0.3,0.34)	0.48 (0.45,0.49)	0.2 (0.19,0.22)	0.27 (0.26,0.28)	0.43 (0.4,0.45)

Table 2.100: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE0, CO, MS4}. Results are stratified by number of clusters and cluster variance.

Model	48 Clusters			72 Clusters			96 Clusters		
	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	0.23 (0.22,0.23)	0.27 (0.27,0.29)	0.42 (0.4,0.42)	0.18 (0.18,0.19)	0.25 (0.24,0.26)	0.35 (0.34,0.35)	0.17 (0.15,0.19)	0.2 (0.19,0.22)	0.34 (0.31,0.37)
Strata	0.22 (0.22,0.23)	0.27 (0.27,0.29)	0.41 (0.38,0.43)	0.18 (0.18,0.19)	0.25 (0.24,0.26)	0.35 (0.34,0.35)	0.17 (0.15,0.18)	0.2 (0.19,0.22)	0.34 (0.31,0.37)
True	0.2 (0.19,0.21)	0.24 (0.25,0.27)	0.41 (0.37,0.44)	0.16 (0.15,0.17)	0.23 (0.22,0.24)	0.34 (0.34,0.36)	0.14 (0.13,0.15)	0.19 (0.18,0.21)	0.32 (0.3,0.33)
NKM	0.21 (0.2,0.22)	0.26 (0.26,0.28)	0.39 (0.36,0.44)	0.17 (0.16,0.18)	0.24 (0.23,0.24)	0.34 (0.33,0.34)	0.15 (0.14,0.17)	0.19 (0.17,0.21)	0.33 (0.31,0.36)
Standard	0.24 (0.23,0.24)	0.27 (0.27,0.29)	0.41 (0.39,0.42)	0.19 (0.18,0.2)	0.26 (0.25,0.27)	0.35 (0.34,0.36)	0.17 (0.16,0.18)	0.21 (0.19,0.23)	0.34 (0.31,0.38)
GEE True	0.2 (0.19,0.2)	0.27 (0.26,0.29)	0.43 (0.4,0.45)	0.16 (0.15,0.17)	0.22 (0.21,0.23)	0.34 (0.33,0.36)	0.14 (0.13,0.14)	0.2 (0.19,0.2)	0.3 (0.28,0.31)
GEE TRM	0.23 (0.22,0.24)	0.3 (0.29,0.31)	0.45 (0.42,0.46)	0.19 (0.18,0.2)	0.24 (0.23,0.26)	0.36 (0.34,0.38)	0.16 (0.15,0.17)	0.22 (0.21,0.22)	0.31 (0.29,0.32)

Table 2.101: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE1, CO, MS4}. Results are stratified by number of clusters and cluster variance.

Model	24 Clusters			36 Clusters			48 Clusters		
	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	0.3 (0.28,0.32)	0.4 (0.36,0.45)	0.59 (0.55,0.64)	0.24 (0.22,0.26)	0.32 (0.3,0.34)	0.5 (0.44,0.55)	0.21 (0.2,0.22)	0.28 (0.26,0.3)	0.44 (0.42,0.46)
Strata	0.3 (0.27,0.32)	0.4 (0.36,0.46)	0.6 (0.54,0.64)	0.23 (0.21,0.25)	0.32 (0.29,0.34)	0.5 (0.46,0.54)	0.21 (0.2,0.22)	0.28 (0.26,0.29)	0.43 (0.41,0.46)
True	0.29 (0.27,0.31)	0.4 (0.35,0.45)	0.6 (0.55,0.64)	0.23 (0.21,0.25)	0.32 (0.29,0.34)	0.49 (0.46,0.54)	0.21 (0.19,0.22)	0.27 (0.26,0.3)	0.43 (0.41,0.46)
NKM	0.29 (0.27,0.31)	0.4 (0.36,0.44)	0.6 (0.55,0.65)	0.23 (0.21,0.25)	0.32 (0.29,0.35)	0.49 (0.46,0.53)	0.21 (0.19,0.22)	0.28 (0.26,0.3)	0.43 (0.41,0.45)
Standard	0.35 (0.31,0.38)	0.44 (0.4,0.46)	0.63 (0.59,0.68)	0.28 (0.26,0.3)	0.35 (0.32,0.38)	0.52 (0.48,0.58)	0.25 (0.23,0.27)	0.3 (0.28,0.33)	0.45 (0.43,0.47)
GEE True	0.29 (0.27,0.31)	0.39 (0.37,0.41)	0.61 (0.58,0.64)	0.23 (0.22,0.25)	0.32 (0.3,0.34)	0.49 (0.46,0.52)	0.2 (0.19,0.21)	0.28 (0.26,0.29)	0.43 (0.41,0.45)
GEE TRM	0.33 (0.3,0.37)	0.42 (0.39,0.45)	0.63 (0.6,0.66)	0.27 (0.25,0.29)	0.35 (0.33,0.36)	0.51 (0.49,0.54)	0.24 (0.22,0.25)	0.3 (0.29,0.32)	0.45 (0.43,0.47)

Table 2.102: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE1, CO, MS4}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
TRM	0.33 (0.3,0.36)	0.4 (0.38,0.41)	0.5 (0.48,0.52)	0.29 (0.28,0.3)	0.33 (0.3,0.34)	0.41 (0.39,0.43)	0.25 (0.23,0.27)	0.29 (0.26,0.29)	0.35 (0.34,0.37)
Strata	0.33 (0.31,0.36)	0.4 (0.38,0.41)	0.5 (0.48,0.52)	0.29 (0.29,0.3)	0.33 (0.3,0.35)	0.41 (0.4,0.43)	0.25 (0.23,0.27)	0.28 (0.26,0.29)	0.35 (0.33,0.37)
True	0.19 (0.18,0.21)	0.27 (0.26,0.29)	0.42 (0.41,0.44)	0.17 (0.16,0.18)	0.24 (0.22,0.25)	0.36 (0.34,0.38)	0.14 (0.13,0.15)	0.19 (0.18,0.19)	0.3 (0.28,0.3)
NKM	0.26 (0.24,0.28)	0.33 (0.31,0.35)	0.45 (0.43,0.47)	0.23 (0.22,0.24)	0.28 (0.26,0.29)	0.38 (0.37,0.4)	0.2 (0.19,0.21)	0.24 (0.22,0.25)	0.33 (0.32,0.34)
Standard	0.34 (0.31,0.37)	0.4 (0.38,0.41)	0.51 (0.49,0.53)	0.29 (0.28,0.3)	0.33 (0.3,0.35)	0.42 (0.4,0.43)	0.25 (0.23,0.27)	0.29 (0.26,0.29)	0.35 (0.33,0.37)
GEE True	0.2 (0.19,0.21)	0.28 (0.27,0.3)	0.43 (0.41,0.45)	0.17 (0.16,0.18)	0.23 (0.22,0.24)	0.35 (0.33,0.37)	0.14 (0.14,0.15)	0.19 (0.18,0.2)	0.31 (0.29,0.32)
GEE TRM	0.36 (0.33,0.38)	0.41 (0.38,0.44)	0.52 (0.49,0.54)	0.29 (0.28,0.31)	0.32 (0.3,0.34)	0.42 (0.4,0.44)	0.25 (0.24,0.26)	0.28 (0.26,0.29)	0.37 (0.35,0.38)

2.7.8 Binary Outcomes Model Specification 4 Results

Table 2.103: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE0, BO, MS4}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	95 (91,97)	94 (91,97)	94 (91,97)	95 (94,97)	96 (95,97)	95 (94,97)	96 (94,97)	97 (94,98)	96 (94,100)
Strata	95 (93,97)	93 (91,95)	95 (94,97)	94 (92,97)	95 (93,97)	94 (93,97)	96 (94,97)	96 (94,100)	95 (94,97)
True	95 (92,97)	93 (91,97)	94 (94,97)	95 (94,98)	95 (94,97)	95 (94,97)	95 (94,97)	96 (94,97)	94 (93,97)
NKM	95 (94,97)	93 (91,97)	94 (93,96)	94 (94,97)	95 (94,97)	94 (91,97)	95 (94,97)	96 (94,100)	95 (94,97)
Standard	97 (95,100)	96 (95,99)	96 (94,97)	96 (94,98)	98 (97,100)	97 (97,97)	97 (95,99)	97 (97,100)	98 (97,100)
GEE True	95 (94,97)	95 (93,97)	94 (92,96)	95 (94,96)	95 (94,96)	95 (93,96)	96 (94,98)	95 (94,96)	95 (94,97)
GEE TRM	96 (94,97)	96 (94,98)	94 (92,96)	95 (94,96)	95 (94,97)	95 (93,96)	96 (96,98)	95 (94,96)	95 (94,97)

Table 2.104: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE0, BO, MS4}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	95 (93,97)	98 (97,100)	95 (92,97)	95 (94,97)	96 (97,97)	94 (91,97)	94 (94,96)	95 (94,97)	96 (93,99)
Strata	94 (94,96)	96 (94,98)	95 (94,98)	95 (93,97)	97 (97,98)	93 (91,97)	94 (91,96)	96 (95,97)	96 (94,98)
True	93 (91,95)	96 (94,97)	94 (91,96)	94 (91,96)	97 (96,98)	91 (87,94)	93 (91,96)	96 (94,97)	95 (91,97)
NKM	93 (92,94)	96 (94,97)	95 (93,100)	94 (94,95)	96 (94,98)	91 (88,94)	92 (91,94)	94 (92,96)	95 (92,97)
Standard	98 (97,100)	98 (97,100)	96 (94,98)	98 (97,100)	98 (97,99)	96 (94,98)	97 (96,97)	98 (97,100)	97 (95,99)
GEE True	95 (94,96)	95 (94,96)	95 (94,97)	94 (92,96)	95 (94,98)	95 (94,97)	95 (94,97)	96 (95,97)	95 (92,97)
GEE TRM	96 (95,97)	96 (95,98)	96 (95,97)	96 (95,98)	97 (95,98)	96 (95,98)	96 (95,98)	97 (96,98)	96 (94,98)

Table 2.105: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR1, HTE1, BO, MS4}. Results are stratified by β_9 and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	96 (95,98)	97 (94,98)	96 (94,97)	95 (94,98)	94 (93,97)	96 (94,99)	95 (94,97)	95 (93,97)	95 (94,97)
Strata	96 (94,97)	96 (94,97)	96 (94,97)	95 (92,97)	94 (91,97)	96 (94,97)	94 (92,97)	94 (91,97)	95 (91,97)
True	96 (94,97)	95 (92,98)	95 (94,97)	94 (91,97)	94 (92,97)	96 (94,97)	93 (91,97)	94 (91,97)	95 (93,97)
NKM	96 (94,97)	96 (93,100)	96 (94,97)	95 (92,97)	94 (91,97)	96 (94,98)	94 (93,97)	94 (91,97)	95 (93,98)
Standard	98 (97,100)	98 (97,100)	97 (95,100)	97 (95,100)	96 (96,98)	97 (94,100)	97 (95,97)	97 (94,100)	96 (94,98)
GEE True	94 (93,95)	94 (92,96)	94 (92,96)	94 (93,96)	95 (93,97)	94 (93,96)	94 (93,96)	95 (94,96)	95 (93,96)
GEE TRM	94 (93,96)	94 (93,96)	94 (92,96)	95 (93,96)	95 (94,97)	94 (93,95)	95 (93,96)	96 (94,97)	95 (94,96)

Table 2.106: Mean and interquartile range for coverage rate of 95% interval estimates for each model for Simulation Set {SR2, HTE1, BO, MS4}. Results are stratified by β_9 and cluster variance.

	$\beta_9 = 0.5$			$\beta_9 = 1$		
Model	$J = 48$	$J = 72$	$J = 96$	$J = 48$	$J = 72$	$J = 96$
TRM	97 (95,98)	96 (94,97)	96 (93,98)	95 (94,98)	96 (94,98)	96 (93,99)
Strata	97 (94,98)	96 (94,97)	96 (94,98)	95 (94,97)	96 (94,97)	96 (94,97)
True	96 (94,98)	94 (92,97)	95 (92,98)	94 (94,97)	94 (92,97)	94 (91,96)
NKM	96 (94,98)	94 (93,97)	96 (93,97)	94 (94,97)	94 (92,96)	95 (93,96)
Standard	98 (97,100)	98 (97,100)	97 (95,100)	98 (97,100)	98 (97,100)	97 (96,98)
GEE True	95 (94,97)	95 (93,96)	95 (94,97)	95 (94,97)	95 (94,97)	95 (94,96)
GEE TRM	96 (95,97)	96 (95,97)	97 (95,98)	96 (95,98)	97 (95,98)	96 (95,98)

Table 2.107: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE0, BO, MS4}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.29 (0.27,0.31)	0.3 (0.29,0.32)	0.35 (0.34,0.37)	0.24 (0.22,0.25)	0.25 (0.24,0.26)	0.29 (0.27,0.3)	0.2 (0.19,0.22)	0.21 (0.2,0.23)	0.24 (0.23,0.26)
Strata	0.28 (0.26,0.29)	0.29 (0.28,0.31)	0.34 (0.33,0.36)	0.23 (0.21,0.24)	0.24 (0.23,0.25)	0.28 (0.27,0.29)	0.2 (0.18,0.21)	0.2 (0.19,0.21)	0.24 (0.23,0.25)
True	0.28 (0.27,0.3)	0.29 (0.28,0.31)	0.34 (0.33,0.36)	0.23 (0.21,0.24)	0.24 (0.23,0.25)	0.28 (0.27,0.29)	0.19 (0.18,0.2)	0.2 (0.19,0.21)	0.24 (0.23,0.24)
NKM	0.28 (0.26,0.29)	0.29 (0.28,0.3)	0.34 (0.33,0.36)	0.22 (0.21,0.24)	0.24 (0.23,0.25)	0.28 (0.26,0.28)	0.19 (0.18,0.2)	0.2 (0.19,0.21)	0.24 (0.23,0.24)
Standard	0.33 (0.31,0.35)	0.34 (0.32,0.36)	0.38 (0.37,0.39)	0.27 (0.24,0.28)	0.28 (0.26,0.29)	0.31 (0.29,0.32)	0.23 (0.22,0.24)	0.24 (0.22,0.25)	0.26 (0.25,0.28)
GEE True	0.29 (0.27,0.3)	0.31 (0.28,0.32)	0.36 (0.33,0.38)	0.23 (0.21,0.24)	0.25 (0.23,0.26)	0.29 (0.27,0.3)	0.2 (0.18,0.2)	0.21 (0.19,0.22)	0.25 (0.23,0.26)
GEE TRM	0.29 (0.28,0.31)	0.31 (0.29,0.33)	0.36 (0.34,0.38)	0.23 (0.22,0.25)	0.25 (0.23,0.27)	0.29 (0.27,0.3)	0.2 (0.19,0.21)	0.22 (0.2,0.23)	0.25 (0.23,0.26)

Table 2.108: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE0, BO, MS4}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.24 (0.24,0.24)	0.25 (0.25,0.25)	0.28 (0.27,0.28)	0.19 (0.19,0.19)	0.2 (0.2,0.2)	0.23 (0.22,0.23)	0.17 (0.16,0.17)	0.17 (0.17,0.18)	0.2 (0.2,0.2)
Strata	0.23 (0.23,0.23)	0.24 (0.24,0.24)	0.27 (0.27,0.27)	0.18 (0.18,0.19)	0.2 (0.19,0.2)	0.22 (0.22,0.22)	0.16 (0.16,0.16)	0.17 (0.17,0.17)	0.19 (0.19,0.19)
True	0.21 (0.21,0.22)	0.23 (0.23,0.23)	0.26 (0.26,0.26)	0.17 (0.17,0.18)	0.19 (0.18,0.19)	0.21 (0.21,0.21)	0.15 (0.15,0.15)	0.16 (0.16,0.16)	0.18 (0.18,0.19)
NKM	0.21 (0.21,0.22)	0.23 (0.23,0.23)	0.26 (0.26,0.26)	0.17 (0.17,0.18)	0.19 (0.18,0.19)	0.21 (0.21,0.21)	0.15 (0.15,0.15)	0.16 (0.16,0.16)	0.18 (0.18,0.19)
Standard	0.27 (0.26,0.27)	0.27 (0.27,0.28)	0.29 (0.29,0.3)	0.22 (0.21,0.22)	0.22 (0.22,0.23)	0.24 (0.24,0.25)	0.19 (0.18,0.19)	0.19 (0.19,0.2)	0.21 (0.21,0.21)
GEE True	0.22 (0.21,0.22)	0.23 (0.23,0.24)	0.27 (0.27,0.27)	0.17 (0.17,0.18)	0.19 (0.18,0.19)	0.22 (0.21,0.22)	0.15 (0.15,0.15)	0.16 (0.16,0.16)	0.19 (0.18,0.19)
GEE TRM	0.24 (0.23,0.24)	0.25 (0.25,0.25)	0.28 (0.28,0.28)	0.19 (0.19,0.19)	0.2 (0.2,0.2)	0.23 (0.22,0.23)	0.16 (0.16,0.17)	0.17 (0.17,0.18)	0.2 (0.19,0.2)

Table 2.109: Mean and interquartile range of interval widths for each model for Simulation Set {SR1, HTE1, BO, MS4}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.24 (0.22,0.27)	0.26 (0.24,0.27)	0.3 (0.28,0.33)	0.2 (0.17,0.22)	0.2 (0.18,0.22)	0.24 (0.22,0.26)	0.17 (0.15,0.19)	0.17 (0.16,0.19)	0.21 (0.19,0.23)
Strata	0.24 (0.22,0.25)	0.25 (0.24,0.27)	0.3 (0.28,0.32)	0.19 (0.17,0.21)	0.2 (0.18,0.22)	0.23 (0.22,0.25)	0.17 (0.15,0.18)	0.17 (0.16,0.19)	0.2 (0.19,0.22)
True	0.24 (0.22,0.25)	0.25 (0.24,0.26)	0.3 (0.28,0.31)	0.19 (0.17,0.2)	0.2 (0.18,0.21)	0.23 (0.22,0.25)	0.16 (0.15,0.18)	0.17 (0.16,0.18)	0.2 (0.19,0.22)
NKM	0.24 (0.22,0.25)	0.25 (0.24,0.26)	0.3 (0.28,0.32)	0.19 (0.17,0.21)	0.2 (0.19,0.21)	0.23 (0.22,0.25)	0.16 (0.15,0.18)	0.17 (0.16,0.18)	0.2 (0.19,0.22)
Standard	0.28 (0.25,0.3)	0.29 (0.27,0.31)	0.33 (0.31,0.35)	0.22 (0.2,0.25)	0.23 (0.2,0.25)	0.26 (0.24,0.28)	0.19 (0.18,0.21)	0.2 (0.18,0.22)	0.22 (0.2,0.24)
GEE True	0.24 (0.22,0.26)	0.25 (0.23,0.28)	0.3 (0.28,0.33)	0.19 (0.17,0.21)	0.2 (0.19,0.22)	0.24 (0.22,0.26)	0.16 (0.15,0.18)	0.17 (0.16,0.19)	0.2 (0.19,0.22)
GEE TRM	0.24 (0.21,0.27)	0.26 (0.23,0.28)	0.3 (0.27,0.33)	0.19 (0.17,0.21)	0.21 (0.19,0.23)	0.24 (0.22,0.26)	0.17 (0.15,0.18)	0.18 (0.16,0.2)	0.21 (0.19,0.23)

Table 2.110: Mean and interquartile range of interval widths for each model for Simulation Set {SR2, HTE1, BO, MS4}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.22 (0.22,0.23)	0.24 (0.23,0.24)	0.27 (0.26,0.27)	0.18 (0.18,0.19)	0.19 (0.19,0.2)	0.22 (0.21,0.22)	0.16 (0.15,0.16)	0.17 (0.16,0.17)	0.19 (0.19,0.19)
Strata	0.22 (0.22,0.22)	0.23 (0.23,0.24)	0.26 (0.26,0.26)	0.18 (0.18,0.18)	0.19 (0.19,0.19)	0.22 (0.21,0.22)	0.16 (0.15,0.16)	0.16 (0.16,0.17)	0.19 (0.18,0.19)
True	0.2 (0.2,0.21)	0.22 (0.22,0.22)	0.25 (0.25,0.26)	0.17 (0.16,0.17)	0.18 (0.18,0.18)	0.21 (0.2,0.21)	0.14 (0.14,0.15)	0.15 (0.15,0.16)	0.18 (0.17,0.18)
NKM	0.2 (0.2,0.21)	0.22 (0.22,0.22)	0.25 (0.25,0.25)	0.17 (0.16,0.17)	0.18 (0.18,0.18)	0.21 (0.2,0.21)	0.14 (0.14,0.15)	0.15 (0.15,0.16)	0.18 (0.18,0.18)
Standard	0.25 (0.25,0.26)	0.26 (0.25,0.26)	0.28 (0.28,0.29)	0.21 (0.2,0.21)	0.21 (0.21,0.22)	0.23 (0.23,0.24)	0.18 (0.17,0.18)	0.18 (0.18,0.19)	0.2 (0.2,0.2)
GEE True	0.2 (0.2,0.21)	0.22 (0.21,0.23)	0.26 (0.25,0.26)	0.16 (0.16,0.17)	0.18 (0.17,0.18)	0.21 (0.2,0.21)	0.14 (0.14,0.14)	0.15 (0.15,0.16)	0.18 (0.17,0.18)
GEE TRM	0.22 (0.22,0.23)	0.24 (0.23,0.24)	0.27 (0.26,0.27)	0.18 (0.18,0.18)	0.19 (0.19,0.2)	0.22 (0.21,0.22)	0.16 (0.15,0.16)	0.16 (0.16,0.17)	0.19 (0.18,0.19)

Table 2.111: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE0, BO, MS4}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.07 (0.06,0.07)	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.06 (0.05,0.06)	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.05 (0.04,0.05)
Strata	0.06 (0.05,0.06)	0.06 (0.06,0.07)	0.07 (0.06,0.07)	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.06 (0.05,0.06)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.04,0.05)
True	0.06 (0.05,0.06)	0.06 (0.06,0.06)	0.07 (0.07,0.07)	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.06 (0.05,0.06)	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.04,0.05)
NKM	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.07 (0.06,0.07)	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.06 (0.05,0.06)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.04,0.05)
Standard	0.06 (0.05,0.06)	0.06 (0.06,0.07)	0.07 (0.07,0.07)	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.06 (0.05,0.06)	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.05 (0.04,0.05)
GEE True	0.05 (0.05,0.06)	0.06 (0.06,0.06)	0.07 (0.07,0.07)	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.06 (0.05,0.06)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.05 (0.04,0.05)
GEE TRM	0.06 (0.05,0.06)	0.06 (0.05,0.06)	0.07 (0.07,0.07)	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.06 (0.05,0.06)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.05)

Table 2.112: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE0, BO, MS4}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.04 (0.04,0.05)	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.04 (0.03,0.04)	0.03 (0.03,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.04)
Strata	0.04 (0.04,0.05)	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.04 (0.03,0.04)	0.03 (0.03,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.04)	0.03 (0.03,0.04)	0.04 (0.04,0.04)
True	0.04 (0.05,0.05)	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.04 (0.03,0.04)	0.03 (0.03,0.03)	0.05 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.04)
NKM	0.04 (0.04,0.05)	0.04 (0.04,0.04)	0.05 (0.05,0.06)	0.04 (0.03,0.04)	0.03 (0.03,0.03)	0.05 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.04)
Standard	0.04 (0.04,0.05)	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.04 (0.03,0.04)	0.03 (0.03,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.04)
GEE True	0.04 (0.04,0.05)	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.04 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.04,0.04)
GEE TRM	0.04 (0.04,0.05)	0.05 (0.05,0.05)	0.05 (0.05,0.05)	0.04 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.04,0.04)

Table 2.113: Mean and interquartile range of absolute biases for each model for Simulation Set {SR1, HTE1, BO, MS4}. Results are stratified by number of clusters and cluster variance.

	24 Clusters			36 Clusters			48 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.04 (0.04,0.05)	0.05 (0.04,0.05)	0.06 (0.05,0.06)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.04 (0.04,0.05)	0.03 (0.03,0.04)	0.03 (0.03,0.04)	0.04 (0.03,0.05)
Strata	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.06 (0.05,0.06)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.04,0.05)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
True	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.06 (0.05,0.06)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.04,0.05)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
NKM	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.06 (0.05,0.06)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.04,0.05)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
Standard	0.05 (0.04,0.05)	0.05 (0.05,0.06)	0.06 (0.05,0.07)	0.04 (0.03,0.05)	0.04 (0.04,0.05)	0.05 (0.04,0.05)	0.04 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
GEE True	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.06 (0.05,0.07)	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.04)	0.03 (0.03,0.04)	0.04 (0.04,0.05)
GEE TRM	0.05 (0.04,0.05)	0.05 (0.04,0.05)	0.06 (0.05,0.07)	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.03 (0.03,0.04)	0.03 (0.03,0.04)	0.04 (0.04,0.05)

Table 2.114: Mean and interquartile range of absolute biases for each model for Simulation Set {SR2, HTE1, BO, MS4}. Results are stratified by number of clusters and cluster variance.

	48 Clusters			72 Clusters			96 Clusters		
Model	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
TRM	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.05 (0.05,0.06)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.03,0.04)
Strata	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)
True	0.04 (0.03,0.04)	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)
NKM	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.03 (0.03,0.04)
Standard	0.04 (0.04,0.04)	0.05 (0.04,0.05)	0.05 (0.05,0.05)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.03 (0.03,0.04)
GEE True	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)
GEE TRM	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)

2.7.9 GEE Bias-Adjusted Methods vs. No Bias Adjustment

Table 2.115: Mean and interquartile range for coverage rate of 95% interval estimates for GEE models with different Bias-Adjustments for Simulation {SR1, HTE0, CO, MS1}. "M&D" represents Mancl and DeRouen (2001) bias-adjustment, "K&C" represents Kauermann and Carroll (2001) bias-adjustment, and "None" represents GEE results with no bias-adjustment implemented.

		24 Clusters			36 Clusters			48 Clusters		
	Bias-Adjustment	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
GEE True	M&D	95 (94,96)	96 (96,97)	95 (94,96)	95 (94,96)	96 (94,96)	95 (93,97)	95 (94,96)	95 (94,96)	95 (94,96)
	K&C	93 (92,94)	95 (94,96)	93 (92,95)	94 (93,96)	95 (94,96)	94 (92,95)	95 (94,96)	95 (93,96)	94 (93,96)
	None	92 (89,94)	92 (91,94)	90 (85,94)	94 (91,97)	94 (91,97)	92 (89,96)	94 (92,97)	94 (91,97)	94 (91,97)
GEE TRM	M&D	98 (96,99)	98 (96,99)	96 (94,98)	98 (97,100)	97 (96,98)	95 (94,97)	99 (98,100)	97 (96,99)	96 (94,97)
	K&C	97 (96,99)	97 (96,98)	95 (93,97)	98 (96,100)	96 (95,97)	95 (93,97)	98 (98,100)	97 (95,98)	95 (94,97)
	None	98 (95,100)	96 (94,98)	93 (91,95)	97 (96,100)	97 (94,100)	94 (91,97)	98 (97,100)	97 (94,100)	95 (93,97)

Table 2.116: Mean and interquartile range for coverage rate of 95% interval estimates for GEE models with different Bias-Adjustments for Simulation Set {SR2, HTE0, CO, MS1}. "M&D" represents Mancl and DeRouen (2001) bias-adjustment, "K&C" represents Kauermann and Carroll (2001) bias-adjustment, and "None" represents GEE results with no bias-adjustment implemented.

		48 Clusters			72 Clusters			96 Clusters		
	Bias-Adjustment	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
GEE True	M&D	96 (94,98)	95 (94,96)	95 (94,96)	96 (94,97)	95 (95,97)	95 (94,97)	95 (94,97)	95 (94,97)	95 (94,96)
	K&C	95 (93,97)	94 (93,95)	94 (93,95)	95 (94,96)	95 (94,96)	94 (93,96)	95 (93,96)	95 (93,97)	95 (93,97)
	None	93 (89,96)	93 (92,94)	93 (91,95)	94 (91,97)	94 (91,95)	93 (91,95)	94 (92,97)	94 (93,96)	96 (93,98)
GEE TRM	M&D	99 (98,100)	98 (96,99)	96 (95,97)	99 (98,100)	98 (97,98)	96 (95,98)	99 (98,100)	98 (96,98)	96 (95,98)
	K&C	99 (98,99)	97 (96,99)	95 (94,97)	99 (98,100)	97 (97,98)	96 (94,97)	99 (98,100)	97 (96,98)	96 (95,98)
	None	98 (97,99)	96 (95,97)	95 (94,97)	98 (97,98)	97 (95,98)	95 (94,96)	97 (95,99)	97 (96,98)	97 (94,100)

Table 2.117: Mean and interquartile range for coverage rate of 95% interval estimates for GEE models with different Bias-Adjustments for Simulation Set {SR1, HTE0, BO, MS1}. "M&D" represents Mancl and DeRouen (2001) bias-adjustment, "K&C" represents Kauermann and Carroll (2001) bias-adjustment, and "None" represents GEE results with no bias-adjustment implemented.

		24 Clusters			36 Clusters			48 Clusters		
	Bias-Adjustment	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
GEE True	M&D	94 (93,96)	95 (92,96)	95 (93,96)	95 (94,96)	95 (93,96)	95 (94,96)	95 (94,96)	95 (94,97)	95 (93,97)
	K&C	93 (92,95)	93 (90,96)	93 (92,94)	94 (93,95)	94 (92,96)	94 (93,95)	94 (93,95)	95 (94,96)	94 (92,96)
	None	93 (91,94)	91 (88,94)	91 (87,96)	91 (88,94)	92 (88,94)	93 (91,97)	94 (91,96)	92 (88,94)	93 (91,96)
GEE TRM	M&D	95 (94,96)	95 (93,98)	95 (93,96)	95 (94,97)	95 (94,96)	95 (94,96)	96 (94,97)	96 (94,97)	95 (93,97)
	K&C	94 (93,95)	94 (92,97)	94 (92,95)	95 (94,96)	95 (94,96)	94 (93,95)	95 (94,96)	95 (94,97)	94 (92,97)
	None	94 (92,97)	92 (90,94)	92 (88,97)	93 (91,95)	93 (91,97)	94 (91,97)	96 (94,97)	93 (88,97)	94 (91,97)

Table 2.118: Mean and interquartile range for coverage rate of 95% interval estimates for GEE models with different Bias-Adjustments for Simulation Set {SR2, HTE0, BO, MS1}. Results are stratified by number of clusters and cluster variance. "M&D" represents Mancl and DeRouen (2001) bias-adjustment, "K&C" represents Kauermann and Carroll (2001) bias-adjustment, and "None" represents GEE results with no bias-adjustment implemented.

		48 Clusters			72 Clusters			96 Clusters		
	Bias-Adjustment	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
GEE True	M&D	95 (94,96)	95 (94,97)	96 (94,97)	95 (94,97)	95 (94,96)	95 (94,96)	95 (94,96)	95 (94,97)	95 (94,96)
	K&C	95 (93,96)	94 (93,96)	95 (93,96)	95 (93,96)	94 (93,96)	94 (92,95)	94 (93,96)	95 (93,96)	94 (93,96)
	None	94 (91,96)	93 (90,97)	93 (90,95)	94 (94,96)	94 (93,97)	94 (93,97)	93 (90,95)	95 (94,97)	94 (91,97)
GEE TRM	M&D	97 (96,98)	97 (96,98)	97 (96,98)	97 (96,99)	96 (95,98)	96 (94,97)	97 (96,98)	97 (95,98)	96 (95,97)
	K&C	97 (95,98)	97 (96,98)	96 (95,97)	97 (96,98)	96 (95,98)	95 (94,96)	97 (95,98)	97 (95,98)	96 (95,97)
	None	96 (95,97)	95 (93,98)	95 (94,97)	97 (96,100)	97 (94,100)	95 (94,97)	95 (93,97)	96 (95,97)	95 (94,97)

Table 2.119: Mean and interquartile range for interval widths for GEE models with different Bias-Adjustments for Simulation Set {SR1, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance. "M&D" represents Mancl and DeRouen (2001) bias-adjustment, "K&C" represents Kauermann and Carroll (2001) bias-adjustment, and "None" represents GEE results with no bias-adjustment implemented.

		24 Clusters			36 Clusters			48 Clusters		
	Bias-Adjustment	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
GEE True	M&D	1.05 (1.04,1.07)	1.74 (1.72,1.75)	2.97 (2.94,3)	0.84 (0.83,0.85)	1.4 (1.38,1.41)	2.37 (2.35,2.39)	0.72 (0.72,0.73)	1.19 (1.18,1.19)	2.03 (2.01,2.05)
	K&C	0.98 (0.97,0.99)	1.62 (1.61,1.64)	2.78 (2.75,2.8)	0.8 (0.8,0.81)	1.34 (1.32,1.35)	2.27 (2.25,2.29)	0.7 (0.69,0.7)	1.15 (1.14,1.15)	1.97 (1.95,1.98)
	None	0.92 (0.9,0.93)	1.52 (1.49,1.55)	2.58 (2.53,2.62)	0.77 (0.77,0.78)	1.27 (1.25,1.28)	2.18 (2.15,2.2)	0.68 (0.67,0.69)	1.12 (1.11,1.13)	1.9 (1.88,1.92)
GEE TRM	M&D	1.43 (1.13,1.75)	1.97 (1.77,2.18)	3.08 (2.95,3.2)	1.14 (0.91,1.39)	1.6 (1.44,1.78)	2.48 (2.4,2.61)	0.99 (0.79,1.2)	1.37 (1.23,1.53)	2.13 (2.06,2.23)
	K&C	1.36 (1.08,1.67)	1.88 (1.69,2.09)	2.95 (2.82,3.06)	1.11 (0.89,1.35)	1.55 (1.4,1.73)	2.41 (2.33,2.53)	0.96 (0.77,1.17)	1.34 (1.2,1.5)	2.09 (2.02,2.18)
	None	1.3 (1.04,1.56)	1.82 (1.62,2.02)	2.79 (2.69,2.89)	1.08 (0.86,1.31)	1.5 (1.35,1.66)	2.34 (2.26,2.42)	0.94 (0.76,1.15)	1.32 (1.19,1.46)	2.03 (1.97,2.1)

Table 2.120: Mean and interquartile range for interval widths for GEE models with different Bias-Adjustments for Simulation Set {SR2, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance. "M&D" represents Mancl and DeRouen (2001) bias-adjustment, "K&C" represents Kauermann and Carroll (2001) bias-adjustment, and "None" represents GEE results with no bias-adjustment implemented.

		48 Clusters			72 Clusters			96 Clusters		
	Bias-Adjustment	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
GEE True	M&D	0.73 (0.72,0.73)	1.21 (1.21,1.23)	2.07 (2.05,2.1)	0.59 (0.59,0.59)	0.97 (0.97,0.98)	1.66 (1.65,1.67)	0.51 (0.51,0.51)	0.84 (0.83,0.84)	1.43 (1.42,1.44)
	K&C	0.7 (0.69,0.7)	1.16 (1.16,1.17)	1.97 (1.95,2)	0.57 (0.57,0.57)	0.95 (0.94,0.95)	1.61 (1.6,1.62)	0.5 (0.49,0.5)	0.82 (0.82,0.82)	1.4 (1.39,1.4)
	None	0.67 (0.66,0.67)	1.1 (1.1,1.11)	1.9 (1.89,1.92)	0.56 (0.55,0.56)	0.92 (0.92,0.93)	1.56 (1.55,1.57)	0.48 (0.48,0.49)	0.8 (0.8,0.8)	1.37 (1.36,1.38)
GEE TRM	M&D	1.23 (1.1,1.37)	1.56 (1.45,1.69)	2.26 (2.19,2.34)	1 (0.89,1.12)	1.27 (1.18,1.37)	1.84 (1.79,1.91)	0.86 (0.77,0.97)	1.09 (1.02,1.18)	1.59 (1.54,1.64)
	K&C	1.2 (1.07,1.34)	1.53 (1.42,1.66)	2.2 (2.13,2.28)	0.98 (0.88,1.11)	1.25 (1.16,1.35)	1.81 (1.76,1.88)	0.85 (0.76,0.96)	1.08 (1.01,1.16)	1.57 (1.52,1.62)
	None	1.08 (1.03,1.14)	1.39 (1.36,1.4)	2.1 (2.08,2.11)	0.9 (0.86,0.95)	1.16 (1.14,1.17)	1.73 (1.71,1.75)	0.76 (0.74,0.78)	1.01 (0.99,1.01)	1.51 (1.49,1.53)

Table 2.121: Mean and interquartile range for interval widths for GEE models with different Bias-Adjustments for Simulation Set {SR1, HTE0, BO, MS1}. Results are stratified by number of clusters and cluster variance. "M&D" represents Mancl and DeRouen (2001) bias-adjustment, "K&C" represents Kauermann and Carroll (2001) bias-adjustment, and "None" represents GEE results with no bias-adjustment implemented.

		24 Clusters			36 Clusters			48 Clusters		
	Bias-Adjustment	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
GEE True	M&D	0.24 (0.22,0.25)	0.27 (0.25,0.28)	0.33 (0.31,0.34)	0.19 (0.18,0.2)	0.21 (0.2,0.22)	0.26 (0.24,0.28)	0.17 (0.15,0.17)	0.18 (0.17,0.19)	0.23 (0.21,0.23)
	K&C	0.22 (0.21,0.24)	0.25 (0.23,0.26)	0.31 (0.29,0.32)	0.18 (0.17,0.19)	0.21 (0.19,0.22)	0.25 (0.23,0.26)	0.16 (0.15,0.17)	0.18 (0.16,0.19)	0.22 (0.2,0.23)
	None	0.21 (0.19,0.22)	0.23 (0.21,0.25)	0.28 (0.26,0.3)	0.17 (0.16,0.19)	0.19 (0.18,0.2)	0.24 (0.22,0.26)	0.15 (0.14,0.16)	0.17 (0.15,0.18)	0.21 (0.19,0.22)
GEE TRM	M&D	0.25 (0.23,0.26)	0.27 (0.26,0.29)	0.33 (0.31,0.35)	0.2 (0.18,0.21)	0.22 (0.2,0.24)	0.27 (0.25,0.28)	0.17 (0.16,0.18)	0.19 (0.17,0.2)	0.23 (0.21,0.24)
	K&C	0.24 (0.22,0.25)	0.26 (0.24,0.28)	0.32 (0.3,0.34)	0.19 (0.18,0.21)	0.21 (0.19,0.23)	0.26 (0.24,0.28)	0.17 (0.15,0.18)	0.18 (0.17,0.2)	0.22 (0.21,0.24)
	None	0.22 (0.2,0.24)	0.24 (0.22,0.26)	0.3 (0.28,0.32)	0.18 (0.17,0.2)	0.2 (0.18,0.22)	0.25 (0.23,0.26)	0.16 (0.15,0.18)	0.18 (0.16,0.19)	0.22 (0.2,0.23)

Table 2.122: Mean and interquartile range for interval widths for GEE models with different Bias-Adjustments for Simulation Set {SR2, HTE0, BO, MS1}. Results are stratified by number of clusters and cluster variance. "M&D" represents Mancl and DeRouen (2001) bias-adjustment, "K&C" represents Kauermann and Carroll (2001) bias-adjustment, and "None" represents GEE results with no bias-adjustment implemented.

		48 Clusters			72 Clusters			96 Clusters		
	Bias-Adjustment	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
GEE True	M&D	0.18 (0.18,0.18)	0.2 (0.2,0.2)	0.25 (0.24,0.25)	0.15 (0.14,0.15)	0.16 (0.16,0.16)	0.2 (0.2,0.2)	0.12 (0.12,0.13)	0.14 (0.14,0.14)	0.17 (0.17,0.17)
	K&C	0.17 (0.17,0.18)	0.19 (0.19,0.2)	0.24 (0.23,0.24)	0.14 (0.14,0.14)	0.16 (0.15,0.16)	0.19 (0.19,0.2)	0.12 (0.12,0.12)	0.14 (0.13,0.14)	0.17 (0.17,0.17)
	None	0.17 (0.17,0.17)	0.19 (0.19,0.19)	0.23 (0.23,0.23)	0.14 (0.14,0.14)	0.16 (0.15,0.16)	0.19 (0.19,0.19)	0.12 (0.12,0.12)	0.14 (0.13,0.14)	0.17 (0.17,0.17)
GEE TRM	M&D	0.21 (0.2,0.21)	0.22 (0.22,0.23)	0.26 (0.26,0.27)	0.17 (0.16,0.17)	0.18 (0.18,0.19)	0.21 (0.21,0.22)	0.14 (0.14,0.15)	0.16 (0.15,0.16)	0.18 (0.18,0.19)
	K&C	0.2 (0.2,0.21)	0.22 (0.21,0.22)	0.26 (0.25,0.26)	0.16 (0.16,0.17)	0.18 (0.17,0.18)	0.21 (0.21,0.21)	0.14 (0.14,0.15)	0.15 (0.15,0.16)	0.18 (0.18,0.18)
	None	0.19 (0.19,0.2)	0.21 (0.21,0.21)	0.25 (0.25,0.25)	0.16 (0.16,0.16)	0.17 (0.17,0.17)	0.21 (0.2,0.21)	0.14 (0.14,0.14)	0.15 (0.15,0.15)	0.18 (0.18,0.18)

Table 2.123: Mean and interquartile range for absolute biases for GEE models with different Bias-Adjustments for Simulation Set {SR1, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance. "M&D" represents Mancl and DeRouen (2001) bias-adjustment, "K&C" represents Kauermann and Carroll (2001) bias-adjustment, and "None" represents GEE results with no bias-adjustment implemented.

		24 Clusters			36 Clusters			48 Clusters		
	Bias-Adjustment	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
GEE True	M&D	0.2 (0.19,0.21)	0.33 (0.31,0.34)	0.58 (0.55,0.6)	0.16 (0.15,0.17)	0.27 (0.25,0.28)	0.48 (0.45,0.5)	0.14 (0.13,0.15)	0.23 (0.21,0.25)	0.4 (0.38,0.43)
	K&C	0.2 (0.19,0.21)	0.33 (0.31,0.34)	0.58 (0.55,0.6)	0.16 (0.15,0.17)	0.27 (0.25,0.28)	0.48 (0.45,0.5)	0.14 (0.13,0.15)	0.23 (0.21,0.25)	0.4 (0.38,0.43)
	None	0.2 (0.19,0.22)	0.34 (0.32,0.37)	0.59 (0.54,0.64)	0.16 (0.15,0.17)	0.26 (0.25,0.28)	0.48 (0.44,0.52)	0.14 (0.13,0.15)	0.24 (0.22,0.26)	0.39 (0.37,0.42)
GEE TRM	M&D	0.22 (0.2,0.24)	0.33 (0.32,0.35)	0.58 (0.55,0.61)	0.18 (0.16,0.19)	0.28 (0.27,0.28)	0.48 (0.46,0.51)	0.15 (0.14,0.16)	0.24 (0.22,0.26)	0.41 (0.39,0.43)
	K&C	0.22 (0.2,0.24)	0.33 (0.32,0.35)	0.58 (0.55,0.61)	0.18 (0.16,0.19)	0.28 (0.27,0.28)	0.48 (0.46,0.51)	0.15 (0.14,0.16)	0.24 (0.22,0.26)	0.41 (0.39,0.43)
	None	0.22 (0.2,0.23)	0.35 (0.32,0.38)	0.59 (0.54,0.62)	0.17 (0.15,0.19)	0.27 (0.26,0.29)	0.48 (0.44,0.52)	0.15 (0.14,0.17)	0.25 (0.23,0.26)	0.39 (0.38,0.41)

Table 2.124: Mean and interquartile range for absolute biases for GEE models with different Bias-Adjustments for Simulation Set {SR2, HTE0, CO, MS1}. Results are stratified by number of clusters and cluster variance. "M&D" represents Mancl and DeRouen (2001) bias-adjustment, "K&C" represents Kauermann and Carroll (2001) bias-adjustment, and "None" represents GEE results with no bias-adjustment implemented.

		48 Clusters			72 Clusters			96 Clusters		
	Bias-Adjustment	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1$	$\sigma_c^2 = 3$
GEE True	M&D	0.14 (0.14,0.15)	0.24 (0.23,0.25)	0.4 (0.38,0.42)	0.12 (0.11,0.12)	0.19 (0.19,0.2)	0.34 (0.32,0.36)	0.1 (0.09,0.11)	0.17 (0.16,0.18)	0.29 (0.28,0.31)
	K&C	0.14 (0.14,0.15)	0.24 (0.23,0.25)	0.4 (0.38,0.42)	0.12 (0.11,0.12)	0.19 (0.19,0.2)	0.34 (0.31,0.36)	0.1 (0.09,0.11)	0.17 (0.16,0.18)	0.29 (0.28,0.31)
	None	0.14 (0.13,0.15)	0.23 (0.21,0.25)	0.41 (0.39,0.43)	0.12 (0.11,0.13)	0.19 (0.18,0.2)	0.34 (0.33,0.35)	0.1 (0.1,0.11)	0.17 (0.15,0.18)	0.27 (0.23,0.31)
GEE TRM	M&D	0.18 (0.17,0.19)	0.27 (0.26,0.28)	0.42 (0.4,0.43)	0.15 (0.14,0.16)	0.22 (0.21,0.23)	0.35 (0.33,0.38)	0.13 (0.12,0.14)	0.19 (0.18,0.2)	0.3 (0.29,0.32)
	K&C	0.18 (0.17,0.19)	0.27 (0.26,0.28)	0.42 (0.39,0.44)	0.15 (0.14,0.16)	0.22 (0.21,0.23)	0.35 (0.33,0.38)	0.13 (0.12,0.14)	0.19 (0.18,0.2)	0.3 (0.29,0.32)
	None	0.18 (0.17,0.19)	0.26 (0.25,0.28)	0.43 (0.39,0.46)	0.15 (0.14,0.17)	0.21 (0.2,0.22)	0.35 (0.32,0.37)	0.13 (0.12,0.14)	0.19 (0.18,0.2)	0.28 (0.25,0.32)

Table 2.125: Mean and interquartile range for absolute biases for GEE models with different Bias-Adjustments for Simulation Set {SR1, HTE0, BO, MS1}. Results are stratified by number of clusters and cluster variance. "M&D" represents Mancl and DeRouen (2001) bias-adjustment, "K&C" represents Kauermann and Carroll (2001) bias-adjustment, and "None" represents GEE results with no bias-adjustment implemented.

		24 Clusters			36 Clusters			48 Clusters		
	Bias-Adjustment	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
GEE True	M&D	0.05 (0.04,0.05)	0.05 (0.05,0.06)	0.06 (0.06,0.07)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.06)	0.03 (0.03,0.03)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
	K&C	0.05 (0.04,0.05)	0.05 (0.05,0.06)	0.06 (0.06,0.07)	0.04 (0.04,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.06)	0.03 (0.03,0.03)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
	None	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.06 (0.05,0.07)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
GEE TRM	M&D	0.05 (0.04,0.05)	0.05 (0.05,0.06)	0.06 (0.06,0.07)	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
	K&C	0.05 (0.04,0.05)	0.05 (0.05,0.06)	0.06 (0.06,0.07)	0.04 (0.04,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.03 (0.03,0.04)	0.04 (0.03,0.04)	0.04 (0.04,0.05)
	None	0.04 (0.04,0.05)	0.05 (0.05,0.05)	0.06 (0.05,0.07)	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.06)	0.03 (0.03,0.03)	0.04 (0.03,0.04)	0.04 (0.04,0.05)

Table 2.126: Mean and interquartile range for absolute biases for GEE models with different Bias-Adjustments for Simulation Set {SR2, HTE0, BO, MS1}. Results are stratified by number of clusters and cluster variance. "M&D" represents Mancl and DeRouen (2001) bias-adjustment, "K&C" represents Kauermann and Carroll (2001) bias-adjustment, and "None" represents GEE results with no bias-adjustment implemented.

		48 Clusters			72 Clusters			96 Clusters		
	Bias-Adjustment	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$	$\sigma_c^2 = 1/3$	$\sigma_c^2 = 1/2$	$\sigma_c^2 = 1$
GEE True	M&D	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.04,0.04)	0.02 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)
	K&C	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.04,0.04)	0.02 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)
	None	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.04,0.04)	0.02 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.03)
GEE TRM	M&D	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)
	K&C	0.04 (0.03,0.04)	0.04 (0.04,0.04)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.03)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.03 (0.03,0.04)
	None	0.04 (0.03,0.04)	0.04 (0.04,0.05)	0.05 (0.05,0.05)	0.03 (0.03,0.03)	0.03 (0.03,0.04)	0.04 (0.04,0.04)	0.03 (0.02,0.03)	0.03 (0.03,0.03)	0.04 (0.03,0.04)

Chapter 3

Designing an Adaptive Trial to Address Noncompliance

3.1 Introduction

Randomized clinical trials can be placed on a continuum between explanatory and pragmatic. (Schwartz and Lellouch, 1967; Gamerman et al., 2019). Trials that are close to the explanatory end aim to examine the magnitude and effectiveness of an intervention under ideal conditions. Explanatory trials are designed to adjust for factors that increase the heterogeneity of the intervention to maximize the standardized intervention effect. Trials that are closer to the pragmatic end aim to investigate if an intervention is effective in real-world settings (Roland and Torgerson, 1998; Godwin et al., 2003; Patsopoulos, 2011; Ware and Hamel, 2011). These trials are usually implemented in clinical settings in which the patients are expected to receive the intervention rather than specific research sites (Gamerman et al., 2019). They comprise of heterogeneous population that the intervention is intended for and they include outcomes that inform optimal healthcare decisions (Gamerman et al., 2019). In addition, interventions in pragmatic trials are commonly complex and include multiple components (Ford and Norrie, 2016). Some of these components aim to ensure adherence and compliance with the active intervention and some components are the active ingredients that will improve efficacy if administered.

The estimand of interest in pragmatic trials commonly follows the intention-to-treat (ITT) principle (Zuidgeest et al., 2017; Hernán and Robins, 2017). This is because the goal is to estimate the effects of interventions in practice, where participants are subject to different degrees of compliance with the intervention. When participants do not comply with the assigned intervention, the intention-to-treat principle may not provide an unbiased estimate of the effectiveness and harms of the intervention if all participants would have complied with the assigned intervention (Ten Have et al., 2008; Hernán and Robins, 2017). Moreover, when the compliance in the trial differs from the compliance that would be observed in the population, the ITT effects estimated in the trial may not represent the effects that would be observed in the population (Hernán and Robins, 2017).

With a binary intervention, noncompliance can be one- or two-sided (Imbens and Rubin, 2015). In a one-sided noncompliance scenario, individuals that are assigned to the treatment-arm may not comply with the treatment, while those in the control-arm do not have access to the intervention and adhere to the assigned protocol. In a two-sided noncompliance scenario, individuals may not adhere to either protocol. We will concentrate on trials that can only experience one sided noncompliance.

As noncompliance increases, summary statistics of the effect of an intervention will tend towards the null, requiring larger sample sizes to achieve a pre-defined power. Common procedures to adjust for noncompliance at the design stage involve increasing the full compliance required sample size by a function of the expected noncompliance proportion (Hickey et al., 2018; Lachin and Foulkes, 1986; Wittes, 2002). However, these procedures may lead to studies with insufficient power when the expected non-compliance rate is higher than expected.

In the analysis stage, randomized trials with one-sided noncompliance can also be considered as Randomized Encouragement Designs (Imbens and Rubin, 2015). Considering randomization as the instrument, instrumental variable analysis can be used to estimate the complier average causal effect (CACE) (Angrist et al., 1996). CACE estimates the effect of an intervention within a sub-population of individuals who complied with their intervention assignment (Imbens and Rubin, 1997). Imbens and Rubin (1997) developed a Bayesian framework to estimate the CACE. Frangakis and Rubin (2002) provide an extended framework to address two-sided noncompliance. Using the principal stratification framework other extensions involved addressing missing outcomes (Mealli and Rubin, 2002), partial noncompliance (Jin and Rubin, 2008), and multi-arm studies (Long et al., 2010). Under one-sided noncompliance, the compliance of individuals to the active intervention is commonly unknown for participants in the control group. Identifying the sub-group of individuals who would have complied with the assigned intervention requires assumptions. These assumptions cannot always be verified from the data and are not controlled by randomization. Addressing noncompliance at the design stage could limit the need for such assumptions.

Prespecified interim analyses offer researchers an early assessment of intervention performance. Interim analyses may include summary statistics on the patient population, safety data on adverse events, efficacy data of the intervention, and reasons for early stopping (Armitage, 1991). Determining the number of interim analyses to perform depends on multiple factors such as the expected effect size, the expected power, and the sample size (McPherson, 1982). As the number of interim analyses increases, the probability of a spurious significant result increases (Armitage et al., 1969). To maintain appropriate Type I error rates, it is advised that researchers adjust their significance levels in accordance with the number of interim analyses planned or performed (Armitage et al., 1969; Geller and Pocock, 1987; Pocock, 1977). It has been argued that no such adjustment is required when working under a Bayesian framework (Berry, 1985). Interim analyses can be used to assess compliance in the study in order to modify the administration of the interventions.

Adaptive designs aim to pre-specify possible modifications to the trial's design, while maintaining the trial's validity and integrity (Chang, 2014). In these designs, investigators evaluate interventions as data are accrued and modify the trial. To ensure the validity and the integrity of the trials, the modifications should be prospectively defined in the trial protocol (Guidance, 2018; Chow et al., 2005). Current adaptive designs adjust different components of the trial (Chang, 2014). For example, group sequential designs are an adaptive design that allow premature termination of the trial because of efficacy or futility. Sample size reestimation designs enable sample-size adjustments based on interim analyses. Drop-loser designs stop assigning to inferior interventions after interim analyses according to a predefined criteria. Adaptive randomization designs adjust the randomization schedules during the conduct of the trial, and it may be based on observed responses. Adaptive dose-finding designs adjust the dose of the next patient based on the toxicity that is observed in previous patients. Treatment-switching designs allow switching of patient's initial assignment if there is evidence for lack of efficacy or safety concerns. Biomarker adaptive designs allow for adaption based on information obtained from biomarkers.

None of these adaptive designs specifically target lack of compliance with the assigned intervention, and many of them require interim analyses of the outcome data. We propose a new adaptive design for multi-component interventions, in which some components influence the compliance, and some components include the active ingredients. We describe the design of the adaptive trial, the corresponding analysis methods, and their required causal assumptions. Using simulations, we show that the proposed adaptive randomized trial result is statistically valid and with shorter interval estimates, compared to trials that do not adjust for noncompliance.

The paper proceeds as follows: Section 2 introduces the design, notations, and assumptions of the adaptive trial. Section 3 describes the design stage of the proposed trial and Section 4 describes its analysis stage. Possible analysis models, and simulations that show the operating characteristics of the design under these models are presented in Section 5. Section 6 implements the methods on a study that adjusted the intervention to address compliance. Section 7 provides discussion and conclusions.

3.2 Background

3.2.1 Design Setting and Notation

Let the number of individuals in the study be $N = N_S + N_A + N_C$, where N_S is the number of individuals assigned to the standard active intervention ("standard intervention"), N_A is the number of individuals assigned to the augmented active intervention ("augmented intervention"), and N_C is the number of individuals assigned to the control intervention. Let $N_T = N_S + N_A$ be the total number of individuals assigned to the active intervention arms, such that under 1:1 randomization $N_T = N_C$. For individual $i \in \{1, \ldots, N\}$, let $Z_i \in \{(T_i, A_i)\}$ be the treatment assignment indicators for individuals in the trial, where $T_i = 1$ if individual i is assigned to the active ingredient of the intervention and $T_i = 0$ otherwise, and $A_i = 1$ if individual i is assigned to augmented compliance and $A_i = 0$

otherwise. Because we only consider one-sided noncompliance, $A_i = 1$ can only occur in the active intervention arm, such that $Z_i \neq (0, 1), \forall i$. Thus, there are three possible intervention assignments for unit i, $Z_i \in \{(0,0),(1,0),(1,1)\}$ representing the control, standard, and augmented interventions, respectively.

The proposed adaptive design consists of three steps:

1. Randomize individuals to the control ($Z_i = (0,0)$) or the standard intervention group ($Z_i = (1,0)$) using individual-level randomization scheme (e.g. simple randomization, stratified randomization).

2. At a pre-defined point, examine the rates of compliance within the active intervention arm without unblinding the primary outcome data. If the compliance is at or above the expected compliance, continue the randomization in Step 1. If the compliance is below the threshold, adjust the compliance components of the multi-component intervention to increase compliance, and randomize individuals to the control arm ($Z_i = (0,0)$) or the augmented intervention arm ($Z_i = (1,1)$).

3. Estimate the ITT estimand for the effect of the standard or augmented intervention arm over the entire study population. This population comprises individuals in the control arm, individuals in the standard intervention arm, and individuals in the augmented intervention arm.

Let $n_S \in \{1, \ldots, N_S\}$ be the pre-defined number of individuals that would be observed before compliance rates are examined. Thus, let $N_S \in \{n_S, N_T\}$, such that if the augmented intervention is *not* implemented, then $n_S < N_S = N_T$. If the augmented intervention *is* implemented then $0 < N_A$, and $n_S = N_S < N_T$.

3.2.2 Potential Outcomes

To describe the ITT estimands, causal assumptions, and analysis methods, we will use the potential outcome framework (Rubin, 1978). This framework posits that for each unit i we have three primary potential outcomes $\mathbf{Y}_i(Z) = (Y_i(0,0), Y_i(1,0), Y_i(1,1))$ representing the outcome under each of the three possible interventions. In addition, each individual has three potential compliance statuses $\mathbf{W}_i(Z) = (W_i(0,0), W_i(1,0), W_i(1,1))$, where $W_i(Z_i) = 1$ if individual i complied with assignment Z_i and 0 otherwise. Because we assume that individuals in the control arm always comply with their assigned intervention $W_i(0,0) = 1, \forall i$. For each unit i we record P pre-intervention covariates, $\mathbf{X}_i = (X_{i1}, \ldots, X_{iP})$.

Only one of the potential outcomes and one of the compliance statuses are realized and observed for unit i. Under the stable unit treatment value assumption (Rubin, 1980, 1990), the observed and missing compliance outcomes for unit i can be written as,

$$W_i^{obs} = W_i(Z_i) = \begin{cases} W_i(0,0) = 1 & \text{if } Z_i = (0,0) \\ W_i(1,0), W_i(0,0) & \text{if } Z_i = (1,0) \\ W_i(1,1), W_i(0,0) & \text{if } Z_i = (1,1) \end{cases} \quad W_i^{mis} = \begin{cases} W_i(1,0), W_i(1,1) & \text{if } Z_i = (0,0) \\ W_i(1,1) & \text{if } Z_i = (1,0) \\ W_i(1,0) & \text{if } Z_i = (1,1) \end{cases}$$

Similarly, the observed and missing primary outcomes for unit i can be written as,

$$Y_i^{obs} = Y_i(Z_i) = \begin{cases} Y_i(0,0) & \text{if } Z_i = (0,0) \\ Y_i(1,0) & \text{if } Z_i = (1,0) \\ Y_i(1,1) & \text{if } Z_i = (1,1) \end{cases} \quad Y_i^{mis} = \begin{cases} \big(Y_i(1,0), Y_i(1,1)\big) & \text{if } Z_i = (0,0) \\ \big(Y_i(0,0), Y_i(1,1)\big) & \text{if } Z_i = (1,0) \\ \big(Y_i(0,0), Y_i(1,0)\big) & \text{if } Z_i = (1,1) \end{cases}$$

Let $\mathbf{X}_l^h = \{\mathbf{X}_i\}_l^h, \mathbf{T}_l^h = \{T_i\}_l^h, \mathbf{A}_l^h = \{A_i\}_l^h$, the matrix of covariates, assignment to active ingredient, and assignment to augmented compliance status for participants i such that $l \leq i \leq h$, respectively. In addition, let $\mathbf{Z}_l^h = (\mathbf{T}_l^h, \mathbf{A}_l^h)$, $\mathbf{W}_l^h = \{W_i(Z)\}_l^h$ and $\mathbf{Y}_l^h = \{Y_i(Z)\}_l^h$, the complete intervention assignment, the potential compliances matrix and the potential outcomes matrix for participants i such that $l \leq i \leq h$, respectively. $\mathbf{W}_l^h$ can be partitioned into $\mathbf{W}_l^{h,obs} = \{W_i^{obs}\}_l^h$ the matrix of observed compliance statuses and $\mathbf{W}_l^{h,mis} = \{W_i^{mis}\}_l^h$ the matrix of missing compliance statuses, similarly $\mathbf{Y}_l^h$ can be partitioned into

$\mathbf{Y}_l^{h,obs} = \{Y_i^{obs}\}_l^h$ the vector of observed outcomes and $\mathbf{Y}_l^{h,mis} = \{Y_i^{mis}\}_l^h$ the matrix of missing outcomes.

3.2.3 Estimands

Our proposed adaptive trial comprises two distinct stages. The design stage, which establishes the unit-level randomization scheme, the sample size of the trial, the compliance threshold, and the pre-defined point to examine rates of compliance within the active intervention arm. The second stage is the analysis stage, which provides estimation procedures for the rate of compliance of n_S standard intervention units, and the estimation of the intervention effect using the outcome data of the entire study population following the completion of the trial.

Compliance Estimand

The trial begins by selecting $2 \cdot n_S$ units from a pool of available units, and randomizing half of them to the control arm and half to the standard intervention arm. Without unblinding the outcomes, investigators use the observed $W_i(1,0)$ for the n_S units assigned to the standard intervention arm to summarize the super-population compliance rate, $CR \coloneqq E(W_i(1,0))$.

Primary Outcome Estimands

Pragmatic trials intend to examine the effectiveness of intervention among all individuals in the population. A commonly used estimand to summarize this effectiveness is the super-population difference between the mean of the outcomes if all units were assigned to the standard intervention and if all units were assigned to the control (ITT). When compliance to the standard intervention is high and no augmented intervention is employed the estimand of interest is $ITT^S \coloneqq E(Y_i(1,0) - Y_i(0,0))$. If compliance is low and the augmented intervention is introduced, the estimand of interest is $ITT^A \coloneqq E(Y_i(1,1) - Y_i(0,0))$. The expectation in both estimands is over the super-population of units.

Bayesian Methodology Overview

The analysis stage of the proposed adaptive trial relies on Bayesian model-based imputation (Imbens and Rubin (2015)). Bayesian inference for causal estimand τ considers $\mathbf{X}, \mathbf{T}, \mathbf{A}, \mathbf{Y}^{obs}$ and $\mathbf{W}^{obs}$ as a realization of random variables and $\mathbf{Y}^{mis}$ and $\mathbf{W}^{mis}$ as unobserved random variables. This perspective explicitly confronts the missing values by conditioning on observed variables and sampling from the posterior predictive distribution of τ,

$$\begin{aligned} &p(\tau|\mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N) \\ &\quad = \int p(\tau|\mathbf{Y}_1^{N,obs}, \mathbf{Y}_1^{N,mis}, \mathbf{W}_1^{N,obs}, \mathbf{W}_1^{N,mis}, \mathbf{X}_1^N, \mathbf{Z}_1^N) \\ &\qquad\qquad \cdot p(\mathbf{Y}_1^{N,mis}, \mathbf{W}_1^{N,mis}|\mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N) \;\; d\mathbf{Y}_1^{N,mis} d\mathbf{W}_1^{N,mis} \end{aligned} \tag{3.1}$$

Equation (3.1) shows that estimating the posterior distribution of τ involves integrating over $p(\mathbf{Y}_1^{N,mis}, \mathbf{W}_1^{N,mis}|\mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N)$.

3.2.4 Simplifying Assumptions

We assume that the N study participants are available at the onset of the study, and that all participants are randomized to either control intervention, $T_i = 0$, or active intervention, $T_i = 1$. We assume there is no effect of time on the compliance, outcomes, or assignment of the first $2 \cdot n_S$ units or the $N - 2 \cdot n_S$ units assigned to an intervention following the compliance assessment,

Assumption 1. $P(\mathbf{Y}_1^{2n_S}, \mathbf{W}_1^{2n_S}, \mathbf{Z}_1^{2n_S}, \mathbf{X}_1^{2n_S}) = P(\mathbf{Y}_{2n_S+1}^N, \mathbf{W}_{2n_S+1}^N, \mathbf{Z}_{2n_S+1}^N, \mathbf{X}_{2n_S+1}^N)$

We express the joint distribution of potential outcomes, intervention assignment, and covariates as the following,

Assumption 2.

$$
\begin{aligned}
p(\mathbf{Z},\mathbf{Y},\mathbf{W},\mathbf{X}) &= p(\mathbf{A},\mathbf{T},\mathbf{Y},\mathbf{W},\mathbf{X}) \\
&= p(\mathbf{A}_1^{2n_S},\mathbf{T}_1^{2n_S},\mathbf{Y}_1^{2n_S},\mathbf{W}_1^{2n_S},\mathbf{X}_1^{2n_S}) \\
&\quad \cdot p(\mathbf{A}_{2n_S+1}^{N},\mathbf{T}_{2n_S+1}^{N},\mathbf{Y}_{2n_S+1}^{N},\mathbf{W}_{2n_S+1}^{N},\mathbf{X}_{2n_S+1}^{N} \mid \mathbf{A}_1^{2n_S},\mathbf{T}_1^{2n_S},\mathbf{Y}_1^{2n_S},\mathbf{W}_1^{2n_S},\mathbf{X}_1^{2n_S}) \\
&= p(\mathbf{A}_1^{2n_S} \mid \mathbf{T}_1^{2n_S},\mathbf{Y}_1^{2n_S},\mathbf{W}_1^{2n_S},\mathbf{X}_1^{2n_S}) \cdot p(\mathbf{T}_1^{2n_S} \mid \mathbf{Y}_1^{2n_S},\mathbf{W}_1^{2n_S},\mathbf{X}_1^{2n_S}) \\
&\quad \cdot p(\mathbf{Y}_1^{2n_S} \mid \mathbf{W}_1^{2n_S},\mathbf{X}_1^{2n_S}) \cdot p(\mathbf{W}_1^{2n_S} \mid \mathbf{X}_1^{2n_S}) \cdot p(\mathbf{X}_1^{2n_S}) \\
&\quad \cdot p(\mathbf{A}_{2n_S+1}^{N} \mid \mathbf{T}_{2n_S+1}^{N},\mathbf{Y}_{2n_S+1}^{N},\mathbf{W}_{2n_S+1}^{N},\mathbf{X}_{2n_S+1}^{N},\mathbf{A}_1^{2n_S},\mathbf{T}_1^{2n_S},\mathbf{Y}_1^{2n_S},\mathbf{W}_1^{2n_S},\mathbf{X}_1^{2n_S}) \\
&\quad \cdot p(\mathbf{T}_{2n_S+1}^{N} \mid \mathbf{Y}_{2n_S+1}^{N},\mathbf{W}_{2n_S+1}^{N},\mathbf{X}_{2n_S+1}^{N},\mathbf{A}_1^{2n_S},\mathbf{T}_1^{2n_S},\mathbf{Y}_1^{2n_S},\mathbf{W}_1^{2n_S},\mathbf{X}_1^{2n_S}) \\
&\quad \cdot p(\mathbf{Y}_{2n_S+1}^{N} \mid \mathbf{W}_{2n_S+1}^{N},\mathbf{X}_{2n_S+1}^{N},\mathbf{A}_1^{2n_S},\mathbf{T}_1^{2n_S},\mathbf{Y}_1^{2n_S},\mathbf{W}_1^{2n_S},\mathbf{X}_1^{2n_S}) \\
&\quad \cdot p(\mathbf{W}_{2n_S+1}^{N} \mid \mathbf{X}_{2n_S+1}^{N},\mathbf{A}_1^{2n_S},\mathbf{T}_1^{2n_S},\mathbf{Y}_1^{2n_S},\mathbf{W}_1^{2n_S},\mathbf{X}_1^{2n_S}) \\
&\quad \cdot p(\mathbf{X}_{2n_S+1}^{N} \mid \mathbf{A}_1^{2n_S},\mathbf{T}_1^{2n_S},\mathbf{Y}_1^{2n_S},\mathbf{W}_1^{2n_S},\mathbf{X}_1^{2n_S}) \\
&= p(\mathbf{A}_1^{2n_S} \mid \mathbf{T}_1^{2n_S},\mathbf{X}_1^{2n_S}) \cdot p(\mathbf{T}_1^{2n_S} \mid \mathbf{X}_1^{2n_S}) \\
&\quad \cdot p(\mathbf{Y}_1^{2n_S} \mid \mathbf{W}_1^{2n_S},\mathbf{X}_1^{2n_S}) \cdot p(\mathbf{W}_1^{2n_S} \mid \mathbf{X}_1^{2n_S}) \cdot p(\mathbf{X}_1^{2n_S}) \\
&\quad \cdot p(\mathbf{A}_{2n_S+1}^{N} \mid \mathbf{T}_{2n_S+1}^{N},\mathbf{X}_{2n_S+1}^{N},\mathbf{W}_1^{2n_S,obs}) \cdot p(\mathbf{T}_{2n_S+1}^{N} \mid \mathbf{X}_{2n_S+1}^{N}) \\
&\quad \cdot p(\mathbf{Y}_{2n_S+1}^{N} \mid \mathbf{W}_{2n_S+1}^{N},\mathbf{X}_{2n_S+1}^{N}) \cdot p(\mathbf{W}_{2n_S+1}^{N} \mid \mathbf{X}_{2n_S+1}^{N}) \cdot p(\mathbf{X}_{2n_S+1}^{N}) \\
&= \prod_{i=1}^{N} p(A_i \mid T_i,\mathbf{W}_{n_S}^{obs},\mathbf{X}_i) \cdot p(T_i \mid \mathbf{X}_i) \\
&\qquad \cdot p(\mathbf{Y}_i(Z) \mid \mathbf{W}_i(Z),\mathbf{X}_i) \cdot p(\mathbf{W}_i(Z) \mid \mathbf{X}_i) \cdot p(\mathbf{X}_i)
\end{aligned}
$$

where for any individual i,

$$A_i = \begin{cases} 0 & \text{if } T_i = 0 \\ 0 & \text{if } T_i = 1, \sum_{j \leq i} T_j \leq n_S \\ 0 & \text{if } T_i = 1, \sum_{j \leq i} T_j > n_S, \mathbf{W}_{n_S}^{obs} \geq \phi \\ 1 & \text{if } T_i = 1, \sum_{j \leq i} T_j > n_S, \mathbf{W}_{n_S}^{obs} < \phi \end{cases}$$

where $\sum_{j \leq i} T_j \leq n_S$ identifies the original n_S individuals randomized to the standard intervention to observe compliance, $\mathbf{W}_{n_S}^{obs}$ represents the observed compliance statuses of these n_S standard intervention individuals, and ϕ is some compliance threshold where $\mathbf{W}_{n_S}^{obs} < \phi$ represents CR estimation below the compliance threshold and $\mathbf{W}_{n_S}^{obs} \geq \phi$ represents CR estimation at or above the compliance threshold. We note that the distribution of A_i is a degenerate distribution conditional on T_i, $\mathbf{W}_{n_S}^{obs}$, and $\mathbf{X}_i$.

Assumption 2 implies ignorability of the assignment mechanism of $\mathbf{Z}$, where $\mathbf{A}$ depends only on $\mathbf{T}$, $\mathbf{W}_{n_S}^{obs}$, and $\mathbf{X}$, but not the missing compliance statuses or any missing or observed primary outcomes. Because we assume $\mathbf{T}$ is randomized prior to the start of the trial and that the outcome data remains blinded during the compliance rate estimation, $\mathbf{T}$ is independent of all missing and observed primary and compliance outcomes.

The ignorable assignment mechanism is of particular importance in the outcome estimation procedures, such that we assume the results of the analysis would be the same had A_i been randomized prior to the trial for all i in the active intervention arms. **?** details an example where assignment to the active intervention depends on an individuals covariates and thereby leads to covariate imbalance between the intervention arms. **?** shows that when there is covariate imbalance, and in this particular example, ignoring a nonignorable intervention assignment, analysis of the outcomes can actually lead to the opposite and incorrect conclusions.

We assume that the potential primary outcomes of the active intervention arms are equal in expectation when the potential compliance statuses of these interventions are also equal. Formally,

Assumption 3. *For all* $i = 1, \ldots, N$, $E[Y_i(1,0) \mid W_i(1,0) = W_i(1,1) = w] = E[Y_i(1,1) \mid W_i(1,0) = W_i(1,1) = w]$ *for* $w = 0, 1$.

3.3 Addressing Noncompliance in the Design Stage

3.3.1 Background

Current methods to address noncompliance at the design stage rely on expected proportion of compliance, p_0, to adjust the required sample size to achieve predefined power with $p_0 = 1$. (Hickey et al., 2018; Lachin and Foulkes, 1986; Lachin, 1981; Wittes, 2002). For example, in a trial with 1:1 randomization scheme, when comparing the means under the active intervention and control arms, N_T can be estimated from the following,

$$N_T = 2 \cdot \left[\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{p_0 \cdot ES} \right]^2 \tag{3.2}$$

where α is the Type 1 Error, $1 - \beta$ is the Power, $ES = \frac{\mu_T - \mu_C}{\sigma}$ is the assumed Effect Size, with μ_T and μ_C the means of the active intervention and control arms, respectively, σ common standard deviation of the two populations, and where $Z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$, $Z_{1-\beta} = \Phi^{-1}(1 - \beta)$, with $\Phi^{-1}(\cdot)$ the inverse of the Cumulative Distribution Function (CDF) of the Standard Normal Distribution.

We demonstrate via simulation that Equation (3.2) is an appropriate sample size formula in settings of one-sided noncompliance, such that it achieves expected Power given the data generating mechanism is defined by p_0, ES, α, and the estimated N_T (Appendix 3.8.1). When the noncompliance is two-sided, where p_0 is the proportion of compliance in both the active intervention arms and control arm, Equation (3.2) does not achieve expected Power (Appendix 3.8.1). When the noncompliance is two-sided, Lachin and Foulkes (1986) suggest replacing p_0 in Equation (3.2) with $1 - q_T - q_C$, where q_T and q_C are the assumed proportions of noncompliance in the active intervention arms and control arm, respectively. Using simulation, we show that this formulation achieves expected Power under the estimated N_T and correct data generating mechanism (Appendix 3.8.1).

3.3.2 Type 1 Error

Our three-step adaptive trial tests two hypotheses during the analysis stage: one for the compliance rate of the n_S individuals assigned to the standard intervention and one for the ITT effect of the primary potential outcomes for assignment to the active intervention compared to control. By performing multiple tests, there may be increased probability of a spurious significant ITT result due to increased Type 1 Error (Pocock et al., 1987). There exists substantial literature addressing inflated Type 1 Error due to multiple tests, or multiple comparisons (Hochberg, 1988; Tukey, 1991; Benjamini and Hochberg, 1995; Benjamini and Braun, 2002). A key component of this literature adjusts the significance value for each test or comparison, to control the overall Type 1 Error at some pre-specified value.

However, we argue that under the null hypothesis, H_0, that there is no effect of the active interventions, the Type 1 Error for estimating the ITT remains unaffected by the CR estimation. Formally, consider Fisher's sharp null hypothesis (Imbens and Rubin, 2015) and let $H_0 : Y_i(0,0) = Y_i(1,0) = Y_i(1,1)$. For individual i, $W_i(1,0)$ influences $Y_i(1,0)$ through the effect of the active intervention $T_i = 1$, such that the active intervention effect may only be observed when $W_i(1,0) = 1$. However, under H_0, there cannot be an effect of the active intervention because $T_i = 0$ when $Z_i = (0,0)$ for all i, and thus, under H_0, compliance to the active intervention does not influence $Y_i(1,0)$. This holds similarly for $W_i(1,1)$ with $Y_i(1,1)$. We show both analytically and numerically, that the ITT Type 1 Error remains unchanged at the pre-defined rate of α regardless of the CR results and estimation itself (see Appendix 3.8.2).

3.3.3 Defining Compliance Thresholds

It is possible that p_0 will differ from the true compliance of the study population, p_{true}. When $p_{true} \neq p_0$, the trial may be under- or over-powered. Table 3.1 presents the Power of a trial for possible values of p_{true}, while ES, N_T, and α remain constant. The Power is calculated as follows,

$$1 - \beta = \Phi(Z_{1-\beta}) = \Phi\left(p_{true} \cdot ES\sqrt{\frac{N_T}{2}} - Z_{1-\alpha/2}\right) \tag{3.3}$$

where $\Phi(\cdot)$ is the Standard Normal CDF.

Table 3.1: The Power, $1 - \beta$, of a trial with $\alpha = 0.05$, given N_T and ES, and the true compliance being p_{true}.

	$N_T = 50$			$N_T = 100$			$N_T = 500$		
p_{true}	$ES = 1.1$	$ES = 1.3$	$ES = 1.5$	$ES = 0.8$	$ES = 0.9$	$ES = 1.0$	$ES = 0.4$	$ES = 0.5$	$ES = 0.6$
0.1	0.079	0.095	0.113	0.082	0.093	0.105	0.092	0.121	0.156
0.2	0.195	0.255	0.323	0.204	0.246	0.293	0.243	0.352	0.475
0.3	0.378	0.496	0.614	0.396	0.480	0.564	0.475	0.660	0.812
0.4	0.595	0.739	0.851	0.619	0.721	0.807	0.716	0.885	0.967
0.5	0.785	0.901	0.963	0.807	0.889	0.942	0.885	0.977	0.997
0.6	0.910	0.974	0.995	0.924	0.968	0.989	0.967	0.997	1
0.7	0.971	0.995	1	0.977	0.994	0.999	0.993	1	1
0.8	0.993	0.999	1	0.995	0.999	1	0.999	1	1
0.9	0.999	1	1	0.999	1	1	1	1	1

We observe in Table 3.1 that if p_0 is much greater than p_{true}, the trial is likely to be under-powered at the conclusion of the study. Similarly, if p_0 is much smaller than p_{true}, the trial is likely to be over-powered at the study conclusion.

Defining compliance thresholds to compare against estimated rates of compliance can protect against large differences between p_0 and p_{true}. Factors in defining compliance thresholds may include recruitment limitations or acceptable levels of Power.

For example, let $N_{max} > N_T$ be the maximum achievable intervention arm size of a trial and N_T be the estimated intervention arm size of the same trial using Equation (3.2). The minimum compliance threshold, δ_L, needed to obtain α and $1 - \beta$ with N_{max} intervention arm size can be determined by rearranging Equation (3.2) as follows,

$$\delta_L = \frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES \cdot \sqrt{N_{max}/2}}. \tag{3.4}$$

Alternatively, replacing N_{max} in Equation (3.4) with N_T and replacing $Z_{1-\beta}$ with $Z_{1-\beta_{min}}$, where $1-\beta > 1-\beta_{min}$, derives δ_L for a minimally acceptable power, given the estimated N_T.

Similar to Equation (3.4), a maximum compliance threshold, δ_U, given a maximum desired power, $1-\beta_{max} > 1-\beta$, α, ES, and N_T, is derived from the following,

$$\delta_U = \frac{Z_{1-\alpha/2} + Z_{1-\beta_{max}}}{ES \cdot \sqrt{N_T/2}} \tag{3.5}$$

Replacing N_T in Equation (3.5) with $N_{min} < N_T$ and replacing $Z_{1-\beta_{max}}$ with $Z_{1-\beta}$ derives δ_U for the minimal intervention arm size needed to achieve Power of $1-\beta$.

3.3.4 Defining Point of Compliance Estimation

Investigators determine n_S by balancing two criteria. First, as λ_S increases, so too does n_S, resulting in increased precision of the estimate for CR. Second, increasing n_S decreases N_A, resulting in worsened precision of the estimate for ITT^A if the augmented intervention is implemented. We consider $\frac{n_S}{N_T} \in [0.1, 0.2]$ for the remainder of the paper.

3.4 Addressing Noncompliance in the Analysis Stage

3.4.1 Rate of Compliance Estimation

Following the observation of n_S individuals in the standard intervention arm, the rate of compliance is estimated and compared to the compliance thresholds. When the rate of compliance is below the pre-defined threshold, investigators randomize remaining individuals to the augmented intervention or control.

Berry et al. (2010) outline Bayesian procedures to estimate the probability of success of a binary intervention at an interim analyses of an adaptive trial. We extend these methods to estimate CR (Section 3.2.3), using the n_S unblinded compliance statuses. The compliance thresholds, δ_L and δ_U, define decision rules in order to implement the augmented intervention.

For the following compliance estimation procedures, we assume $W_i(1,0) \sim Bernoulli(p)$ and $\sum_{i=1}^{n_S} W_i(1,0) \sim Binomial(n_S, p), \forall i$, with a non-informative prior distribution on the compliance probability, $p \sim Beta(1,1)$. We obtain the following posterior probability for p,

$$p | \sum_{i=1}^{n_S} W_i(1,0) \sim Beta(1 + \sum_{i=1}^{n_S} W_i(1,0), 1 + n_S - \sum_{i=1}^{n_S} W_i(1,0)). \tag{3.6}$$

We utilize Equation (3.6) for each of the following methods. Appendix 3.8.3 provides an example illustrating the application of the following methods when $p_0 = 0.5$, $n_S = 10$, $N_T = 50$, $ES = 1.3$, and $\alpha = 0.05$, for multiple values of δ_L and δ_U.

Posterior Probability Method

The posterior probability method utilizes the compliance thresholds, δ_L and δ_U (Section 3.3.3), and Equation (3.6). Additionally, investigators must apriori define probability thresholds required to trigger a decision to switch to the augmented intervention or maintain the standard intervention. Specifically, define the probability threshold for switching to the augmented intervention as π_s, and maintaining the standard intervention as π_m, where $\pi_s, \pi_m \in [0,1]$. Threshold values closer to 1 indicate greater probability required to trigger these decisions. We use π_s, π_m, δ_L, and δ_U to determine the following,

$$P\left(p < \delta_L \mid \sum_{i=1}^{n_S} W_i(1,0)\right) \tag{3.7}$$

$$P\left(p > \delta_U \mid \sum_{i=1}^{n_S} W_i(1,0)\right) \tag{3.8}$$

We adhere to the following decision rules:

- If Equation (3.7) $> \pi_s$, switch to the augmented intervention, else
- if Equation (3.8) $> \pi_m$, maintain the standard intervention with option for N_T reduction, otherwise
- maintain standard intervention with pre-defined N_T for the remainder of the trial.

Predictive Probability Method

The Predictive Probability Method extends the Posterior Probability Method in Section 3.4.1 by considering the future compliance statuses of individuals assigned to the active intervention, $W_i(1,0)$ for $i \in \{n_S+1, \ldots, N_T\}$. Additional to pre-defining n_S, δ_L, δ_U, π_s, and π_m, this method requires the investigator to specify an additional threshold value, $\theta_T \in [0,1]$, to inform the indicator function in Equations (3.9) and (3.10) which define the predictive probabilities of compliance below δ_L and above δ_U, respectively,

$$\sum_{j=0}^{N_T-n_S} P\Big(\sum_{i=n_S+1}^{N_T} W_i(1,0) = j \mid \sum_{i=1}^{n_S} W_i(1,0)\Big) \cdot \mathbb{1}_j\Big[P\Big(p < \delta_L \mid \sum_{i=1}^{n_S} W_i(1,0), \sum_{i=n_S+1}^{N_T} W_i(1,0) = j\Big) > \theta_T\Big] \tag{3.9}$$

$$\sum_{j=0}^{N_T-n_S} P\Big(\sum_{i=n_S+1}^{N_T} W_i(1,0) = j \mid \sum_{i=1}^{n_S} W_i(1,0)\Big) \cdot \mathbb{1}_j\Big[P\Big(p > \delta_U \mid \sum_{i=1}^{n_S} W_i(1,0), \sum_{i=n_S+1}^{N_T} W_i(1,0) = j\Big) > \theta_T\Big] \tag{3.10}$$

We have the following distributions on $\sum_{i=n_S+1}^{N_T} W_i(1,0)$ and p,

$$\sum_{i=n_S+1}^{N_T} W_i(1,0) \mid \sum_{i=1}^{n_S} W_i(1,0) \sim BetaBin\Big(N_T - n_S, 1 + \sum_{i=1}^{n_S} W_i(1,0), 1 + n_S - \sum_{i=1}^{n_S} W_i(1,0)\Big) \tag{3.11}$$

$$p \mid \sum_{i=1}^{n_S} W_i(1,0), \sum_{i=n_S+1}^{N_T} W_i(1,0) = j \sim Beta\Big(1 + j + \sum_{i=1}^{n_S} W_i(1,0), 1 + N_T - j - \sum_{i=1}^{n_S} W_i(1,0)\Big) \tag{3.12}$$

We adhere to the following decision rules:

- If Equation (3.9) $> \pi_s$, switch to the augmented intervention, else
- if Equation (3.10) $> \pi_m$, maintain the standard intervention with option for N_T reduction, otherwise
- maintain the standard intervention with pre-defined N_T for the remainder of the trial.

Indifference Zone Method

The indifference zone method utilizes Equation (3.6) to determine the lower and upper $100(1-\alpha)\%$ posterior credible interval bounds, p_L and p_U, respectively, for p, defined in Equations (3.13) and (3.14),

$$P\Big(p < p_L \mid \sum_{i=1}^{n_S} W_i(1,0)\Big) = \alpha/2 \tag{3.13}$$

$$P\Big(p > p_U \mid \sum_{i=1}^{n_S} W_i(1,0)\Big) = 1 - \alpha/2 \tag{3.14}$$

Using the compliance thresholds, δ_L and δ_U, and p_L and p_U, we adhere to the following decision rules:

- If $p_L < \delta_L$ and $p_U < \delta_U$, switch to the augmented intervention, or

- if $p_L > \delta_L$ and $p_U > \delta_U$, maintain standard intervention with option for N_T reduction, or

- if $p_L > \delta_L$ and $p_U < \delta_U$, maintain standard intervention with pre-defined N_T for the remainder of the trial, or

- if $p_L < \delta_L$ and $p_U > \delta_U$, follow apriori preference of investigator to maintain standard intervention for the remainder of the trial or switch to augmented intervention.

The final scenario of $p_L < \delta_L$ and $p_U > \delta_U$ indicates that the posterior credible interval of p contains both the upper and lower compliance thresholds. Therefore, the uncertainty of p does not lead to a decision through this method. Investigators should consider defining an active intervention switch decision if this scenario arises. In trials considering multiple interim analyses, a decision rule for this scenario would be well-suited to maintain standard intervention method until the subsequent interim compliance analysis.

3.4.2 Primary Outcomes Estimation

Bayesian Modelling

There are four steps required for our Bayesian model-based imputation approach. Recall we define our joint distributions as the product of N unit-level distributions from Assumption 2, however, for ease of notation we summarize the steps in vector form.

Step 1: Derive $p(\mathbf{Y}_1^{N,mis}, \mathbf{W}_1^{N,mis} \mid \mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N, \boldsymbol{\theta})$. This requires the specification of the joint distribution of the potential outcomes and potential compliance statuses given the covariates and $\boldsymbol{\theta}$, $p(\mathbf{Y}_1^N, \mathbf{W}_1^N \mid \mathbf{X}_1^N, \boldsymbol{\theta})$, and the specification of the intervention assignment mechanism $p(\mathbf{Z}_1^N \mid \mathbf{X}_1^N, \mathbf{W}_1^N, \mathbf{Y}_1^N)$ to obtain $p(\mathbf{Y}_1^N, \mathbf{W}_1^N, \mathbf{Z}_1^N \mid \mathbf{X}_1^N, \boldsymbol{\theta})$ and then $p(\mathbf{Y}_1^N, \mathbf{W}_1^N \mid \mathbf{X}_1^N, \mathbf{Z}_1^N, \boldsymbol{\theta})$. We use the transformations $h(\mathbf{Y}_i(Z), Z_i)$ and $g(\mathbf{W}_i(Z), Z_i)$ to obtain $p(\mathbf{Y}_1^{N,obs}, \mathbf{Y}_1^{N,mis}, \mathbf{W}_1^{N,obs}, \mathbf{W}_1^{N,mis} \mid \mathbf{X}_1^N, \mathbf{Z}_1^N, \boldsymbol{\theta})$ and subsequently derive $p(\mathbf{Y}_1^{N,mis}, \mathbf{W}_1^{N,mis} \mid \mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N, \boldsymbol{\theta})$.

Step 2: Derive the posterior distribution of the parameters given the observed data, which is proportional to,

$$p(\boldsymbol{\theta} \mid \mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N) \propto p(\boldsymbol{\theta}) \int p(\mathbf{Y}_1^{N,obs}, \mathbf{Y}_1^{N,mis}, \mathbf{W}_1^{N,obs}, \mathbf{W}_1^{N,mis} \mid \mathbf{X}_1^N, \mathbf{Z}_1^N, \boldsymbol{\theta}) d\mathbf{Y}_1^{N,mis} d\mathbf{W}_1^{N,mis} \tag{3.15}$$

Step 3: Derive the posterior predictive distribution of the missing outcomes and missing compliance statuses using Steps 1 and 2,

$$\begin{aligned} p(\mathbf{Y}_1^{N,mis}, &\mathbf{W}_1^{N,mis} \mid \mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N) \\ &= \int p(\mathbf{Y}_1^{N,mis}, \mathbf{W}_1^{N,mis} \mid \mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta} \mid \mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N)\, d\boldsymbol{\theta} \end{aligned} \tag{3.16}$$

Step 4: Derive the posterior distribution of the estimand of interest given the observed data, as defined in Equation (3.1).

Metropolis-Gibbs Sampling Algorithm

We utilize a combination of Metropolis and Gibbs sampling algorithms to appromixate the posterior distribution of the estimand of interest. The algorithm iterates between drawing from the posterior predictive distribution of $(\mathbf{Y}_1^{N,mis}, \mathbf{W}_1^{N,mis})$ given the observed data to impute the missing values and then drawing new parameter values for $\boldsymbol{\theta}$ from its posterior distribution given all other variables. The algorithm comprises the following steps (**?**),

1. Initialize parameters $\boldsymbol{\theta}^{(0)} = \{\theta_r^{(0)}\}$ for $r \in 1, \ldots, R$.

2. For $t = 1, 2, \ldots$:

    (a) Draw $(\mathbf{Y}_1^{N,mis,(t)}, \mathbf{W}_1^{N,mis,(t)}) \sim p(\mathbf{Y}_1^{N,mis}, \mathbf{W}_1^{N,mis} \mid \mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N, \boldsymbol{\theta}^{(t-1)})$.

    (b) Draw $\theta_r^{(t)} \sim p(\theta_m \mid \boldsymbol{\theta}_{-r}^{(t-1)}, \mathbf{Y}_1^{N,mis,(t)}, \mathbf{W}_1^{N,mis,(t)}, \mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N)$ for $r \in 1, \ldots, R$.

    (c) Calculate $\tau^{(t)}$.

(d) Repeat steps 2(a) - 2(d) sufficiently large number of times.

where $\boldsymbol{\theta}_{-r}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_{r-1}^{(t)}, \theta_{r+1}^{(t-1)}, \ldots, \theta_R^{(t-1)})$.

For any θ_r, $r \in 1, \ldots, R$, with a conditional posterior distribution that is not of closed form, we utilize a Metropolis-Hastings algorithm (**?**) to draw $\theta_r^{(t)}$ in Step 2(b) of the algorithm above, as follows,

i. Draw a starting point $\theta_r^{(0)}$ for which $f(\theta_r^0 \mid \boldsymbol{\theta}_{-r}^{(t)}, \mathbf{Y}_1^{N,mis,(t)}, \mathbf{W}_1^{N,mis,(t)}, \mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N) > 0$ from a starting distribution $f_0(\theta_r)$.

ii. For $t = 1, 2, \ldots$:

(a) Sample a proposal θ_r^* from a jumping distribution at time t, $J_t(\theta_r^* \mid \theta_r^{t-1})$, for a symmetric jumping distribution that satisfies $J_t(\theta_a \mid \theta_b) = J_t(\theta_b \mid \theta_a)$ for all θ_a, θ_b, and t.

(b) Calculate $q = \frac{f(\theta_r^* | \boldsymbol{\theta}_{-r}^{(t)}, \mathbf{Y}_1^{N,mis,(t)}, \mathbf{W}_1^{N,mis,(t)}, \mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N)}{f(\theta_r^{(t-1)} | \boldsymbol{\theta}_{-r}^{(t)}, \mathbf{Y}_1^{N,mis,(t)}, \mathbf{W}_1^{N,mis,(t)}, \mathbf{Y}_1^{N,obs}, \mathbf{W}_1^{N,obs}, \mathbf{X}_1^N, \mathbf{Z}_1^N)}$.

(c) Set $\theta_r^t = \begin{cases} \theta_r^* & \text{with probability min(q,1)} \\ \theta_r^{(t-1)} & \text{otherwise.} \end{cases}$

(d) Repeat steps ii(a) - ii(d) sufficiently large number of times.

Non-Imputation Estimation

We consider three possible non-imputation based estimators for our three step adaptive trial, to estimate the estimands of interest in Section 3.2.3. When compliance to the standard intervention is low and the augmented intervention is employed, let $\widehat{ITT^{A1}} = \frac{1}{N_A} \sum_{i:Z_i=(1,1)} Y_i^{obs} - \frac{1}{N_C} \sum_{i:Z_i=(0,0)} Y_i^{obs}$ and $\widehat{ITT^{A2}} = \frac{1}{N_A+N_S} \big(\sum_{i:Z_i=(1,1)} Y_i^{obs} + \sum_{i:Z_i=(1,0)} Y_i^{obs} \big) - \frac{1}{N_C} \sum_{i:Z_i=(0,0)} Y_i^{obs}$. When compliance is high and no augmented intervention is employed, let $\widehat{ITT^S} = \frac{1}{N_S} \sum_{i:Z_i=(1,0)} Y_i^{obs} - \frac{1}{N_C} \sum_{i:Z_i=(0,0)} Y_i^{obs}$.

3.5 Simulation Study

3.5.1 Data Generating Mechanism

We perform simulation analysis to test the methodology outlined in Section 3.4.2 and to test the assumptions of our three step adaptive trial. The simulation study focusses on the setting when compliance to the standard intervention is low and the augmented intervention is implemented.

We first define the size of the study population, N, and the point to examine the compliance rates, $n_S = N_S$. Let p_0 be the assumed rate of compliance to the standard intervention. Under a 1:1 simple randomization scheme, we randomize T_i for all i. Therefore, $N_C = N/2$ and $N_A = N - N_S - N_C$. We assume $A_i = 1$ for any unit i randomized to $T_i = 1$ entering the trial subsequent to the n_S standard intervention unit.

For each unit i, we generate $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})$ from $X_{i1} \sim N(0,1)$, $X_{i2} \sim N(0,1)$, and $X_{i3} \sim N(1,1)$, where X_{i1}, X_{i2}, and X_{i3} are mutually independent. We generate potential compliance statuses under the active interventions from two independent Probit models, $W_i(1,0) \sim Ber(p_i)$ and $W_i(1,1) \sim Ber(q_i)$, where $p_i = \Phi\big(\beta_{p0} + \beta_{p1}X_{i1} + \beta_{p2}X_{i2} + \beta_{p3}X_{i3}\big)$, $q_i = \Phi\big(\beta_{q0} + \beta_{q1}X_{i1} + \beta_{q2}X_{i2} + \beta_{q3}X_{i3}\big)$, β_{uv} are the pre-defined parameter values for all $u \in \{p, q\}$ and $v \in \{0, 1, 2, 3\}$, and $\Phi(\cdot)$ is the standard normal CDF. We define $\beta_{p0} = \Phi^{-1}(\lambda_p p_0)$, and $\beta_{q0} = \Phi^{-1}((1+\lambda_q)p_0)$ for $(1+\lambda_q)p_0 < 1$, where $\lambda_p = \frac{p_{true}}{p_0}$ and p_{true} is the true population rate of compliance to the standard intervention, and λ_q represents the proportion of increased compliance relative to the assumed compliance that investigators achieve via the augmented intervention.

We generate the potential outcomes of unit i such that $\mathbf{Y}_i(Z) \mid \mathbf{W}_i(Z), \mathbf{X}_i, \boldsymbol{\theta} \sim \mathbf{G}_3(\tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}})$, for trivariate distribution $\mathbf{G}$, with 3-dimensional location vector $\tilde{\boldsymbol{\mu}}_i$, and 3×3 matrix $\tilde{\boldsymbol{\Sigma}}$.

Specifically, we define $\tilde{\boldsymbol{\mu}}_i = E[\mathbf{Y}_i(Z) \mid \mathbf{W}_i(Z), \mathbf{X}_i, \boldsymbol{\theta}]$ to be the following,

$$\begin{aligned}
E[Y_i(0,0) \mid \mathbf{W}_i(Z), \mathbf{X}_i, \boldsymbol{\theta}] &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 W_i(1,0) + \beta_4 W_i(1,1) \\
&\quad + \beta_5 exp\{X_{i1}\} + \beta_6 X_{i3} \\
E[Y_i(1,0) \mid \mathbf{W}_i(Z), \mathbf{X}_i, \boldsymbol{\theta}] &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 W_i(1,0) + \beta_4 W_i(1,1) \\
&\quad + \beta_5 exp\{X_{i1}\} + \beta_6 X_{i3} + \beta_7 X_{i2} W_i(1,0) T_i + \mu_T W_i(1,0) T_i \\
E[Y_i(1,1) \mid \mathbf{W}_i(Z), \mathbf{X}_i, \boldsymbol{\theta}] &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 W_i(1,0) + \beta_4 W_i(1,1) \\
&\quad + \beta_5 exp\{X_{i1}\} + \beta_6 X_{i3} + \beta_8 X_{i2} W_i(1,1) T_i + \mu_T W_i(1,1) T_i
\end{aligned} \tag{3.17}$$

where β_b, $b = 1, \ldots, 8$ and μ_T are pre-defined parameter values. Additionally, define $\tilde{\boldsymbol{\Sigma}}$ such that $\tilde{\Sigma}_{11} = \sigma_0^2$, $\tilde{\Sigma}_{22} = \sigma_{10}^2$, $\tilde{\Sigma}_{33} = \sigma_{11}^2$, and define $\tilde{\Sigma}_{ij} = \rho\sqrt{\tilde{\Sigma}_{ii}}\sqrt{\tilde{\Sigma}_{jj}}$ for $i, j \in \{1, 2, 3\}$ and $i \neq j$.

We separate these simulations into four groups: Design Stage Assumptions, Potential Outcome Model Assumptions, Compliance Model Assumptions, and Normality Assumptions.

To test the Design Stage Assumptions we vary the variables relevant to the design stage of the trial: p_0, p_{true}, $n_S/N_S = \lambda_{n_S}$, N_T, and the receipt of active ingredient effect size, μ_1.

To test the Potential Outcome Model Assumptions we assign non-zero values to β_5, β_6, β_7, and β_8 in Equation (3.17) for the data generating mechanism to intentionally misspecify our potential outcome models defined in the following section.

To test the Compliance Model Assumptions we vary parameter values related to compliance statuses. In addition, we assign non-zero values to β_{p3} and β_{q3} for the data generating mechanism to intentionally misspecify our compliance statuses model defined in the following section.

Finally, to test the Normality Assumptions we specify parameter values from a combi-

Table 3.2: The pre-defined variables for simulations assessing Design and Modelling Assumptions.

Variable	Design	PO Model	Compliance	Normality
p_0	0.35, 0.50, 0.65	0.50	0.50	0.50
λ_p	0.50, 0.75	0.75	0.75	0.75
λ_q	0.10	0.10	0.10	0.10
λ_{n_S}	0.10, 0.20	0.20	0.20	0.20
N_T	500, 1000	500	500	500
$\mathbf{G}$	$\mathcal{N}_3$	$\mathcal{N}_3$	$\mathcal{N}_3$	t_ν
ν	-	-	-	5
β_0	0	0	0	0
β_1	1	0	1	1
β_2	0.50	0.50	0.50	0.50
β_3	0	0.75	0.25, 1	0.50
β_4	0	0.50	0.25, 0.50	0.50
β_5	0	0.10, 0.25	0	0, 0.25
β_6	0	0, 0.25	0	0, 0.25
β_7	0	0, 0.25	0	0, 0.25
β_8	0	0, 0.25	0	0, 0.25
μ_1	0.50, 1	1	1	1
σ_0	1	1	1	1
σ_{10}	1	1	1	1
σ_{11}	1	1	1	1
ρ	0.50	0.50	0.50	0.50
$g(\cdot)$	Probit	Probit	Probit	Probit
β_{p1}	0.25	0.25	0.25	0.25
β_{p2}	0.25	0.25	0.25	0.25
β_{p3}	0	0	0.10, 0.25	0
β_{q1}	0.25	0.25	0.25	0.25
β_{q2}	0.25	0.25	0.25	0.25
β_{q3}	0	0	0.10, 0.25	0.10

nation of the three previous simulation groups, where the potential outcomes are generated from a t-distribution with ν degrees of freedom. Table 3.2 summarizes the variable values within each simulation.

3.5.2 Simulation Models and Estimation Procedures

We define four Bayesian imputation-based models and two non-imputation-estimation procedures for our simulation analysis. We follow the assumptions outlined in Sections 3.2.4.

Primary Outcome Models

To implement the methods outlined in Sections 3.4.2 and 3.4.2, the Bayesian models require the specification of $p(\mathbf{Y}, \mathbf{W}|\mathbf{X}, \boldsymbol{\theta}) = p(\mathbf{Y} \mid \mathbf{W}, \mathbf{X}, \boldsymbol{\theta})p(\mathbf{W}|\mathbf{X}, \boldsymbol{\theta})$. We assume $\mathbf{X}_i = \{X_{i1}, X_{i2}\}$, such that investigators do not record or do not have knowledge of X_{i3}. We first define the unit-level conditional distribution of the potential outcomes for any individual i as $\mathbf{Y}_i(Z) \mid \mathbf{W}_i(Z), \mathbf{X}_i, \boldsymbol{\theta} \sim \mathbf{N}_3(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ for trivariate normal distribution $\mathbf{N}_3$ with expectation $\boldsymbol{\mu}_i$ and 3×3 covariance matrix $\boldsymbol{\Sigma}$. We define $\boldsymbol{\mu}_i = E[\mathbf{Y}_i(Z) \mid \mathbf{W}_i(Z), \mathbf{X}_i, \boldsymbol{\theta}]$ for all four Bayesian models as,

$$
\begin{aligned}
&E[Y_i(0,0) \mid \mathbf{W}_i(Z), \mathbf{X}_i, \boldsymbol{\theta}] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 W_i(1,0) + \beta_4 W_i(1,1) \\
&E[Y_i(1,0) \mid \mathbf{W}_i(Z), \mathbf{X}_i, \boldsymbol{\theta}] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 W_i(1,0) + \beta_4 W_i(1,1) + \mu_T W_i(1,0) T_i \\
&E[Y_i(1,1) \mid \mathbf{W}_i(Z), \mathbf{X}_i, \boldsymbol{\theta}] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 W_i(1,0) + \beta_4 W_i(1,1) + \mu_T W_i(1,1) T_i
\end{aligned}
\tag{3.18}
$$

where β_b, $b = 1, \ldots, 4$ and μ_T represent unknown parameters. We distinguish between the four Bayesian models through the specification of their covariance matrix, $\boldsymbol{\Sigma}$, defined in Table 3.3. We denote these models by their covariance matrix assumptions: Independent (abbreviated as Indep), Equal Variance and constant correlation (abbreviated as EqVar), Differing Variance for compliance to active interventions versus noncompliance and constant correlation (abbreviated as DifVar), and Differing Variance and Correlation for compliance to active intervetions versus noncompliance (abbreviated as DifCov). The motivation for DifCov is presented in Appendix 3.8.4. In Table 3.3, for DifVar we define $\sigma_{10}^2 = \sigma_1^2 \cdot W_i(1,0) + \sigma_0^2 \cdot (1 - W_i(1,0))$ and $\sigma_{11}^2 = \sigma_1^2 \cdot W_i(1,1) + \sigma_0^2 \cdot (1 - W_i(1,1))$.

Compliance Models

For all four Bayesian imputation-based models we define the unit-conditional distribution of the potential compliance statuses, for all i, as $p(\mathbf{W}_i(Z) \mid \mathbf{X}_i, \boldsymbol{\theta}) = p(W_i(1,0) \mid \mathbf{X}_i, \boldsymbol{\theta})p(W_i(1,1) \mid$

Table 3.3: Conditional covariance matrices of the potential response outcomes, given the potential compliance outcomes, the covariates, and the model parameters.

Model Name	$\boldsymbol{\Sigma}$
Indep	$\begin{pmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_{10}^2 & 0 \\ 0 & 0 & \sigma_{11}^2 \end{pmatrix}$
EqVar	$\begin{pmatrix} \sigma_0^2 & \rho\sigma_0^2 & \rho\sigma_0^2 \\ \rho\sigma_0^2 & \sigma_0^2 & \rho\sigma_0^2 \\ \rho\sigma_0^2 & \rho\sigma_0^2 & \sigma_0^2 \end{pmatrix}$
DifVar	$\begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_{10} & \rho\sigma_0\sigma_{11} \\ \rho\sigma_0\sigma_{10} & \sigma_{10}^2 & \rho\sigma_{10}\sigma_{11} \\ \rho\sigma_0\sigma_{11} & \rho\sigma_{10}\sigma_{11} & \sigma_{11}^2 \end{pmatrix}$
DifCov	$\begin{pmatrix} \sigma_0^2 & \sigma_0^2 & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2 + \tau_\delta^2 + W_i(1,0)\cdot\tau_1^2 & \sigma_0^2 + W_i(1,0)\cdot W_i(1,1)\cdot\tau_1^2 \\ \sigma_0^2 & \sigma_0^2 + W_i(1,0)\cdot W_i(1,1)\cdot\tau_1^2 & \sigma_0^2 + \tau_\epsilon^2 + W_i(1,1)\cdot\tau_1^2 \end{pmatrix}$

$\mathbf{X}_i, \boldsymbol{\theta})$. Recall, we assume $W_i(0,0) = 1,\ \forall i$. Similar to the data generating mechanism, we define $W_i(1,0) \mid \mathbf{X}_i, \boldsymbol{\theta} \sim Ber(p_i)$ and $W_i(1,1) \mid \mathbf{X}_i, \boldsymbol{\theta} \sim Ber(q_i)$, where $p_i = \Phi\big(\beta_{p0} + \beta_{p1}X_{i1} + \beta_{p2}X_{i2}\big)$, $q_i = \Phi\big(\beta_{q0} + \beta_{q1}X_{i1} + \beta_{q2}X_{i2}\big)$, with β_{uv} representing the unknown parameter values for all $u \in \{p, q\}$ and $v \in \{0, 1, 2\}$.

However, conditional independence of the compliance statuses for unit i given their covariates may not be a preferred assumption by all investigators. It may be preferred to make assumptions that relate the observed compliance to the missing compliance statuses. For example, investigators may believe that an individual who has complied to the standard intervention would then also comply to the augmented intervention, such that $p(\mathbf{W}_i(Z) \mid \mathbf{X}_i, \boldsymbol{\theta}) = p(W_i(1,1) \mid W_i(1,0), \mathbf{X}_i, \boldsymbol{\theta})p(W_i(1,0) \mid \mathbf{X}_i, \boldsymbol{\theta})$. Under this assumption, when $W_i^{obs} = W_i(1,0) = 1$ for unit i, investigators would then directly impute $W_i^{mis} = W_i(1,1) = 1$. This improves the efficiency of the compliance estimation, compared to the conditional independence assumption, as there is no uncertainty of the augmented intervention compliance when unit i complies to the standard intervention. When this assumption is true, this may lead to overall efficiency gains in compliance estimation, however when the assumption is untrue, it may lead to poor compliance estimation.

Investigators may weaken the previous assumption by assuming that an individual's compliance probability is larger in expectation for the augmented intervention compared to the standard intervention, such that $P(W_i(1,1) = 1 \mid W_i(1,0), \mathbf{X}_i, \boldsymbol{\theta}) = p_i + \lambda W_i(1,0)$ and $P(W_i(1,0) = 1 \mid \mathbf{X}_i, \boldsymbol{\theta}) = q_i$, where $p_i > q_i$ and $\lambda \in [0, 1 - p_i]$ represents the proportion of increased compliance probability to the augmented intervention, given unit i complied to the standard intervention. Again, there may be efficiency gains in the compliance estimation, compared to the conditional independence assumption, when this assumption is true, however, the impact on efficiency of ITT estimation may not be substantial.

These two alternative assumptions for the potential compliance status distribution may lead to efficiency gains in the compliance estimation, compared to the conditional independence assumption. However, they rely on the augmented intervention improving compliance in specific ways, which may not happen in practice. Conditional independence given an individual's covariates protects against these violations, while possibly losing efficiency if the alternative assumptions are true. Ultimately, these assumptions are often impossible to validate with observed data because an individual will only observe a single compliance status during a trial. Moreover, any efficiency gains in the compliance estimation may not lead to efficiency gains in the ITT estimation which focusses on the primary potential outcomes. Investigators must decide the modelling assumptions they wish to make about compliance behavior while considering the possible gains or losses they may encounter from those assumptions.

Priors, Sampling Algorithm, and Alternative Estimation Procedures

We define a prior distribution for each parameter with the assumption of mutual independence, such that, $p(\boldsymbol{\theta}) = p(\theta_1, \theta_2, \ldots, \theta_R) = p(\theta_1)p(\theta_2)\ldots p(\theta_R)$ for all R parameters in $\boldsymbol{\theta}$. Specifically, we define $\beta_b \sim N(0, 100^2)$ for all $b \in \{0, 1, 2, 3, 4\}$, $\mu_T \sim N(0, 100^2)$, $p(\beta_{uv}) \propto 1$

for all $u \in \{p, q\}$ and $v \in \{0, 1, 2\}$, $p(\rho) \sim Beta(1, 10)$, and then all remaining terms within $\boldsymbol{\Sigma}$ as σ_0^2, σ_1^2, σ_{10}^2, σ_{11}^2, τ_1^2, τ_ϵ^2, and τ_δ^2 following a Half-Cauchy distribution with support $[0, \infty)$ and scale 25.

For each Bayesian model, we employ the Metropolis-Gibbs algorithm defined in Section 3.4.2. We set the number of Gibbs iterations to 3500 and the number of Metropolis-Hastings iterations to 250. We utilize the Metropolis-Hastings algorithm to draw each parameter in Σ defined in Table 3.3.

We implement two non-imputation-based estimation procedures outlined in Section 3.4.2, and define AugInt $:= \widehat{ITT^{A1}}$ and BothInt $:= \widehat{ITT^{A2}}$.

3.5.3 Simulation Evaluation Metrics

Table 3.2 presents 96 unique combinations of variable values across the simulation groups. We denote each combination as configuration l for $l \in \{1, \ldots, L\}$, where $L = 96$. The data generating mechanism defined by configuration l is simulated and analyzed 100 times, for all l. We denote the simulated dataset for configuration l by $m \in \{1, \ldots, M\}$, where $M = 100$.

Because our simulations focus on the setting when compliance to the standard intervention is low and the augmented intervention is implemented, we define the estimand of interest to be $\tau_l = ITT^A$, for configuration l. We consider four evaluation metrics for $\hat{\tau}_l$, the estimate of τ_l: coverage probability of τ_l, mean variance, mean bias of $\hat{\tau}_l$, and mean absolute bias of $\hat{\tau}_l$ (MAB).

3.5.4 Simulation Study Results

Design Stage Assumptions

As shown in Table 3.2, there are 48 configurations that are each replicated 100 times. Table 3.4 summarizes the coverage probability, mean variance, mean bias of $\hat{\tau}_l$, and mean absolute bias of $\hat{\tau}_l$ (MAB), averaging across the 48 configurations, for each of the four Bayesian imputation-based models (Indep, EqVar, DifVar, DifCov) and the two non-imputation-based procedures (BothInt, AugInt).

Each of the methods have approximately 90% or higher coverage, with the exception of Indep, which has the lowest coverage. The remaining methods achieve nominal coverage via the distribution of coverage probabilities across configurations. The Bayesian models are similar in mean variance ranging from 0.0031 to 0.0035 on average, whereas the mean variance under the non-imputation methods is twice as large. AugInt has the largest mean variance at 0.0080 on average from all configurations.

The mean bias of $\hat{\tau}_l$ is approximately 0 for all methods, with the greatest difference from 0 as -0.007 for Indep and 0.025 for BothInt. The Bayesian models have the smallest mean absolute bias of $\hat{\tau}_l$ compared to the non-imputation methods. DifVar and EqVar have the smallest MAB respectively, with DifCov only 10% larger and Indep nearly 50% larger. BothInt and AugInt have more than a 50% increase in mean absolute bias of $\hat{\tau}_l$ relative to DifVar and EqVar.

Table 3.4: Coverage probability, mean variance, mean bias of $\hat{\tau}_l$, and mean absolute bias of $\hat{\tau}_l$ of each method for the Design Stage Assumptions Simulation, averaged across the 48 configurations. Standard deviations of these values are provided in parentheses.

	Metric			
Model	Coverage	Variance	Bias	MAB
Indep	0.873 (0.04)	0.0035 (0.0011)	-0.007 (0.009)	0.061 (0.01)
EqVar	0.948 (0.02)	0.0031 (0.001)	0.001 (0.006)	0.044 (0.008)
DifVar	0.944 (0.02)	0.0032 (0.0011)	0.001 (0.006)	0.044 (0.008)
DifCov	0.912 (0.03)	0.0032 (0.0011)	0.002 (0.007)	0.049 (0.009)
BothInt	0.934 (0.03)	0.0073 (0.0025)	0.025 (0.019)	0.071 (0.013)
AugInt	0.952 (0.02)	0.0080 (0.0027)	0.000 (0.009)	0.070 (0.013)

Potential Outcome Model Assumptions

Table 3.5 presents the coverage probabilities, mean variance, mean bias of $\hat{\tau}_l$, and mean absolute bias of $\hat{\tau}_l$, averaged across the 16 configurations of this simulation group. Each method has above 90% coverage probability on average, achieving nominal coverage via the distribution of coverage probabilities across configurations. Again, the Bayesian models have the smallest mean variances ranging from 0.0056 to 0.0061 on average. The non-imputation methods, BothInt and AugInt, on average have an increased mean variance by 45% and 67%, respectively, relative to Bayesian models.

The mean bias of $\hat{\tau}_l$ is smallest for AugInt at 0.001 and for DifCov at -0.005, averaged over the configurations. The mean bias of $\hat{\tau}_l$ is similar between the remaining methods ranging from 0.027 to 0.043 in magnitude difference from 0.

Each method has an increase in mean absolute bias of $\hat{\tau}_l$ relative to their respective results in the Design Stage Assumptions simulation from Table 3.4. The non-imputation methods again have the largest mean absolute bias of $\hat{\tau}_l$ ranging from 8% to 20% larger than all of the Bayesian models, on average.

Table 3.5: Coverage probability, mean variance, mean bias of $\hat{\tau}_l$, and mean absolute bias of $\hat{\tau}_l$ of each method for the Potential Outcome Model Assumptions Simulation, averaged across the 16 configurations. Standard deviations of these values are provided in parentheses.

	Metric			
Model	Coverage	Variance	Bias	MAB
Indep	0.91 (0.03)	0.0061 (4e-04)	-0.043 (0.016)	0.072 (0.008)
EqVar	0.917 (0.03)	0.0056 (2e-04)	-0.030 (0.013)	0.067 (0.005)
DifVar	0.921 (0.03)	0.0057 (2e-04)	-0.027 (0.012)	0.066 (0.004)
DifCov	0.916 (0.02)	0.0058 (2e-04)	-0.005 (0.015)	0.066 (0.003)
BothInt	0.946 (0.02)	0.0089 (7e-04)	0.035 (0.011)	0.078 (0.006)
AugInt	0.956 (0.02)	0.0102 (8e-04)	0.001 (0.009)	0.079 (0.006)

Compliance Model Assumptions

Table 3.6 presents the coverage probabilities, mean variance, mean bias of $\hat{\tau}_l$, and mean absolute bias of $\hat{\tau}_l$ of each configuration, averaged across the 16 configurations of this simulation group. Each method has above 90% coverage probability on average, achieving nominal coverage via the interior distribution of coverage probabilities across configurations. The Bayesian models have the smallest mean variances and are similar to one another, ranging from 0.0052 to 0.0054 on average. The non-imputation methods, BothInt and AugInt, on average have an increased mean variance by 132% and 162%, respectively, relative to Bayesian models.

The mean bias of $\hat{\tau}_l$ is smallest for AugInt at 0.001 and largest for BothInt at 0.035, on average. The mean bias of $\hat{\tau}_l$ is similar between the Bayesian models ranging from 0.009 to 0.012 in magnitude difference from 0.

Again, each method has an increase in MAB of $\hat{\tau}_l$ relative to their respective results in the Design Stage Assumptions simulation from Table 3.4, although not as large of an increase relative to the Potential Outcome Model Assumptions results. The non-imputation methods have the largest MAB of $\hat{\tau}_l$ ranging from 37% to 61% larger than all of the Bayesian models, on average.

Table 3.6: Coverage probability, mean variance, mean bias of $\hat{\tau}_l$, and mean absolute bias of $\hat{\tau}_l$ of each method for the Compliance Model Assumptions Simulation, averaged across the 16 configurations. Standard deviations of these values are provided in parentheses.

	Metric			
Model	Coverage	Variance	Bias	MAB
Indep	0.922 (0.03)	0.0054 (6e-04)	-0.012 (0.02)	0.065 (0.005)
EqVar	0.945 (0.02)	0.0052 (4e-04)	-0.01 (0.01)	0.057 (0.005)
DifVar	0.945 (0.02)	0.0053 (4e-04)	-0.009 (0.009)	0.057 (0.004)
DifCov	0.93 (0.02)	0.0054 (3e-04)	0.009 (0.011)	0.061 (0.005)
BothInt	0.95 (0.02)	0.0123 (0.0011)	0.035 (0.014)	0.089 (0.007)
AugInt	0.957 (0.02)	0.0139 (0.0011)	0.001 (0.015)	0.092 (0.005)

Normality Assumptions

We summarize the Normality Assumptions simulation results for coverage probability, mean variance, mean bias of $\hat{\tau}_l$, and MAB of $\hat{\tau}_l$ in Table 3.7, averaged across the 16 configurations for this group. All methods have nominal coverage via the distribution of coverage probabilities across configurations. On average, each method has mean variance more than double compared to their respective mean variance from Table 3.4. Again, on average the Bayesian models have the smallest mean variances, ranging from 0.0076 to 0.0078 for EqVar, DifVar, and DifCov, and at 0.0096 for Indep. The non-imputation methods, BothInt and AugInt, on average have an increased mean variance by 136% and 164%, respectively, relative to DifVar.

The mean bias of $\hat{\tau}_l$ is smallest in magnitude difference from 0 for AugInt at 0.002 and largest for Indep at 0.061, on average. The mean bias of $\hat{\tau}_l$ for the remaining methods ranges from 0.025 to 0.041 in magnitude difference from 0.

The Bayesian models have the smallest MAB of $\hat{\tau}_l$ on average, ranging from 0.073 to 0.074 for EqVar, DifVar, and DifCov, and at 0.084 for Indep. The non-imputation methods have the largest MAB of $\hat{\tau}_l$, on average, ranging from 26% to 51% larger than all of the Bayesian models.

Table 3.7: Coverage probability, mean variance, mean bias of $\hat{\tau}_l$, and mean absolute bias of $\hat{\tau}_l$ of each method for the Normality Assumptions Simulation, averaged across the 16 configurations. Standard deviations of these values are provided in parentheses.

	Metric			
Model	Coverage	Variance	Bias	MAB
Indep	0.933 (0.03)	0.0096 (5e-04)	-0.061 (0.015)	0.084 (0.01)
EqVar	0.925 (0.03)	0.0076 (4e-04)	-0.041 (0.011)	0.074 (0.007)
DifVar	0.935 (0.04)	0.0076 (3e-04)	-0.039 (0.011)	0.073 (0.007)
DifCov	0.929 (0.03)	0.0078 (3e-04)	-0.025 (0.012)	0.073 (0.007)
BothInt	0.948 (0.02)	0.0176 (0.0025)	0.039 (0.013)	0.106 (0.01)
AugInt	0.963 (0.02)	0.0201 (0.0028)	-0.002 (0.012)	0.11 (0.012)

3.6 Real Data Application

We assess our methods for the outcome estimation at the analysis stage on data from the PRagmatic trial of Video Education in Nursing Homes (PROVEN) trial (Mor et al., 2017). PROVEN presented individuals assigned to the active intervention an advanced directive video and examined its effect on an individual's propensity to transfer from a nursing home to the hospital. Assignment to the control intervention did not view any video. The trial was implemented across 360 nursing homes and it was the job of nursing home staff to ensure patients assigned to the active intervention were shown the advanced directive video. Part-way through the trial, some nursing homes assigned to the active intervention began receiving weekly phone calls reminding staff to show the advanced directive video.

For any individual i in the trial, we define the advanced directive video as the standard active intervention, $Z_i(1,0)$, and compliance to this intervention, $W_i(1,0)$, as the act of being shown the advanced directive video. We define the advanced directive video with weekly reminders to be shown the video as the augmented active intervention, $Z_i(1,1)$, and compliance to this intervention, $W_i(1,1)$, again as the act of being shown the advanced directive video. Thus, we let $Z_i(0,0)$ be assignment to control with assumed one-sided noncompliance, $W_i(0,0) = 1$, $\forall i$. We let the outcome of interest $Y_i(Z_i) = 1$ if patient i transferred to hospital under intervention assignment Z_i, and 0 otherwise. We denote the covariates for patient i, $\mathbf{X}_i = \{X_{ip}\}$ for $p = 1, \ldots, 12$. The covariates we consider for each individual i are, age, race identifying as Black, having experienced heart failure, having Alzheimer's disease, having non-Alzheimer's dementia, having asthma/COPD, mortality risk score, ADL score, nursing home total beds, the nursing home healthcare system, the nursing home hospitalization rate the year prior to intervention assignment, and if the nursing home was considered poor. We define our estimand of interest $\tau = ITT^A$ from Section 3.2.3. We evaluate our models with the estimate of the estimand, $\hat{\tau}$, and its corresponding estimated variance, $Var(\hat{\tau})$.

We define one Bayesian imputation-based model and two non-imputation-based estimation procedures for the analysis of the PROVEN data. We define the two non-imputation-based estimation procedures identically to Section 3.5.2 where AugInt $:= \widehat{ITT^{A1}}$ and BothInt $:= \widehat{ITT^{A2}}$. For the Bayesian imputation-based model, for each unit i, we assume independent Probit models for the distributions of potential compliance statuses, $p(\mathbf{W}_i(Z) \mid \mathbf{X}_i, \boldsymbol{\theta}) = p(W_i(1,0) \mid \mathbf{X}_i, \boldsymbol{\theta}) p(W_i(1,1) \mid \mathbf{X}_i, \boldsymbol{\theta})$. Here, each potential outcome is binary and we utilize Probit models to define $p(\mathbf{Y}_i(Z) \mid \mathbf{W}_i(Z), \mathbf{X}_i, \boldsymbol{\theta})$ for all units i. For each intervention assignment, Z_i, we let $E[Y_i(Z_i) \mid \mathbf{W}_i(Z), \mathbf{X}_i, \boldsymbol{\theta}] = \Phi(\beta_{Z_i,0} + \beta_{Z_i,1} W_i(1,0) + \beta_{Z_i,2} W_i(1,1) + \beta_{Z_i,1} X_{i1} + \cdots + \beta_{Z_i,12} X_{i12})$. We denote the Bayesian model as Indep which assumes conditional independence of the potential outcomes given potential compliance statuses, covariates, and parameters. We assume independent prior distributions of parameters where $p(\beta_{uv}) \propto 1$ for all $u \in \{p, q,\}$ and $v \in \{0, 1, 2\}$ and $p(\beta_{Z_i,r}) \propto 1$ for all Z_i and $r \in \{0, 1, 2, 3, 4\}$.

Table 3.8: PROVEN data analysis for the two Bayesian imputation-based models and the two non-imputation-based models, with model evaluation metrics as the estimated ITT^A and estimated variance of ITT^A.

	Model		
Metric	Indep	BothInt	AugInt
$\hat{\tau}$	-0.0445	0.0028	-0.0492
$Var(\hat{\tau})$	0.00073	0.00012	0.00083

Table 3.8 presents the point estimate and estimated variance of ITT^A for each estimation procedure. Indep and AugInt produce similar point estimates identifying an approximate five percent significant decrease in an individual's propensity to transfer to hospital care from their nursing home under the augmented intervention. Opposingly, BothInt results in an approximate one-third percent significant increase in an individual's propensity to transfer to hospital care from their nursing home under the active intervention arms.

Similar to the simulation study, BothInt has improved precision compared to AugInt, approximately one-seventh as large. However, the point estimates of the non-imputation-based estimators result in opposing conclusions. Indep produces a similar point estimate and conclusion compared to AugInt, with increased precision.

3.7 Discussion

We have proposed a three step adaptive trial to address noncompliance in the active intervention arms of a study at the design stage and the analysis stage of the trial. This project is rooted in the importance of addressing noncompliance prior to trial completion and the lack of literature targeting this issue.

We have extended work by Berry et al. (2010) to propose estimation procedures to estimate rate of compliance to the standard intervention at a pre-defined point of the trial, while keeping the outcome of interest blinded. At the design stage of the trial, these procedures establish decision rules and flexible decision criteria in order to maintain the standard intervention when compliance is high or switch to the augmented intervestion when compliance is low. Investigators set a priori thresholds about compliance, corresponding to trial limitations or preferences with power and sample size, defined during the design stage.

During the analysis stage, our compliance rate estimation procedures present three separate methods in which investigators may choose. The Posterior Probability Method allows investigators to define decision rules based exclusively on the observed data and the predefined determined compliance thresholds. The Predictive Probability Method extends the Posterior Probability Method to incorporate the number of individuals remaining in the trial and to consider the possible number of compliers remaining in the trial. Finally, the Indifference Zone Method is generally the most liberal of the estimation methods but flexible enough to accommodate different levels of desired precision of the compliance estimate. Choosing a rate of compliance estimation method requires investigators to determine which set of assumptions they are comfortable making and which method best fits their preferences.

At the analysis stage following the completion of the trial, we have proposed a fully Bayesian imputation-based estimation procedure for the outcomes of interest in the presence of covariates and noncompliance to the active intervention arms. Our methodology retains all information collected from the trial, including the observed compliance and outcome for individuals assigned to the standard intervention arm, to help better estimate our model parameters. Our Bayesian imputation-based methods result in good frequentist properties through coverage probability, mean variance, mean bias, and mean absolute bias when compared to the proposed alternative frequentist estimators. Similar results were observed even after deliberate misspecifications to the assumptions of our Bayesian imputation-based methods through the data generating mechanism.

Among the Bayesian models, EqVar, DifVar, and DifCov perform similarly throughout the simulations and are typically superior in each evaluation metric relative to Indep and the non-imputation-based methods, AugInt and BothInt. Generally, Indep results in larger variance estimation due to the independence assumption between the potential outcomes conditional on the covariates and compliance statuses. This was shown in both simulation and analysis of real-trial data. However, Indep provides investigators a potentially desired conservative estimation approach, relative to the Bayesian models assuming positive correlation between potential primary outcomes. While the conditional independence of potential outcomes assumption may be reasonable, the simulations show that improvements in precision can be made when allowing for non-zero correlation. The equal variance assumption made by EqVar works well when this assumption holds but will not necessarily always be true. In recommending a model, the choice between DifVar and DifCov will depend on the investigators preferences. Both models result in similar evaluation metrics which generally improve upon the remaining models. However, DifVar utilizes the estimation of the correlation coefficient, ρ, which in practice has no observed data to estimate this parameter. This typically should not be an issue as the estimation of the ITT does not depend on ρ. But, it is possible some investigators wish to avoid the estimation of ρ altogether while still gaining the precision improvements over the Indep model. DifCov defines the covariance matrix entirely with parameters that can be estimated from observed data and assumes non-zero correlation. However, DifCov assumes that the conditional variance of the active interventions outcomes will always be larger than the conditional variance of the control outcomes. Although, this does not necessarily result in larger mean variance of the ITT compared to DifVar, as shown in the simulations. However, this property still may be undesirable to some investigators. It is therefore up to the investigator to decide which assumption they are willing to make in order to choose between DifVar and DifCov.

When analyzing the PROVEN data, the Bayesian imputation-based model resulted in a similar point estimate to AugInt, with an improvement in precision. As the simulation study identified, the Indep model produces the most conservative interval estimates of the Bayesian imputation-based methods examined.

The design stage compliance rate estimation method is limited by the assumption of a single interim analysis. Some investigators may wish to extend this method for two or more interim analyses, or possibly continuous monitoring of compliance. In doing so, issues may arise in Type 1 Error, and thus further simulations would be needed to inspect this issue. Additionally, the assumption to examine the rate of compliance to the standard intervention arm after observing between 10-20% of the total active interventions study population may be limiting. While some investigators may wish to observe a greater percentage of individuals in the standard intervention, this does come at the expense of worsened precision in estimating the ITT between the augmented and control interventions, which is the estimand of interest in our simulation study.

The analysis method is limited by simulation-based data generation, as there may be additional plausible scenarios in which to generate the data. However, as the results have similar trends across the configurations and simulation groups, the performance of the Bayesian imputation-based methods relative to the alternative procedures holds up well.

Both the design and analysis stage methods outlined are limited by dependence on the one-sided noncompliance assumption. Additional work would be required to test the performance of the compliance estimation and the analysis methods when this assumption is violated. Potentially more beneficial would be the extension of these methods to the two-sided noncompliance setting.

In conclusion, when there is the possibility for treatment noncompliance in a trial, our methodology gives investigators the ability to prepare for noncompliance before the trial begins, while the trial is ongoing, and after the trial has been completed. Our methods are flexible in that they allow for investigators to define different preferences, assumptions, and various estimands of interest.

Chapter 3 References

J. Angrist, G. Imbens, and D. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. doi: https://doi.org/10.2307/2291629.

P. Armitage. Interim analysis in clinical trials. *Statistics in Medicine*, 10(6):925–937, 1991. doi: 10.1002/sim.4780100613.

P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, 132(2):235, 1969. doi: 10.2307/2343787.

Y. Benjamini and H. Braun. John w. tukey's contributions to multiple comparisons. *The Annals of Statistics*, 30(6), Dec 2002. doi: 10.1214/aos/1043351247.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, Jan 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

D. A. Berry. Interim analyses in clinical trials: Classical vs. bayesian approaches. *Statistics in Medicine*, 4(4):521–526, 1985. doi: 10.1002/sim.4780040412.

S. M. Berry, B. P. Carlin, J. J. Lee, and P. Muller. *Bayesian adaptive methods for clinical trials*, volume 38 of *CRC Biostatistics Series*. Chapman and Hall, 2010.

M. Chang. *Adaptive design theory and implementation using SAS and R.* CRC Press, 2014.

S.-C. Chow, M. Chang, and A. Pong. Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics*, 15(4):575–591, 2005. doi: 10.1081/bip-200062277.

I. Ford and J. Norrie. Pragmatic trials. *New England Journal of Medicine*, 375(5):454–463, 2016. doi: 10.1056/nejmra1510059.

C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. *Biometrics*, 58 (1):21–29, 2002. doi: 10.1111/j.0006-341x.2002.00021.x.

V. Gamerman, T. Cai, and A. Elsäßer. Pragmatic randomized clinical trials: Best practices and statistical guidance. *Health Services and Outcomes Research Methodology*, 19(1): 23–35, 2019. doi: 10.1007/s10742-018-0192-5.

N. L. Geller and S. J. Pocock. Interim analyses in randomized clinical trials: Ramifications and guidelines for practitioners. *Biometrics*, 43(1):213, 1987. doi: 10.2307/2531962.

M. Godwin, L. Ruhland, I. Casson, S. MacDonald, D. Delva, R. Birtwhistle, M. Lam, and R. Seguin. Pragmatic controlled clinical trials in primary care: The struggle between external and internal validity. *BMC Medical Research Methodology*, 3(1), 2003. doi: 10.1186/1471-2288-3-28.

D. Guidance. "adaptive designs for clinical trials of drugs and biologics guidance, adaptive designs for clinical trials of drugs and biologics." center for biologics evaluation and research, 2018. URL `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry`.

M. A. Hernán and J. M. Robins. Per-protocol analyses of pragmatic trials. *New England Journal of Medicine*, 377(14):1391–1398, 2017. doi: 10.1056/nejmsm1605385.

G. L. Hickey, S. W. Grant, J. Dunning, and M. Siepe. Statistical primer: Sample size and power calculations—why, when and how?†. *European Journal of Cardio-Thoracic Surgery*, 54(1):4–9, 2018. doi: 10.1093/ejcts/ezy169.

Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988. doi: 10.1093/biomet/75.4.800.

G. W. Imbens and D. B. Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25(1), 1997. doi: 10.1214/aos/1034276631.

G. W. Imbens and D. B. Rubin. *Causal inference for statistics, social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

H. Jin and D. B. Rubin. Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, 103(481):101–111, 2008. doi: 10.1198/016214507000000347.

J. M. Lachin. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, 2(2):93–113, 1981. doi: 10.1016/0197-2456(81)90001-5.

J. M. Lachin and M. A. Foulkes. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics*, 42(3):507, 1986. doi: 10.2307/2531201.

Q. Long, R. J. Little, and X. Lin. Estimating causal effects in trials involving multitreatment arms subject to non-compliance: A bayesian framework. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 59(3):513–531, 2010. doi: 10.1111/j.1467-9876.2009.00709.x.

K. McPherson. On choosing the number of interim analyses in clinical trials. *Statistics in Medicine*, 1(1):25–36, 1982. doi: 10.1002/sim.4780010105.

F. Mealli and D. B. Rubin. Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. *Health Services and Outcomes Research Methodology*, 3:225–232, Dec 2002. doi: https://doi.org/10.1023/A:1025802028890.

V. Mor, A. E. Volandes, R. Gutman, C. Gatsonis, and S. L. Mitchell. Pragmatic trial of video education in nursing homes: The design and rationale for a pragmatic cluster randomized trial in the nursing home setting. *Clinical Trials*, 14(2):140?151, 2017. doi: 10.1177/1740774516685298.

N. A. Patsopoulos. A pragmatic view on pragmatic trials. *Dialogues in Clinical Neuroscience*, 13(2):217–224, 2011. doi: 10.31887/dcns.2011.13.2/npatsopoulos.

S. J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977. doi: 10.1093/biomet/64.2.191.

S. J. Pocock, M. D. Hughes, and R. J. Lee. Statistical problems in the reporting of clinical trials. *New England Journal of Medicine*, 317(7):426–432, Aug 1987. doi: 10.1056/nejm198708133170706.

M. Roland and D. J. Torgerson. Understanding controlled trials: What are pragmatic trials? *BMJ*, 316(7127):285–285, 1998. doi: 10.1136/bmj.316.7127.285.

D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978. doi: 10.1214/aos/1176344064.

D. B. Rubin. Comment on randomization analysis of experimental data: The fisher randomization test. *Journal of the American Statistical Association*, 75(371):591–593, 1980. doi: 10.2307/2287649.

D. B. Rubin. Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25:279–292, 1990. doi: 10.1016/0378-3758(90)90077-8.

D. Schwartz and J. Lellouch. Explanatory and pragmatic attitudes in therapeutical trials. *Journal of Chronic Diseases*, 20(8):637–648, 1967. doi: 10.1016/0021-9681(67)90041-0.

T. R. Ten Have, S.-L. T. Normand, S. M. Marcus, C. H. Brown, P. Lavori, and N. Duan. Intent-to-treat vs. non-intent-to-treat analyses under treatment non-adherence in mental health randomized trials. *Psychiatric Annals*, 38(12), 2008. doi: 10.3928/00485713-20081201-10.

J. W. Tukey. The philosophy of multiple comparisons. *Statistical Science*, 6(1), Feb 1991. doi: 10.1214/ss/1177011945.

J. H. Ware and M. B. Hamel. Pragmatic trials — guides to better patient care? *New England Journal of Medicine*, 364(18):1685–1687, 2011. doi: 10.1056/nejmp1103502.

J. Wittes. Sample size calculations for randomized controlled trials. *Epidemiologic Reviews*, 24(1):39–53, 2002. doi: 10.1093/epirev/24.1.39.

M. G. Zuidgeest, P. M. Welsing, G. J. van Thiel, A. Ciaglia, R. Alfonso-Cristancho, L. Eckert, M. J. Eijkemans, and M. Egger. Series: Pragmatic trials and real world evidence: Paper 5. usual care and real life comparators. *Journal of Clinical Epidemiology*, 90:92–98, 2017. doi: 10.1016/j.jclinepi.2017.07.001.

3.8 CHAPTER 3 Appendix

3.8.1 Sample Size Estimation

One-Sided Noncompliance Power Simulation

Table 3.9: The true numerical Power, $1-\beta$, of a trial given N_T from Equation (3.2) using $1-\beta_0$ with $\alpha = 0.05$, and ES, and the one-sided true compliance p_0. Each configuration had the data simulated 10,000 times.

	$ES = 0.2$			$ES = 0.5$			$ES = 0.8$		
$1-\beta_0$	$p_0 = 0.7$	$p_0 = 0.8$	$p_0 = 0.9$	$p_0 = 0.7$	$p_0 = 0.8$	$p_0 = 0.9$	$p_0 = 0.7$	$p_0 = 0.8$	$p_0 = 0.9$
0.80	0.801	0.805	0.806	0.787	0.800	0.795	0.789	0.795	0.793
0.85	0.847	0.846	0.852	0.839	0.849	0.853	0.833	0.839	0.846
0.90	0.900	0.895	0.900	0.895	0.894	0.895	0.887	0.880	0.890

Two-Sided Noncompliance Power Simulation Using Equation (1)

Table 3.10: The true numerical Power, $1-\beta$, of a trial given N_T from Equation (3.2) using $1-\beta_0$ with $\alpha = 0.05$, and ES, and the two-sided compliance in both groups is p_0. Each configuration had the data simulated 10,000 times.

	$ES = 0.2$			$ES = 0.5$			$ES = 0.8$		
$1-\beta_0$	$p_0 = 0.7$	$p_0 = 0.8$	$p_0 = 0.9$	$p_0 = 0.7$	$p_0 = 0.8$	$p_0 = 0.9$	$p_0 = 0.7$	$p_0 = 0.8$	$p_0 = 0.9$
0.80	0.354	0.555	0.700	0.350	0.550	0.693	0.334	0.529	0.692
0.85	0.407	0.608	0.762	0.388	0.597	0.747	0.370	0.572	0.736
0.90	0.442	0.678	0.821	0.442	0.665	0.810	0.418	0.641	0.804

Two-Sided Noncompliance Power Simulation Lachin and Foulkes (1986)

Table 3.11: The true numerical Power, $1-\beta$, of a trial given N_T from Equation (3.2) with $p_0 = 1 - q_T - q_C$ using $1-\beta_0$ with $\alpha = 0.05$, and ES, and compliances $q_T = q_C$. Each configuration had the data simulated 10,000 times.

	$ES = 0.2$			$ES = 0.5$			$ES = 0.8$		
$1-\beta_0$	$q_T = 0.7$	$q_T = 0.8$	$q_T = 0.9$	$q_T = 0.7$	$q_T = 0.8$	$q_T = 0.9$	$q_T = 0.7$	$q_T = 0.8$	$q_T = 0.9$
0.80	0.800	0.804	0.801	0.785	0.793	0.793	0.753	0.765	0.778
0.85	0.845	0.847	0.849	0.831	0.839	0.841	0.808	0.813	0.829
0.90	0.900	0.898	0.902	0.885	0.890	0.895	0.870	0.870	0.887

3.8.2 Type 1 Error

Analytical Type 1 Error

Define the null hypothesis, H_0, such that $H_0 : Y_i(0,0) = Y_i(1,0) = Y_i(1,1)$. Recall, the estimand of interest, $E(Y_i(1,1)) - E(Y_i(0,0))$. Additionally, recall the number of individuals assigned to the standard intervention, N_S, the augmented intervention, N_A, and the control intervention, N_C, such that $N_T = N_S + N_A = N_C$, and the number of standard individuals observed prior to compliance rate examination, n_S. We define the following estimator for the estimand of interest,

$$\overline{Y}_\Delta = \frac{1}{N_T} \sum_{i:T_i=1} \widehat{Y}_i(1,1) - \frac{1}{N_C} \sum_{i:T_i=0} Y_i^{obs}$$

where

$$\widehat{Y}_i(1,1) = \begin{cases} Y_i(1,0) = Y_i^{obs} & \text{if } W_{n_S}^{obs} \geq \phi, i \leq n_S \\ Y_i(1,0) = Y_i^{obs} & \text{if } W_{n_S}^{obs} \geq \phi, i > n_S \\ \widetilde{Y}_i(1,1) & \text{if } W_{n_S}^{obs} < \phi, i \leq n_S \\ Y_i(1,1) = Y_i^{obs} & \text{if } W_{n_S}^{obs} < \phi, i > n_S \end{cases}$$

which we can write as follows,

$$\begin{aligned} \widehat{Y}_i(1,1) = & \mathbb{1}(W_{n_S}^{obs} \geq \phi) \cdot \mathbb{1}(i \leq n_S) \cdot Y_i^{obs} \\ & + \mathbb{1}(W_{n_S}^{obs} \geq \phi) \cdot \mathbb{1}(i > n_S) \cdot Y_i^{obs} \\ & + \mathbb{1}(W_{n_S}^{obs} < \phi) \cdot \mathbb{1}(i \leq n_S) \cdot \widetilde{Y}_i(1,1) \\ & + \mathbb{1}(W_{n_S}^{obs} < \phi) \cdot \mathbb{1}(i > n_S) \cdot Y_i^{obs} \end{aligned}$$

and where

$$\widetilde{Y}_i(1,1) \sim N\Bigg(W_i^{obs}\cdot\widetilde{W}_i(1,1)\cdot\mu_{11} + (1-W_i^{obs})\cdot\widetilde{W}_i(1,1)\cdot\mu_{01} + W_i^{obs}\cdot(1-\widetilde{W}_i(1,1))\cdot\mu_{10} + (1-W_i^{obs})\cdot(1-\widetilde{W}_i(1,1))\cdot\mu_{00}\ ,\ \sigma^2\Bigg)$$

with

$$\widetilde{W}_i(1,1) \sim Bernoulli\Bigg(\frac{1}{N_A}\sum_{i:Z_i=(1,1)} W_i^{obs}\Bigg).$$

The mean parameters of this mixture, $\mu_{11},\ \mu_{01},\ \mu_{10}$ and μ_{00}, represent the observed means of the active intervention individuals, conditional on observed compliance statuses. Formally,

$$\mu_{11} = \Bigg(\sum_{\substack{i\le n_S\\ Z_i=(1,0)}} W_i^{obs} + \sum_{\substack{i> n_S\\ Z_i=(1,1)}} W_i^{obs}\Bigg)^{-1}\Bigg[\sum_{\substack{i\le n_S, T_i=1\\ W_i^{obs}=1}} Y_i^{obs} + \sum_{\substack{i> n_S, T_i=1\\ W_i^{obs}=1}} Y_i^{obs}\Bigg]$$

$$\mu_{01} = \Bigg(\sum_{\substack{i\le n_S\\ Z_i=(1,0)}} (1-W_i^{obs}) + \sum_{\substack{i> n_S\\ Z_i=(1,1)}} W_i^{obs}\Bigg)^{-1}\Bigg[\sum_{\substack{i\le n_S, T_i=1\\ W_i^{obs}=0}} Y_i^{obs} + \sum_{\substack{i> n_S, T_i=1\\ W_i^{obs}=1}} Y_i^{obs}\Bigg]$$

$$\mu_{10} = \Bigg(\sum_{\substack{i\le n_S\\ Z_i=(1,0)}} W_i^{obs} + \sum_{\substack{i> n_S\\ Z_i=(1,1)}} (1-W_i^{obs})\Bigg)^{-1}\Bigg[\sum_{\substack{i\le n_S, T_i=1\\ W_i^{obs}=1}} Y_i^{obs} + \sum_{\substack{i> n_S, T_i=1\\ W_i^{obs}=0}} Y_i^{obs}\Bigg]$$

$$\mu_{00} = \Bigg(\sum_{\substack{i\le n_S\\ Z_i=(1,0)}} (1-W_i^{obs}) + \sum_{\substack{i> n_S\\ Z_i=(1,1)}} (1-W_i^{obs})\Bigg)^{-1}\Bigg[\sum_{\substack{i\le n_S, T_i=1\\ W_i^{obs}=0}} Y_i^{obs} + \sum_{\substack{i> n_S, T_i=1\\ W_i^{obs}=0}} Y_i^{obs}\Bigg]$$

We claim that $\frac{1}{M}\sum_{i=1}^{M}\widetilde{Y}_i(1,1) \rightarrow E(Y_i(1,1))$ as $M \rightarrow \infty$ and therefore $E(\widetilde{Y}_i(1,1)) = E(Y_i(1,1))$. Moreover, we claim that under H_0, $E(\overline{Y}_\Delta \mid W_{n_S}^{obs}(\phi), H_0) = 0$. First, when

$W^{obs}_{n_S} \geq \phi$ we have the following,

$$
\begin{aligned}
E\left(\overline{Y}_\Delta \;\middle|\; W^{obs}_{n_S} \geq \phi, H_0\right) &= E\left[\frac{1}{N_T}\sum_{i:T_i=1}\widehat{Y}_i(1,1) - \frac{1}{N_C}\sum_{i:T_i=0} Y^{obs}_i \;\middle|\; W^{obs}_{n_S} \geq \phi, H_0\right] \\
&= \frac{1}{N_T}\sum_{i:T_i=1} E\left[\widehat{Y}_i(1,1) \;\middle|\; W^{obs}_{n_S} \geq \phi, H_0\right] - \frac{1}{N_C}\sum_{i:T_i=0} E\left[Y^{obs}_i \;\middle|\; Z_i = (0,0), W^{obs}_{n_S} \geq \phi, H_0\right] \\
&= \frac{1}{N_T}\sum_{i:T_i=1} E\left[Y^{obs}_i \;\middle|\; Z_i = (1,0), W^{obs}_{n_S} \geq \phi, H_0\right] - \frac{1}{N_C}\sum_{i:T_i=0} E\left[Y^{obs}_i \;\middle|\; Z_i = (0,0), W^{obs}_{n_S} \geq \phi, H_0\right] \\
&= \frac{1}{N_T}\sum_{i:T_i=1} E\left[Y_i(1,0) \;\middle|\; W^{obs}_{n_S} \geq \phi, H_0\right] - \frac{1}{N_C}\sum_{i:T_i=0} E\left[Y_i(0,0) \;\middle|\; W^{obs}_{n_S} \geq \phi, H_0\right] \\
&= \frac{1}{N_T}\sum_{i:T_i=1} E\left[Y_i(0,0) \;\middle|\; W^{obs}_{n_S} \geq \phi, H_0\right] - \frac{1}{N_C}\sum_{i:T_i=0} E\left[Y_i(0,0) \;\middle|\; W^{obs}_{n_S} \geq \phi, H_0\right] \\
&= 0
\end{aligned}
$$

where the difference in expectations results in 0 because $N_T = N_C$ and $E(Y_i(0,0)) = E(Y_i(1,0)) = E(Y_i(1,1))$ for all i under H_0. Second, when $W^{obs}_{n_S} < \phi$ we have the following,

$$
\begin{aligned}
E\left(\overline{Y}_\Delta \;\middle|\; W^{obs}_{n_S} < \phi, H_0\right) &= E\left[\frac{1}{N_T}\sum_{i:T_i=1}\widehat{Y}_i(1,1) - \frac{1}{N_C}\sum_{i:T_i=0} Y^{obs}_i \;\middle|\; W^{obs}_{n_S} < \phi, H_0\right] \\
&= E\left[\frac{1}{N_T}\left(\sum_{\substack{i\leq n_S\\ T_i=1}}\widetilde{Y}_i(1,1) + \sum_{\substack{i> n_S\\ T_i=1}} Y^{obs}_i\right) - \frac{1}{N_C}\sum_{i:T_i=0} Y^{obs}_i \;\middle|\; W^{obs}_{n_S} < \phi, H_0\right] \\
&= \frac{1}{N_T}\left(\sum_{\substack{i\leq n_S\\ T_i=1}} E\left[\widetilde{Y}_i(1,1) \;\middle|\; W^{obs}_{n_S} < \phi, H_0\right] + \sum_{\substack{i> n_S\\ T_i=1}} E\left[Y^{obs}_i \;\middle|\; Z_i = (1,1), W^{obs}_{n_S} < \phi, H_0\right]\right) \\
&\quad - \frac{1}{N_C}\sum_{i:T_i=0} E\left[Y^{obs}_i \;\middle|\; Z_i = (0,0), W^{obs}_{n_S} < \phi, H_0\right] \\
&= \frac{1}{N_T}\left(\sum_{\substack{i\leq n_S\\ T_i=1}} E\left[Y_i(1,1) \;\middle|\; W^{obs}_{n_S} < \phi, H_0\right] + \sum_{\substack{i> n_S\\ T_i=1}} E\left[Y_i(1,1) \;\middle|\; W^{obs}_{n_S} < \phi, H_0\right]\right) \\
&\quad - \frac{1}{N_C}\sum_{i:T_i=0} E\left[Y_i(0,0) \;\middle|\; W^{obs}_{n_S} < \phi, H_0\right] \\
&= \frac{1}{N_T}\sum_{i:T_i=1} E\left[Y_i(1,1) \;\middle|\; W^{obs}_{n_S} < \phi, H_0\right] - \frac{1}{N_C}\sum_{i:T_i=0} E\left[Y_i(0,0) \;\middle|\; W^{obs}_{n_S} < \phi, H_0\right]
\end{aligned}
$$

$$= \frac{1}{N_T} \sum_{i:T_i=1} E\Big[Y_i(0,0) \;\Big|\; W_{n_S}^{obs} < \phi, H_0\Big] - \frac{1}{N_C} \sum_{i:T_i=0} E\Big[Y_i(0,0) \;\Big|\; W_{n_S}^{obs} < \phi, H_0\Big]$$

$$= 0$$

Now denote $S_{\Delta 10}$ and $S_{\Delta 11}$ as the sample standard deviations of our estimator, $\overline{Y}_{\Delta}$, conditional on $W_{n_S}^{obs} \geq \phi$ and $W_{n_S}^{obs} < \phi$, respectively. Therefore, under H_0, we assume

$$\overline{Y}_{\Delta} \sim N\Big(0\ ,\ \mathbb{1}(W_{n_S}^{obs} \geq \phi) \cdot S_{\Delta 10}^2 + \mathbb{1}(W_{n_S}^{obs} < \phi) \cdot S_{\Delta 11}^2\Big)$$

Using the entirety of the above, we may derive the Type 1 Error of our procedure as follows,

$$\begin{aligned}
&P(\text{Reject } H_0 \mid H_0) \\
&\quad = P(\text{Reject } H_0 \mid H_0, \mathbf{W}_{n_S}^{obs} \geq \phi) P(\mathbf{W}_{n_S}^{obs} \geq \phi) + P(\text{Reject } H_0 \mid H_0, \mathbf{W}_{n_S}^{obs} < \phi) P(\mathbf{W}_{n_S}^{obs} < \phi) \\
&\quad = \Big[P\Big(\overline{Y}_{\Delta} - Z_{1-\alpha/2} \cdot S_{\Delta 10} > 0 \;\Big|\; H_0, \mathbf{W}_{n_S}^{obs} \geq \phi\Big) + P\Big(\overline{Y}_{\Delta} + Z_{1-\alpha/2} \cdot S_{\Delta 10} < 0 \;\Big|\; H_0, \mathbf{W}_{n_S}^{obs} \geq \phi\Big)\Big] P\big(\mathbf{W}_{n_S}^{obs} \geq \phi\big) \\
&\qquad + \Big[P\Big(\overline{Y}_{\Delta} - Z_{1-\alpha/2} \cdot S_{\Delta 11} > 0 \;\Big|\; H_0, \mathbf{W}_{n_S}^{obs} < \phi\Big) + P\Big(\overline{Y}_{\Delta} + Z_{1-\alpha/2} \cdot S_{\Delta 11} < 0 \;\Big|\; H_0, \mathbf{W}_{n_S}^{obs} < \phi\Big)\Big] P\big(\mathbf{W}_{n_S}^{obs} < \phi\big) \\
&\quad = \Big[P\Big(\overline{Y}_{\Delta} > Z_{1-\alpha/2} \cdot S_{\Delta 10} \;\Big|\; H_0, \mathbf{W}_{n_S}^{obs} \geq \phi\Big) + P\Big(\overline{Y}_{\Delta} < -Z_{1-\alpha/2} \cdot S_{\Delta 10} \;\Big|\; H_0, \mathbf{W}_{n_S}^{obs} \geq \phi\Big)\Big] P\big(\mathbf{W}_{n_S}^{obs} \geq \phi\big) \\
&\qquad + \Big[P\Big(\overline{Y}_{\Delta} > Z_{1-\alpha/2} \cdot S_{\Delta 11} \;\Big|\; H_0, \mathbf{W}_{n_S}^{obs} < \phi\Big) + P\Big(\overline{Y}_{\Delta} < -Z_{1-\alpha/2} \cdot S_{\Delta 11} \;\Big|\; H_0, \mathbf{W}_{n_S}^{obs} < \phi\Big)\Big] P\big(\mathbf{W}_{n_S}^{obs} < \phi\big) \\
&\quad = \Big[P\Big(\frac{\overline{Y}_{\Delta} - 0}{S_{\Delta 10}} > Z_{1-\alpha/2} \;\Big|\; H_0, \mathbf{W}_{n_S}^{obs} \geq \phi\Big) + P\Big(\frac{\overline{Y}_{\Delta} - 0}{S_{\Delta 10}} < -Z_{1-\alpha/2} \;\Big|\; H_0, \mathbf{W}_{n_S}^{obs} \geq \phi\Big)\Big] P\big(\mathbf{W}_{n_S}^{obs} \geq \phi\big) \\
&\qquad + \Big[P\Big(\frac{\overline{Y}_{\Delta} - 0}{S_{\Delta 11}} > Z_{1-\alpha/2} \;\Big|\; H_0, \mathbf{W}_{n_S}^{obs} < \phi\Big) + P\Big(\frac{\overline{Y}_{\Delta} - 0}{S_{\Delta 11}} < -Z_{1-\alpha/2} \;\Big|\; H_0, \mathbf{W}_{n_S}^{obs} < \phi\Big)\Big] P\big(\mathbf{W}_{n_S}^{obs} < \phi\big) \\
&\quad = \Big[1 - \Phi\Big(Z_{1-\alpha/2}\Big) + \Phi\Big(-Z_{1-\alpha/2}\Big)\Big] P\big(\mathbf{W}_{n_S}^{obs} \geq \phi\big) + \Big[1 - \Phi\Big(Z_{1-\alpha/2}\Big) + \Phi\Big(-Z_{1-\alpha/2}\Big)\Big] P\big(\mathbf{W}_{n_S}^{obs} < \phi\big) \\
&\quad = \Big[1 - \Phi\Big(Z_{1-\alpha/2}\Big) + \Phi\Big(Z_{\alpha/2}\Big)\Big] P\big(\mathbf{W}_{n_S}^{obs} \geq \phi\big) + \Big[1 - \Phi\Big(Z_{1-\alpha/2}\Big) + \Phi\Big(Z_{\alpha/2}\Big)\Big] P\big(\mathbf{W}_{n_S}^{obs} < \phi\big) \\
&\quad = \Big[1 - 1 + \alpha/2 + \alpha/2\Big] P\big(\mathbf{W}_{n_S}^{obs} \geq \phi\big) + \Big[1 - 1 + \alpha/2 + \alpha/2\Big] P\big(\mathbf{W}_{n_S}^{obs} < \phi\big)
\end{aligned}$$

$$= \alpha \cdot P\big(\mathbf{W}_{n_S}^{obs} \geq \phi\big) + \alpha \cdot P\big(\mathbf{W}_{n_S}^{obs} < \phi\big)$$

$$= \alpha \cdot \Big[P\big(\mathbf{W}_{n_S}^{obs} \geq \phi\big) + P\big(\mathbf{W}_{n_S}^{obs} < \phi\big)\Big]$$

$$= \alpha \cdot \Big[1\Big]$$

$$= \alpha$$

Thus we have that the Type 1 Error remains at our pre-defined rate of α regardless of the CR results and estimation itself.

Numerical Type 1 Error

Table 3.12: Numerically derived Type 1 Error of our three-step adaptive trial where N_T is estimated from Equation 3.2 using $ES = \frac{1-0}{\sigma}$, $\alpha = 0.05$, $\beta = 0.2$, and p_0. ROC estimation follows the posterior probability method with $\delta_L = p_0 - 0.05$, $\delta_U = p_0 + 0.05$, and $\pi_S = \pi_M = 0.85$. ITT estimations follows $\widehat{ITT}^{A1}$ when augmented intervention is implemented and $\widehat{ITT}^{S}$ when augmented intervention is not implemented, both defined in Section 3.4.2. Similar results are obtained using Indep for ITT estimation.

	$p_0 = 0.50$				$p_0 = 0.75$			
	$\frac{n_S}{N_T} = 0.1$		$\frac{n_S}{N_T} = 0.2$		$\frac{n_S}{N_T} = 0.1$		$\frac{n_S}{N_T} = 0.2$	
p_{true}	$ES = 0.25$	$ES = 0.50$	$ES = 0.25$	$ES = 0.50$	$ES = 0.25$	$ES = 0.50$	$ES = 0.25$	$ES = 0.50$
$0.75p_0$	0.0484	0.0497	0.0501	0.0494	0.0499	0.0545	0.0494	0.0495
$0.90p_0$	0.0498	0.0490	0.0476	0.0503	0.0542	0.0489	0.0470	0.0483
p_0	0.0511	0.0501	0.0498	0.0516	0.0531	0.0469	0.0523	0.0525

3.8.3 Compliance Estimation Example

For a trial design with assumed $p_0 = 0.5$, $n_S = 10$, $N_T = 50$, $ES = 1.3$, $\alpha = 0.05$, Table 3.13 presents the Power of the trial given the true compliance p_{true}.

Table 3.13: Trial Power, $1 - \beta$, given $n_S = 10$, $N_T = 50$, $ES = 1.3$, $\alpha = 0.05$ and p_{true}.

	p_{true}									
	0.30	0.35	0.40	0.425	0.45	0.50	0.55	0.60	0.65	0.70
$1 - \beta$	0.496	0.624	0.739	0.789	0.833	0.901	0.947	0.974	0.988	0.995

Posterior Probability Method

Table 3.14: Posterior probability that the compliance probability is greater than δ_U or less than δ_L given the observed compliance $\sum_{i=1}^{n_S} W_i(1,0)$.

	δ_L					δ_U				
$\sum_{i=1}^{n_S} W_i(1,0)$	0.30	0.35	0.40	0.425	0.45	0.50	0.55	0.60	0.65	0.70
0	0.980	0.991	0.996	0.998	0.999	0.001	0	0	0	0
1	0.887	0.939	0.970	0.979	0.986	0.006	0.002	0.001	0	0
2	0.687	0.800	0.881	0.911	0.935	0.033	0.015	0.006	0.002	0.001
3	0.430	0.574	0.704	0.760	0.809	0.113	0.061	0.029	0.012	0.004
4	0.210	0.332	0.467	0.536	0.603	0.274	0.174	0.099	0.050	0.022
5	0.078	0.149	0.246	0.304	0.367	0.500	0.367	0.246	0.149	0.078
6	0.022	0.050	0.099	0.133	0.174	0.726	0.603	0.467	0.332	0.210
7	0.004	0.012	0.029	0.043	0.061	0.887	0.809	0.704	0.574	0.430
8	0.001	0.002	0.006	0.009	0.015	0.967	0.935	0.881	0.800	0.687
9	0	0	0.001	0.001	0.002	0.994	0.986	0.970	0.939	0.887
10	0	0	0	0	0	1	0.999	0.996	0.991	0.980

Indifference Zone Method, $\alpha = 0.05$, 95% Credible Interval

Table 3.15: Posterior credible intervals of the compliance probability given the observed compliance $\sum_{i=1}^{n_S} W_i(1,0)$ and an indicator if $p_L < \delta_L$ and $p_U > \delta_U$.

		δ_L					δ_U				
$\sum_{i=1}^{n_S} W_i(1,0)$	CI (p_L, p_U)	0.30	0.35	0.40	0.425	0.45	0.50	0.55	0.60	0.65	0.70
0	(0.002, 0.285)	1	1	1	1	1	0	0	0	0	0
1	(0.023, 0.413)	1	1	1	1	1	0	0	0	0	0
2	(0.060, 0.518)	1	1	1	1	1	1	0	0	0	0
3	(0.109, 0.610)	1	1	1	1	1	1	1	1	0	0
4	(0.167, 0.692)	1	1	1	1	1	1	1	1	1	0
5	(0.234, 0.766)	1	1	1	1	1	1	1	1	1	1
6	(0.308, 0.833)	0	1	1	1	1	1	1	1	1	1
7	(0.390, 0.891)	0	0	1	1	1	1	1	1	1	1
8	(0.482, 0.940)	0	0	0	0	0	1	1	1	1	1
9	(0.587, 0.977)	0	0	0	0	0	1	1	1	1	1
10	(0.715, 0.998)	0	0	0	0	0	1	1	1	1	1

Predictive Probability Method, $\theta_T = 0.90$

Table 3.16: Predictive probability that the compliance probability will be greater than δ_U or less than δ_L given the observed compliance $\sum_{i=1}^{n_S} W_i(1,0)$ and the future $N_T - n_S = 40$ compliance observations.

	δ_L					δ_U				
$\sum_{i=1}^{n_S} W_i(1,0)$	0.30	0.35	0.40	0.425	0.45	0.50	0.55	0.60	0.65	0.70
0	0.952	0.982	0.991	0.994	0.996	0	0	0	0	0
1	0.756	0.880	0.930	0.947	0.961	0	0	0	0	0
2	0.440	0.642	0.752	0.797	0.837	0.003	0.001	0	0	0
3	0.173	0.347	0.475	0.538	0.599	0.020	0.006	0.002	0.001	0
4	0.044	0.131	0.218	0.269	0.325	0.087	0.033	0.015	0.005	0.001
5	0.007	0.033	0.069	0.095	0.125	0.249	0.125	0.069	0.033	0.007
6	0.001	0.005	0.015	0.022	0.033	0.503	0.325	0.218	0.131	0.044
7	0	0.001	0.002	0.003	0.006	0.759	0.599	0.475	0.347	0.173
8	0	0	0	0	0.001	0.922	0.837	0.752	0.642	0.440
9	0	0	0	0	0	0.986	0.961	0.930	0.880	0.756
10	0	0	0	0	0	0.999	0.996	0.991	0.982	0.952

3.8.4 Covariance and Correlation Derivations

Covariance

Consider a model where $Y_i(1,0) = g(Y_i(0,0), \mathbf{W}_i(Z), T_i, \epsilon_i)$ and $Y_i(1,1) = h(Y_i(0,0), \mathbf{W}_i(Z), T_i, \delta_i)$ for some functions $g(\cdot)$ and $h(\cdot)$, where $Y_i(0,0) \sim \mathcal{N}(\mu_0, \sigma_0^2)$, $T_i \sim \mathcal{N}(\mu_1, \tau_1^2)$, $\epsilon_i \sim \mathcal{N}(0, \tau_\epsilon^2)$, and $\delta_i \sim \mathcal{N}(0, \tau_\delta^2)$, and each variable mutually independent for all i. Table 3.17 defines $g(\cdot)$ and $h(\cdot)$.

Table 3.17: Potential response outcomes assumptions conditional on potential compliance outcomes, for a model which sets the potential outcomes as a function of one another.

$W_i(1,0)$	$W_i(1,1)$	$Y_i(1,0)$	$Y_i(1,1)$
1	1	$Y_i(0,0) + T_i + \delta_i$	$Y_i(0,0) + T_i + \epsilon_i$
0	0	$Y_i(0,0) + \delta_i$	$Y_i(0,0) + \epsilon_i$
0	1	$Y_i(0,0) + \delta_i$	$Y_i(0,0) + T_i + \epsilon_i$
1	0	$Y_i(0,0) + T_i + \delta_i$	$Y_i(0,0) + \epsilon_i$

Conditional on the potential compliance outcomes, we can derive the covariance of each pair of potential response outcomes using the assumptions above. For example we derive the covariance of $Y_i(1,0)$ and $Y_i(0,0)$ when $W_i(1,0)=1$ and $W_i(1,1)=1$. The derivations follow similarly for the other covariance values. For notation convenience, we let $\mathbf{W_{i11}} = \big(W_i(1,0)=1, W_i(1,1)=1\big)$.

$$
\begin{aligned}
&Cov\Big(Y_i(1,0), Y_i(0,0)|\mathbf{W_{i11}}\Big) \\
&\quad = E\Big(Y_i(1,0)\cdot Y_i(0,0)|\mathbf{W_{i11}}\Big) - E\Big(Y_i(1,0)|\mathbf{W_{i11}}\Big)\cdot E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big) \\
&\quad = E\Big(\big[Y_i(0,0)+T_i+\delta_i\big]\cdot Y_i(0,0)|\mathbf{W_{i11}}\Big) \\
&\qquad - E\Big(\big[Y_i(0,0)+T_i+\delta_i\big]|\mathbf{W_{i11}}\Big)\cdot E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big) \\
&\quad = E\Big(\big[Y_i(0,0)^2+T_iY_i(0,0)+\delta_iY_i(0,0)\big]|\mathbf{W_{i11}}\Big) \\
&\qquad - E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big)^2 - E\Big(T_i|\mathbf{W_{i11}}\Big)\cdot E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big) \\
&\qquad - E\Big(\delta_i|\mathbf{W_{i11}}\Big)\cdot E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big) \\
&\quad = E\Big(Y_i(0,0)^2|\mathbf{W_{i11}}\Big) + E\Big(T_iY_i(0,0)|\mathbf{W_{i11}}\Big) \\
&\qquad + E\Big(\delta_iY_i(0,0)|\mathbf{W_{i11}}\Big) - E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big)^2 \\
&\qquad - E\Big(T_i|\mathbf{W_{i11}}\Big)\cdot E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big) - E\Big(\delta_i|\mathbf{W_{i11}}\Big)\cdot E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big) \\
&\quad = E\Big(Y_i(0,0)^2|\mathbf{W_{i11}}\Big) + E\Big(T_i|\mathbf{W_{i11}}\Big)\cdot E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big) \\
&\qquad + E\Big(\delta_i|\mathbf{W_{i11}}\Big)\cdot E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big) - E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big)^2 \\
&\qquad - E\Big(T_i|\mathbf{W_{i11}}\Big)\cdot E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big) - E\Big(\delta_i|\mathbf{W_{i11}}\Big)\cdot E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big) \\
&\quad = E\Big(Y_i(0,0)^2|\mathbf{W_{i11}}\Big) - E\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big)^2 \\
&\quad = Var\Big(Y_i(0,0)|\mathbf{W_{i11}}\Big) = \sigma_0^2
\end{aligned}
$$

Correlation

Under the model assumption outlined in Appendix 3.8.4, we can similarly derive the correlation of two potential response outcomes,

$$\begin{aligned}\rho_{Y_i(1,0),Y_i(0,0)|\mathbf{W_{i11}}} &= \frac{Cov\Big(Y_i(1,0), Y_i(0,0)|\mathbf{W_{i11}}\Big)}{\sqrt{Var\big(Y_i(1,0)|\mathbf{W_{i11}}\big)}\sqrt{Var\big(Y_i(0,0)|\mathbf{W_{i11}}\big)}} \\ &= \frac{\sigma_0^2}{\sqrt{\big(\sigma_0^2+\tau_1^2+\tau_\delta^2\big)}\cdot\sigma_0} \\ &= \frac{\sigma_0}{\sqrt{\big(\sigma_0^2+\tau_1^2+\tau_\delta^2\big)}}\end{aligned}$$

3.9 Compliance Model Comparison

Table 3.18: Model comparison between conditional independence of the compliance statuses (IndepW) against the assumption that compliers to the standard intervention always comply to the augmented intervention (CorrW). Four configurations each replicated 100 times and presented as the ITT estimate, its sampling variance, and its mean bias. Corr is the setting defining the compliance generation true to CorrW, Indep follows IndepW, BadAug generates compliance such that augmented intervention has worse compliance than the standard intervention, and finally IncW improves compliance with no effect of covariates from the standard to augmented intervention.

Simulation	Model	$\hat{ITT}$	Var($\hat{ITT}$)	$\hat{ITT}$ Bias
Corr	CorrW	0.751	0.00606	0.0151
	IndepW	0.739	0.00612	0.0028
Indep	CorrW	0.606	0.00542	0.0103
	IndepW	0.600	0.00536	0.0042
BadAug	CorrW	0.303	0.00383	0.0385
	IndepW	0.288	0.00279	0.0239
IncW	CorrW	0.662	0.00606	0.0037
	IndepW	0.665	0.00623	0.0071

Chapter 4

Zero-inflated hierarchical generalized Dirichlet multinomial Bayesian regression model with cyclic splines for analysis of TDP-43 on the ALS disease spectrum

4.1 Introduction

Amyotrophic lateral sclerosis (ALS) is a genetically heterogeneous neurodegenerative disease. Approximately 10% of ALS affected individuals have other affected family member (familial ALS), the remaining individuals have no prior family history (sporadic ALS) (Mejzini et al., 2019). ALS is characterized by motor function impairment , and many ALS patients also experience cognitive and behavioral symptoms that resemble those of frontotemporal dementia (Rusina et al., 2021). The ALS disease spectrum can be characterized by the RNA-binding protein, TDP-43. This characterization exists in both familial and sporadic ALS (Fisher et al., 2023). Defining the role of TDP-43 in the neurodegeneration process can lead to treatment development of ALS.

The complexity and heterogeneity of ALS makes early diagnosis a challenge (Zarei et al., 2015). Moreover, there is no single diagnostic test for ALS, but only a combination of tests that are used to rule out other diagnoses (Zarei et al., 2015). There is currently no known treatment that stops or reverses the progression of ALS, but there are some treatments that can prolong survival, reduce the rate of decline, or help managing symptoms (Zarei et al., 2015). Understanding the genetics and molecular mechanisms underlying ALS has proven challenging because of the obstacles in acquiring living tissue or cells from the central nervous system of ALS patients (Mejzini et al., 2019). One way to learn about human biology and disease is by using animal models. Animal models enable researchers to manipulate genetics and environmental factors to investigate their roles in development, behavior, and health outcomes of a disease. Mice models are frequently used because of their phylogenetic proximity and physiological resemblance to humans. (Perlman, 2016). Knock-in mouse model is a mouse model in which specific genes are replaced or mutated (Doyle et al., 2011). These mouse models enable to clarify gene expression pattern, replace a gene with a related gene to study the effect of subtle variations between genes, or alter a gene to mimic human mutations that lead to a gain or change of function (Nilles and London, 2007). To investigate the contribution of TDP-43 to impaired cognition, White et al. (2018) created a knock-in mouse model with a human mutation in the mouse Tardbp gene.

Many studies that utilize mouse models to examine behavior, motor, and cognitive functions are generally performed using data that are collected over short time periods while the mice undergo a single test (Alfieri et al., 2014; Mar et al., 2013; Heath et al., 2015). For example, in the mouse model developed by Alfieri et al. (2014), one behavioral test assessed the grip strength of mice by having each mouse grasp a wire cage, and recording the time to fall from the cage to a soft bedding below. Commonly, multiple single tests are performed to change multiple behaviors. However, to perform comprehensive phenotypic analyses, it is preferable to consider multiple mouse behaviors simultaneously over longer duration. However, performing these tasks manually are generally costly and may not be reproducible.

The automated home-cage behavioral phenotyping of mice (ACBM) (Jhuang et al., 2010) is a computer vision system that is trained with manually annotated behaviors of interest and tracks these behaviors in freely behaving mice in cages. The ACBM enables investigators to record and analyze multiple behaviors over multiple days for multiple mice. ACBM is trained on manual annotation of behaviors, and is then used to predict the annotated behavior on future camera frames. In cross validation, the average agreement of ACBM with human annotations was 78% for individual behavior and 83% across all individual frames. Using the frame-level predicted annotation, the ACBM algorithm aggregates the data to complete hour by summarizing the number of seconds that a mouse perform a specific behavioral task.

Analysis of Variance (ANOVA) and repeated-measure ANOVA are commonly used statistical tools to compare two or more types of animals across multiple behavioral tasks (Alfieri et al., 2014; White et al., 2018; Heath et al., 2015). The ANOVA model may produce unreliable inference and predictions when the results of a test are correlated over time and within mice, as well as when the outcomes are non-Normally distributed (Monohan, 2008; Rice, 1995). The repeated measures ANOVA, attempts to address correlations within the same mouse, but may still produce unreliable results and predictions when trends are not linear over time (Mundo et al., 2022), or when the data is bounded (Hutmacher et al., 2011; Janicki, 2019). The results produced by the ACBM algorithm are correlated across behaviors, bounded, non-Normally distributed, and are repeated for each mouse across days and hours.

Following a Bayesian framework, we compare the goodness of fit of multiple models to data annotated by the ACBM algorithm over 5 days for 20 mice. We show that ANOVA, repeated measure ANOVA and transformed values repeated measures ANOVA models result in poor model fit. We propose a hierarchical zero-inflated generalized Dirichlet multinomial regression model (abbreviated ZIHGDM) (Tang and Chen, 2018) with cyclic splines to model the time mice spent performing certain behaviors at each circadian hour. This model can integrate periods in which mice do not perform certain behaviors, correlation between the time spent performing the different behaviors, repeated measurement for each mouse and day of the week, and cyclic effects of hours within a day. ZIHGDM outperforms existing models by providing accurate posterior predictive checks of multiple test statistics. Using ZIHGDM we identify phenotypic differences in behavior time between TDP-43 and WT mice at circadian hours that are not detected by the hierarchical Dirichlet regression model with cyclic splines or the hierarchical multivariate Normal-Normal model with cyclic splines.

The paper proceeds as follows: Section 2 provides a description of the mice data. Section 3 and 4 considers separate Bayesian ANOVA models and Bayesian multiple ANOVA models to analyze the data, and describes the procedures that were used to examine the fit of the models. Sections 5 and 6 describe hierarchical models and hierarchical spline models, respectively, as well as their fit with the data. Section 7 introduces the hierarchical Dirichlet regression model with cyclic splines. Section 8 presents the zero-inflated hierarchical generalized Dirichlet multinomial regression model with cyclic splines. Section 9 compares ZIHGDM to hierarchical Normal and Dirichlet models with cyclic splines. Section 10 provides discussion and conclusions.

4.2 Description of Data

Our study comprise of ACBM generated output on 20 mice, aged 180 days, across 5 days. Equal numbers of mice were recorded to have 10 TDP-43 and 10 WT (Wild-Type, control) mice. Each type of mouse had equal numbers of male and female mice. Each recording day consisted of 12 hours of light and darkness. The recording of each mouse was annotated for 9 possible behaviors at each second using ACBM (Jhuang et al., 2010). The 9 mutually exclusive behaviors included drinking, eating, eating by hand (EBH), grooming, hanging, rearing, resting, sniffing, and walking. Using the annotated data per second, the data was aggregated to the number of seconds in hour $j \in \{0, \ldots, 23\}$ on day $k \in \{1, \ldots, 6\}$ that mouse $i \in \{1, \ldots, 20\}$ was performing each of the 9 behaviors, $\mathbf{Y}_{ijk} = \{Y_{ijk1}, \ldots, Y_{ijk9}\}$, where $\sum_{m=1}^{9} Y_{ijkm} = 3600$ and $Y_{ijkm} \geq 0, \ \forall \ i, j, k$. Figure **??** presents the median observed behavior times across the 20 mice at each hour for all 9 behaviors, stratified by TDP-43 and WT mice.

For mouse i, define $T_i = 1$ for genotype TDP-43 and $T_i = 0$ for genotype WT, and define $G_i = 1$ for male and $G_i = 0$ for female mice, for all i. For each hour of observation j, define $H_{ijk} = H_j = j$. We suppress additional indices for T_i, G_i, and H_j due to redundancy. In general, for observation ijk, let $\mathbf{X}_{ijk}$ represent the vector of P covariates.

Each mouse was recorded for 120 hours. Because mice recording did not begin at hour 0 of day 1, 6 calendar days were used to complete these hours. Two mice were missing one hour of observation, one from the test and control group. This missing hour was not related to the 9 behaviors or any other variables in the dataset and were considered missing completely at random. The total number of observations, N, across all mice was 2398 hours. Let $\mathbf{Y} = \{\mathbf{Y}_{ijk}\}$ be the $N \times 9$ matrix of behavior times observed.

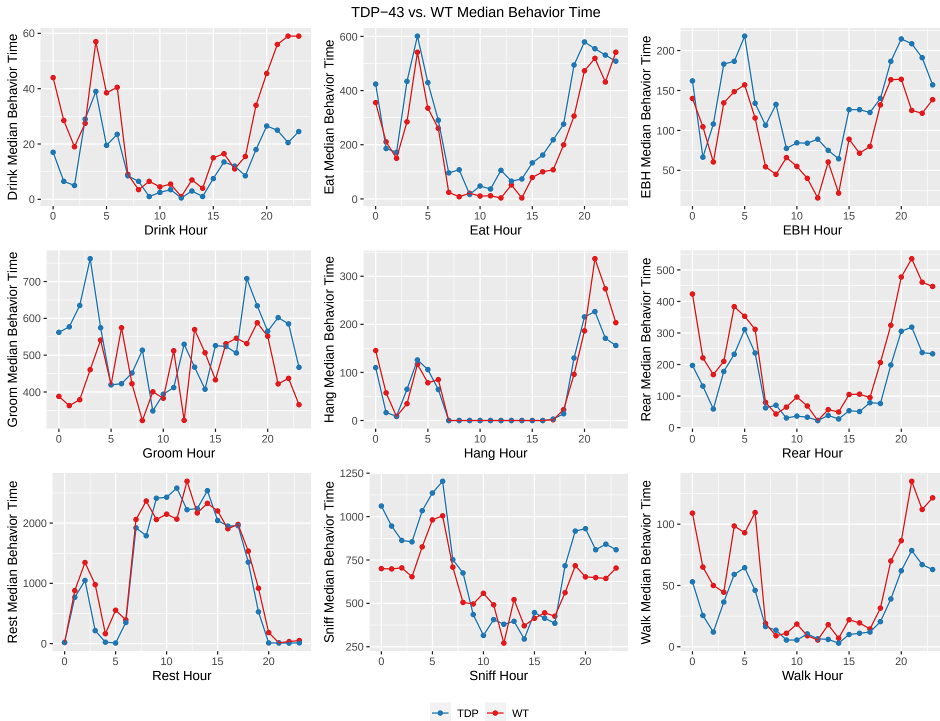

Figure 4.1: Median observed behavior time for each behavior at each hour, stratified by TDP-43 and WT mice

4.3 Bayesian ANOVA Model

Because we were interested in examining the fit of the different models using multiple test statistics as well as having the capability to extend the models with latent structures, we decided to rely on a full Bayesian approach.

To be consistent with previously implemented statistical analysis with ACBM data we relied on models that assumed that either Y_{ijkm} or a transformation of it are conditionally Normally distributed. The two-way ANOVA model considers the effects of two factors on an outcome of interest, where each factor has two or more levels (Rice, 1995). The observation produced by ACBM, Y_{ijkm}, are bounded between 0 and 1. To satisfy the ANOVA normality assumption we applied a transformation $g(x_i) = \left[1 + exp\left\{\frac{log(x_i + \mathbf{1}(x_i=0)) - \bar{x}}{\sigma_x}\right\}\right]^{-1}$ to Y_{ijkm}, where $\bar{x}$ and σ_x are the sample mean and sample standard deviation of x, respectively. Formally, Bayesian ANOVA model that we examined included an indicator variable for

TDP-43 genotype and the time of the day,

$$g(Y_{ijkm}) = \mu_m + \alpha_m T_i + \beta_{jm} + \epsilon_{ijkm} \tag{1}$$

where μ_m is the grand mean response for behavior m, α_m is the association between TDP-43 genotype and behavioral outcome m, β_{jm} is the effect of the hour of observation for $j \in \{1, \ldots, 23\}$, and ϵ_{ijkm} are independently distributed as $N(0, \sigma_m^2)$. To complete the Bayesian model we assume that $p(\mu_m, \alpha_m, \beta_{jm}, \sigma_m) = p(\mu_m|\sigma_m)p(\alpha_m|\sigma_m)p(\beta_{jm}|\sigma_m)p(\sigma_m)$, and $p(\mu_m|\sigma_m) \propto 1$, $p(\alpha_m|\sigma_m) \propto 1$, $p(\beta_{jm}|\sigma_m) \propto 1$ and $\sigma_m \sim half - t_{df=3}$, where $half - t_{df}$ is a half Student-t prior with 3 degrees of freedom (Bürkner, 2017).

4.3.1 Test Statistics

To examine the fit of the different models we implemented a series of posterior predictive checks (Gelman et al., 2014). For behavior m at hour $j \in \{1, \ldots, 23\}$ among TDP-43 and WT mice, we examined the posterior predictive minimum, 25th percentile, median, 75th percentile, and the maximum. These posterior predictive statistics were compared to the corresponding statistics among Y_{ijkm}. The 25th, 50th and 75th percentiles assess the model's ability to capture the most frequently observed values. The minimum and maximum assess the model's ability to approximate boundary conditions and predict the extreme-valued observations.

4.3.2 Computation

Many Bayesian models do not have an analytically closed form or are based on large number of parameters (Gelman et al., 2014). To make inference with such models, a more elaborate computational method, such as Markov chain Monte-Carlo (MCMC) algorithms, are required. We implemented all of the models using the Stan software (Carpenter et al., 2017). Stan allows researchers to define a wide variety of probability models. Posterior samples from these models are obtained using a No-U-Turn Sampler, which is a MCMC algorithm that is similar to the Hamiltonian Monte Carlo but removes the requirement of choosing the number-of-steps parameter (Hoffman and Gelman, 2014). To ensure convergence of the MCMC algorithm we examined the improved $\hat{R}$ statistic (Vehtari et al., 2021) and autocorrelation plots.

4.3.3 ANOVA Results

We generate 4000 samples with from the posterior distribution based on Equation (1). We discard the first 500 samples and conducted posterior predictive checks with the remaining samples. Appropriate convergence is achieved as indicated by the $\hat{R}$ values remaining at or below 1.1 for each of the parameters and the trace plot of all the parameters displaying convergence to stable distribution (data not shown).

The posterior predictive plots for each behavior indicate a poor model fit such that the solid red and blue lines (representing observed quantities of WT and TDP-43 mice, respectively) are frequently outside or on the border of their correspondingly colored shaded regions (representing predicted quantities) (Appendix 4.11.1). In particular, Drink, Eat, and Hang behaviors do not capture the median time at most circadian hours. Although some behaviors do capture the median time at many hours, such as Groom and Sniff, none of the behaviors capture the maximum observed time for most circadian hours. Each behavior has predictions for minimum below 0 and maximum above 1, both of which are impossible for the observed data. Separate ANOVA models could not model the dependency that exists between behaviors within any given hour. Assumptions of normality, independence, and homogeneous variance of the outcomes may not be completely satisfied even after transformation.

4.4 The Multivariate Normal-Normal Model

The ANOVA model does not account for the relationship between behaviors. At each hour j, the sum of seconds performing behaviors equals 3600 seconds, therefore an increase in time spent performing behavior m must result in a decrease in cumulative time performing the remaining 8 behaviors. A possible model to address this limitation is the multivariate normal distribution. Let $\tilde{\mathbf{Y}}_{ijk} = g(\mathbf{Y}_{ijk})$ be the 9-dimensional vector of transformed $\mathbf{Y}_{ijk}$, for $g(\cdot)$ defined in Section 4.3. Let $\mathbf{X}_{ijk} = \{X_{1ijk}, H_j, T_i, G_i\}$, where $X_{1ijk} = 1$ represents an intercept, and H_j, T_i and G_i as in Section 4.2. The likelihood of the multivariate Normal-Normal model is,

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\tilde{\mathbf{Y}}, \mathbf{X}) = \prod_{i=1}^{20}\prod_{j=0}^{23}\prod_{k=1}^{6}(2\pi)^{-\frac{9}{2}}det(\boldsymbol{\Sigma})^{-\frac{9}{2}}e^{-\frac{1}{2}(\tilde{\mathbf{Y}}_{ijk}-\mathbf{X}_{ijk}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\tilde{\mathbf{Y}}_{ijk}-\mathbf{X}_{ijk}\boldsymbol{\beta})}, \qquad (2)$$

where $\tilde{\mathbf{Y}} = \{\tilde{\mathbf{Y}}_{ijk}\}$, $\mathbf{X} = \{\mathbf{X}_{ijk}\}$, $\boldsymbol{\beta} = \{\beta_{pm}, p = 1, \ldots, P; m = 1, \ldots, 9\}$ represents the covariate parameters and $\boldsymbol{\Sigma}$ represents the covariance matrix.

To complete the Bayesian model we assumed that apriori,

$$\boldsymbol{\beta}_m^T \sim N_P(\mathbf{0}, 5^2\boldsymbol{I})$$
$$\boldsymbol{\Sigma} \sim Inv - Wish_9(\boldsymbol{I})$$

for $m = 1, \ldots, 9$ where $\boldsymbol{\beta}_m^T$ represents the P-dimensional vector of parameter effects for the m^{th} behavior. This model allows the covariates to have separate effects for each behavior m and describe the conditional correlations between behaviors with $\boldsymbol{\Sigma}$. We implement the No-U-Turn Sampler through Stan to obtain posterior samples.

4.4.1 Multivariate Normal-Normal Model Results

Models were run for 4000 simulations with the first 500 being warm-up runs. Convergence was assessed using the $\hat{R}$ values and visual inspection of trace plots. The $\hat{R}$ of the parameters were at or below 1.1. Model fit was assessed by the posterior predictive plots for all five test statistics which indicated a poor model fit due to lack of accurate predictions of the observed behavior times (Appendix 4.11.2). Each plot had low containment of the observed test statistics within each 95% credible interval for all behaviors. Model performance was similar across behaviors such that the minimum and 75^{th} percentile contained their respective observed values more frequently than the remaining three test statistics. However, the minimum had credible intervals partially below zero for each behavior. Similarly, the maximum had credible intervals partially above one for each behavior. These violated the constraints of the data to be between zero and one following the transformation described in Section 4.3. For most behaviors, the observed median was generally outside the credible intervals at the earliest and latest hours of the day. For Hang and Groom, the observed median was outside of the credible intervals during the middle circadian hours, as well. For all behaviors, observed maximums were almost entirely outside of their respective credible intervals.

4.5 The Multivariate Hierarchical Normal-Normal Model

In settings where data is collected longitudinally, there may be correlation between observations recorded by the same subject. The multivariate normal model in Section 4.4 does not account for correlation for repeated measurement for the same mouse. Additionally, there may be correlation between observations within the same day. By extending to the multivariate hierarchical normal model we can account for these correlations. We specified additional effects through the mean of the distribution. Define $\mathbf{X}_{ijk}$ as in Section 4.4. The likelihood function is as follows,

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\delta}|\tilde{\mathbf{Y}}, \mathbf{X}) = \prod_{i=1}^{20}\prod_{j=0}^{23}\prod_{k=1}^{6} (2\pi)^{-\frac{9}{2}} det(\boldsymbol{\Sigma})^{-\frac{9}{2}}$$
$$\times\, exp\bigg(-\frac{1}{2}(\tilde{\mathbf{Y}}_{ijk} - (\mathbf{X}_{ijk}\boldsymbol{\beta} + \lambda_i\mathbf{1} + \delta_k\mathbf{1}))^T \qquad (3)$$
$$\boldsymbol{\Sigma}^{-1}(\tilde{\mathbf{Y}}_{ijk} - (\mathbf{X}_{ijk}\boldsymbol{\beta} + \lambda_i\mathbf{1} + \delta_k\mathbf{1})\bigg)$$

where $\boldsymbol{\lambda} = \{\lambda_i\}$ is a 20-dimensional vector representing the effect of each mouse and $\boldsymbol{\delta} = \{\delta_k\}$ is a 6-dimensional vector representing the effect of each day. We set the prior distributions for the effects of $\boldsymbol{\lambda}$ and $\boldsymbol{\delta}$ as follows,

$$\lambda_i \sim N(0, \tau_\lambda)$$
$$\delta_k \sim N(0, \tau_\delta)$$

for all $i \in \{1, \ldots, 20\}$ and $k \in \{1, \ldots, 6\}$. The distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are defined identically as in Section 4.4. Posterior samples from this model are obtained from the No-U-Turn Sampler through Stan, as in Section 4.3.

Again this model achieved appropriate convergence indicated by each of the parameters $\hat{R}$ values and visual inspection of trace plots. The model fit was assessed through posterior predictive plots for the test statistics described in Section 4.3. Model fit was similarly poor to the multivariate normal-normal model for each test statistic, shown in Appendix 4.11.3. For all behaviors, the observed minimum and 75^{th} percentile values were contained within their respective credible intervals at most hours. Containment of the observed values within credible intervals decreased in frequency from the median, to the 25^{th} percentile, and finally to the maximum, for all behaviors. The minimum and maximum credible intervals were again partially below zero and above one, respectively, violating the constraints of the data.

4.6 The Multivariate Hierarchical Normal-Normal Model with Cyclic Splines

The posterior predictive plots from the hierarchical normal-normal model (Section 4.5) indicate that assuming a linear effect of time may not be appropriate. Therefore, we further extend the multivariate hierarchical normal-normal model to allow for a non-linear effect for the hour of observation. Cyclic splines are an appropriate method to capture a non-linear effect that varies over a repeating time-frame (e.g. seasons, months, years). Cyclic splines require endpoints to have similar values to satisfy the periodical nature of the effect. This is suitable, for example, for the hourly effect such that H_{23} from day k and H_0 from day $k+1$ produce similar responses. Let $\mathbf{X}_{ijk} = \{X_{1ijk}, T_i, G_i\}$, where $X_{1ijk} = 1$ represents an intercept, and T_i and G_i as in Section 4.2. Introducing the spline into the mean of the likelihood function takes the following form,

$$\begin{aligned}
&L(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\theta} | \tilde{\mathbf{Y}}, \mathbf{X}, \mathbf{H}) \\
&\quad = \prod_{i=1}^{20} \prod_{j=0}^{23} \prod_{k=1}^{6} (2\pi)^{-\frac{9}{2}} det(\boldsymbol{\Sigma})^{-\frac{9}{2}} \\
&\quad \times exp\Bigg(-\frac{1}{2} \Big(\tilde{\mathbf{Y}}_{ijk} - \Big(\mathbf{X}_{ijk}\boldsymbol{\beta} + \lambda_i \mathbf{1} + \delta_k \mathbf{1} + \boldsymbol{\psi}(H_j) + \boldsymbol{\alpha} \odot \boldsymbol{\psi}(H_j) T_i \Big) \Big)^T \\
&\qquad \boldsymbol{\Sigma}^{-1} \Big(\tilde{\mathbf{Y}}_{ijk} - \Big(\mathbf{X}_{ijk}\boldsymbol{\beta} + \lambda_i \mathbf{1} + \delta_k \mathbf{1} + \boldsymbol{\psi}(H_j) + \boldsymbol{\alpha} \odot \boldsymbol{\psi}(H_j) T_i \Big) \Big) \Bigg)
\end{aligned} \tag{4}$$

where $\mathbf{H} = \{H_j\}$, $\boldsymbol{\psi}(H_j)$ is a 9-dimensional vector of cyclic splines on the hour of observation, $\boldsymbol{\alpha} = \{\alpha_m\}$ represents the 9-dimensional vector of parameter effects for the interaction between the cyclic splines, $\boldsymbol{\psi}(H_j)$, and genotype, T_i, and $\odot$ represents element-wise multiplication. We can express the cyclic splines as,

$$\psi_m(H_j) = \sum_{l=1}^{L} \theta_{lm} \, C_l(H_j) \tag{5}$$

for all $m \in \{1, \ldots, 9\}$, where $\boldsymbol{\theta}_m$ is an L-dimensional vector of basis coefficients for behavior m, and $C_l(\cdot)$ is the basis function for the cyclic spline. To complete the Bayesian regression model we set prior distributions for $\boldsymbol{\theta} = \{\theta_{lm}\}$ and $\boldsymbol{\alpha}$,

$$\theta_{lm} \sim N\big(0, 1\big)$$
$$\alpha_m \sim N\big(0, 5^2\big)$$

for all $l \in \{1, \ldots, L\}$ and $m \in \{1, \ldots, 9\}$. The remaining parameters, $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\delta}$ are defined identically to the previous model in Section 4.5. The No-U-Turn Sampler through Stan is implemented to obtain posterior samples, as in Section 4.3.

Again the model converged as indicated by the $\hat{R}$ value being at or less than 1.1 for each parameter and trace plots indicating a stable distribution. The posterior predictive plots identified this to be a better fitting model such that the observed values were more frequently contained in the credible intervals compared to previous models for all behaviors, shown in Appendix 4.11.4. The test statistics characterizing the interior of the distribution contain the observed value within their credible intervals at many hours, for several behaviors, such as EBH, Groom, and Sniff. However, across test statistics and behaviors, such as Drink and Hang, many observed values were outside the credible intervals. For all behaviors, results for the minimum test statistic was similar to the previous models. The maximum test statistic improved compared to previous models for each behavior, having some observed values correctly predicted by the model, although most were overestimated. However, both minimum and maximum credible intervals violated the constraints of the data, predicting values below zero and above one, respectively.

4.7 Hierarchical Dirichlet Model with Cyclic Splines

The multivariate normal-normal models fail to account for $\sum_{m=1}^{M} \tilde{Y}_{ijkm} = 1$ for all i, j, and k. This shortcoming leads us to consider extending our methods to an alternative distribution.

The Dirichlet distribution imposes an appropriate constraint on the sum of the outcomes for a multivariate response. This model is suitable following a transformation of our data to be proportional values, such that $\tilde{Y}_{ijkm} = g(Y_{ijkm}) = \frac{\tilde{Y}_{ijkm}}{\sum_{m=1}^{9} \tilde{Y}_{ijkm}}\ \forall\ i, j, k, m$. We let $\mathbf{X}_{ijk}$ be defined as in Section 4.6. For observation $\tilde{\mathbf{Y}}_{ijk}$ the Dirichlet distribution takes the following form,

$$f\left(\tilde{Y}_{ijk1}, \ldots, \tilde{Y}_{ijk9} | \gamma_1, \ldots, \gamma_M\right) = \frac{1}{B(\boldsymbol{\gamma})} \prod_{m=1}^{M} \tilde{Y}_{ijkm}^{(\gamma_m - 1)} \tag{6}$$

where $\boldsymbol{\gamma} = \{\gamma_m\}$, $\gamma_m > 0$ for all m, $\sum_{m=1}^{M} \tilde{Y}_{ijkm} = 1$ for all i, and $B(\boldsymbol{\gamma}) = \frac{\prod_{m=1}^{M} \Gamma(\gamma_m)}{\Gamma\left(\sum_{m=1}^{M} \gamma_m\right)}$ represents the multivariate beta function. Note that the mean response of each outcome is $E(\tilde{Y}_{ijkm}) = \frac{\gamma_m}{\gamma_0}$, where $\gamma_0 = \sum_{m=1}^{M} \gamma_m$, and variance $Var(\tilde{Y}_{ijkm}) = \frac{\gamma_m(\gamma_0 - \gamma_m)}{\gamma_0^2(\gamma_0 + 1)}$ (Douma and Weedon, 2019).

An alternative parameterization of the Dirichlet distribution allows direct specification of the mean responses. We obtain this paramterization by making the transformations $\gamma_m = \mu_{ijkm}\phi_{ijk}$ and $\gamma_0 = \phi_{ijk}$, where $E(\tilde{Y}_{ijkm}) = \mu_{ijkm}$ and $Var(\tilde{Y}_{ijkm}) = \frac{\mu_{ijkm}(1 - \mu_{ijkm})}{\phi_{ijk} + 1}$ (Douma and Weedon, 2019).

Dirichlet regression links the linear predictor, $\boldsymbol{\eta}_{ijk} = \{\eta_{ijkm}\}$, with the mean responses, μ_{ijkm}, through some link function. The multinomial logit function is an appropriate link for this regression, such that

$$\mu_{ijkm} = \frac{e^{\eta_{ijkm}}}{\sum_{m=1}^{M} e^{\eta_{ijkm}}} \tag{7}$$

$$\mu_{ijkM} = \frac{1}{\sum_{m=1}^{M} e^{\eta_{ijkm}}} \tag{8}$$

where the first $M - 1$ outcome means are specified through the function (7) of the linear predictor η_{ijkm} and the M^{th} linear predictor $\eta_{ijkM} = 0$ to impose the constraint on the sums, $\sum_{m=1}^{M} \mu_{ijkm} = 1$. The precision parameter ϕ_{ijk} is modelled through a log link function such that $log(\phi_{ijk}) = \mathbf{X}_{ijk}\boldsymbol{\xi}$, for parameters $\boldsymbol{\xi} = \{\xi_p\},\ p = 1, \ldots, P$.

We specify the linear predictor of the Dirichlet regression to be the mean specification as defined in the hierarchical multivariate normal-normal model with cyclic splines from Section 4.6,

$$\boldsymbol{\eta}_{ijk} = \mathbf{X}_{ijk}\boldsymbol{\beta} + \lambda_i \mathbf{1} + \delta_k \mathbf{1} + \boldsymbol{\psi}\left(H_j\right) + \boldsymbol{\alpha} \odot \boldsymbol{\psi}\left(H_j\right) T_i \tag{9}$$

where each variable is defined identically to Section 4.6 with the same prior distributions. This model accounts for the covariates of interest, correlated observations due to repeated observations within mice and days, a non-linear effect for hour of observation, and an interaction between the spline for hour of observation and genotype.

The test statistics and computation for this model followed the methods described in Section 4.3, respectively, with further guidance for model implementation in Stan by Sennhenn-Reulen (2018). Here we define $\tilde{Y}_{ijkm} = g(Y_{ijkm}) = \frac{Y_{ijkm}+10}{\sum_{m=1}^{9}(Y_{ijkm}+10)} \; \forall \; i, j, k, m$, to satisfy the constraint of the Dirichlet distribution such that $\sum_{m=1}^{9} \tilde{Y}_{ijkm} = 1 \; \forall \; i, j, k$ and enforce non-zero probabilities for each behavior m for all i, j, and k.

The model reached appropriate convergence determined by appropriate $\hat{R}$ values and trace plot inspection (data not shown), and was assessed for fit using the posterior predictive plots across each hour of observation (Appendix 4.11.5). Overall, specific behaviors, such as Eat and Groom, contained the observed value within their credible intervals at most hours, across all test statistics. However, there was varied model success within each test statistic across behaviors. For the interior distribution test statistics, the model failed to accurately predict values that were observed at more extreme ends of the distribution, i.e. very large or small amounts of time performing a behavior. For these interior test statistics, behaviors Drink, Hang, Rear, and Walk had predictions too large during the middle hours of the day. Whereas the Rest behavior had predictions too small during the same timeframe and same test statistics. Across behaviors, minimum values were generally predicted well with some observed values being outside the credible intervals due to the additional 10 seconds added to each observation, although this difference was small. Observed maximum values were poorly predicted but still improved compared to previous models. The observed Drink and Walk maximums for each hour were not contained in all but four combined hours of their credible intervals. However, neither the minimum or maximum credible intervals predicted values outside of the constraints requiring observations to be between zero and one, inclusive.

4.7.1 Skewed Link Function

The logistic function is a symmetric link function which may not be appropriate for all data. The hierarchical Dirichlet model with cyclic splines under-performs when predicting values close to zero which may be indicative of an inappropriate link function. Skewed, or asymmetrical, link functions may be a suitable method to improve this performance. Caron et al. (2018) have studied the performance of and compared several symmetric and asymmetric link functions. Each function was considered in the binary and multinomial settings, with the Weibull function performing the best across evaluation metrics for multinomial data. However, the Weibull link function requires a non-negative linear predictor which we cannot guarantee. An appropriate, asymmetric alternative that does not restrict the linear predictor values is the log-log link, $g^{-1}(\eta_{ijkm}) = exp(-exp(\eta_{ijkm}))$. We can use the log-log link for the Dirichlet regression such that the mean responses are defined as,

$$\mu_{ijkm} = \frac{e^{-e^{\eta_{ijkm}}}}{\sum_{m=1}^{M} e^{-e^{\eta_{ijkm}}}} \tag{10}$$

$$\mu_{ijkM} = \frac{e^{-1}}{\sum_{m=1}^{M} e^{-e^{\eta_{ijkm}}}} \tag{11}$$

where we again set the M^{th} linear predictor equal to zero to constrain the sum of the means to equal one. The model was run identically in construction, computation, and analysis as the hierarchical Dirichlet model with cyclic splines.

Overall, model performance for predicting the observed values was similar to the previous hierarchical Dirichlet model (Appendix 4.11.6). Observed Eat and Groom behaviors were predicted well across most test statistics whereas Hang and Rear, for example, had varying success. Notably, the credible intervals for Rest and Walk contained the observed values much less frequently with this asymmetric link function compared to the symmetric link.

From hours 7 to 17 the model predicted values too low for Rest at each test statistic except the minimum. The same held true for Walk with predicted values too large during the same timeframe and test statistics.

4.8 Zero-Inflated Hierarchical Generalized Dirichlet Multinomial (ZIHGDM) Model with Cyclic Splines

To account for the high frequency of zero responses we consider the zero-inflated generalized Dirichlet multinomial regression model (ZIGDM) (Tang and Chen, 2018). Let $\mathbf{X}_{ijk}$ be defined as in Section 4.6. We begin by summarizing the ZIGDM model from Tang and Chen (2018) in the context of our data. Let $\boldsymbol{\Delta_{ijk}} = \{\Delta_{ijkm}\}$ represent the vector of binary (zero/non-zero) outcomes for the first 8 behaviors. We assume

$$\Delta_{ijkm} \sim Bernoulli\left(\pi_{ijkm}\right) \tag{12}$$

where π_{ijkm} represents the probability that mouse ijk will have a zero response for behavior $m \in \{1, \ldots, 8\}$. Similar to Section 4.7, we do not estimate any parameter effects for the last outcome to satisfy the constraint of the Dirichlet distribution probabilities summing to one.

We postulate that $\mathbf{Y}_{ijk}$ follows a multinomial distribution with probabilities given by $\mathbf{P}_{ijk} = \{P_{ijk1}, \ldots, P_{ijk9}\}$. We construct a prior on $\mathbf{P}_{ijk}$ using the Generalized Dirichlet (GD) distribution (Tang and Chen, 2018). Specifically, let $\mathbf{Z}_{ijk} = \{Z_{ijk1}, \ldots, Z_{ijk8}\}$ represent a set of mutually independent random variables, such that each Z_{ijkm} follows a Beta distribution. The zero-inflated GD multinomial model (ZIGDM) is,

$$Z_{ijkm} = 0 \text{ if } \Delta_{ijkm} = 1,\ Z_{ijkm} \mid \Delta_{ijkm} = 0 \sim Beta\big(a_{ijkm}, b_{ijkm}\big),\ \ m = 1, \ldots, 8 \tag{13}$$

$$P_{ijk1} = Z_{ijk1}, \quad P_{ijkm} = Z_{ijkm} \prod_{h=1}^{m-1} (1 - Z_{ijkh}),\ \ m = 2, \ldots, 8 \tag{14}$$

$$\mathbf{Y}_{ijk} \sim Multinomial(\mathbf{P}_{ijk}, N) \tag{15}$$

where the behavioral outcomes, $\mathbf{Y}$, are integers and satisfy $\sum_{m=1}^{9} Y_{ijkm} = 3600$, and $\sum_{m=1}^{9} P_{ijkm} = 1$ for all i, j, and k. An alternative parameterization of the Beta distribution defines

$$\mu_{ijkm} = \frac{a_{ijkm}}{a_{ijkm} + b_{ijkm}} \tag{16}$$

$$\sigma_{ijkm} = \frac{1}{1 + a_{ijkm} + b_{ijkm}} \tag{17}$$

where μ_{ijkm} and σ_{ijkm} represent the mean and dispersion of the Beta distribution. We can then link μ_{ijkm}, σ_{ijkm}, and π_{ijkm} to a matrix of covariates using the logit link function. We specify μ_{ijkm} and π_{ijkm} similar to the mean specification for the hierarchical Dirichlet model such that

$$log\left(\frac{\mu_{ijkm}}{1-\mu_{ijkm}}\right) = \mathbf{X}_{ijk}\boldsymbol{\beta}_{1m} + \lambda_{1i} + \delta_{1k} + \psi_{1m}\left(H_j\right) + \alpha_{1m}\ \psi_{1m}\left(H_j\right) T_i \tag{18}$$

$$log\left(\frac{\pi_{ijkm}}{1-\pi_{ijkm}}\right) = \mathbf{X}_{ijk}\boldsymbol{\beta}_{2m} + \lambda_{2i} + \delta_{2k} + \psi_{2m}\left(H_j\right) + \alpha_{2m}\ \psi_{2m}\left(H_j\right) T_i \tag{19}$$

for all $m \in \{1, \ldots, 8\}$ and where each of the parameters on the right-hand-side of equations (18) and (19) are defined similarly to Section 4.7. We specify the dispersion parameter, σ_{ijkm}, such that,

$$log\left(\frac{\sigma_{ijkm}}{1-\sigma_{ijkm}}\right) = \mathbf{X}_{ijk}\boldsymbol{\phi}_m \tag{20}$$

where $\boldsymbol{\phi} = \{\phi_{pm}, p = 1, \ldots, P; m = 1, \ldots, 8\}$ represents a parameter matrix. It was not required for the covariate matrix, $\mathbf{X}_{ijk}$, to be specified identically for equations (18), (19), and (20), but was chosen as such due to the a small number of covariates.

Let Ω represent all of the parameters in the model. We define the log likelihood to be,

$$
\begin{aligned}
&log(L(\mathbf{\Omega} \mid \mathbf{Y},\mathbf{X})) \\
&\qquad = \sum_i \sum_j \sum_k log\{f(\mathbf{Y}_{ijk} \mid \mathbf{P}_{ijk})\} \\
&\qquad + \sum_i \sum_j \sum_k \sum_m \Big\{ \Delta_{ijkm} log(\pi_{ijkm}) + (1-\Delta_{ijkm}) log(1-\pi_{ijkm}) \\
&\qquad + (1-\Delta_{ijkm}) log\{f(Z_{ijkm} \mid a_{ijkm}, b_{ijkm})\} \Big\}
\end{aligned}
\tag{21}
$$

where $f(Z_{ijkm} \mid a_{ijkm}, b_{ijkm})$ represents the Beta distribution with parameters, $a_{ijkm} = \mu_{ijkm}(1/\sigma_{ijkm} - 1)$ and $b_{ijkm} = (1-\mu_{ijkm})(1/\sigma_{ijkm} - 1)$ (Tang and Chen, 2018).

The test statistics and computation of this model are similar to those described in Section 4.3 with two exceptions. First, the ZIHGDM model required greater than 4000 simulations to provide stable estimates. Thus, the model was run for 12,000 iterations, with 2000 warm-up iterations, and results were analyzed on a subsample of 2500 of these simulation runs. Second, the data was transformed to ensure observations were integer values and still satisfy $\sum_{m=1}^{9} Y_{ijkm} = 3600$ for all i, j, k.

The model reached appropriate convergence following 10,000 iterations as identified by $\hat{R}$, trace plot inspection, and the additional confirmation of sufficient effective sample size for each of the parameters. For every behavior, and every test statistic, and at nearly every hour of observation, the observed values were contained within their respective credible intervals (Appendix 4.11.7). The model was able to accurately predict values across the distribution of observed values. No predictions violated the constraints of the data, requiring values to be between 0 and 3600. Unlike all previous models, for individual behaviors there were many hours that had credible intervals which did not overlap between different genotypes. This was the first model to produce non-overlapping credible intervals for any behavior between the WT and TDP-43 mice. The ZIHGDM model was able to appropriately capture the relationships within the data as evidenced by the posterior predictive plots. Small, moderate, and large time spent performing a behavior was accurately predicted for all behaviors. These results lead us to trust this model to predict the observed data and to subsequently predict future outcomes.

4.9 Model Comparisons

4.9.1 Estimands and Evaluation Metrics

We are interested in mouse behavior time for both TDP-43 and WT genotypes, and their differences in behavior time. We define the following estimands: $\boldsymbol{\tau}_{1j} = E[\mathbf{Y}_{ijk} \mid T_i = 1, H_j = j]$ represents the mean behavior time for TDP-43 mice at hour j, $\boldsymbol{\tau}_{0j} = E[\mathbf{Y}_{ijk} \mid T_i = 0, H_j = j]$ represents the mean behavior time for WT mice at hour j, and $\boldsymbol{\tau}_{10j} = E[\mathbf{Y}_{ijk} \mid T_i = 1, H_j = j] - E[\mathbf{Y}_{ijk} \mid T_i = 0, H_j = j]$ represents the mean difference in behavior time between TDP-43 and WT mice at hour j, for all j.

We obtain point estimates, $\widehat{\boldsymbol{\tau}}_{1j}$, and 95% central credible intervals for $\boldsymbol{\tau}_{1j}$ directly from the posterior distribution of the estimand of interest $p(\boldsymbol{\tau}_{1j} \mid \mathbf{Y}, T = 1, H = j)$ for all $j \in \{0, \ldots, 23\}$. We obtain this posterior distribution for hour j as follows,

$$
\begin{aligned}
p(\boldsymbol{\tau}_{1j} \mid &\mathbf{Y}, T = 1, H = j) \\
&= \int p(\boldsymbol{\tau}_{1j} \mid \mathbf{Y}, \mathbf{X}, \boldsymbol{\Omega}, T = 1, H = j) \\
&\qquad \cdot p(\boldsymbol{\Omega} \mid \mathbf{Y}, \mathbf{X}, T = 1, H = j) \cdot p(\mathbf{X}) \, d\boldsymbol{\Omega} \, d\mathbf{X}
\end{aligned}
$$

where $\boldsymbol{\Omega}$ represents each parameter in the model. We repeat this procedure similarly for $\boldsymbol{\tau}_{0j}$ and $\boldsymbol{\tau}_{10j}$ from their respective posterior distributions.

4.9.2 Results

We perform the procedure outlined in Section 4.9.1 for the ZIHGDM, hierarchical Dirichlet (HDIR), and hierarchical multivariate normal-normal models (HMVNN), each with cyclic splines. To compare estimates of the estimands of interest between models, we perform the inverse transformation defined in Section 4.3 to the HMVNN posterior distribution to return to the original outcome range.

Table 4.1 displays the posterior mean behavior times and 95% credible interval for TDP-43 and WT mice at hour 12 specifically, where responses are recorded in seconds per hour. For TDP-43 mice, the ZIHGDM model estimates less time drinking, eating, eating by hand, rearing, and resting than the HDIR and HMVNN models. Each of the corresponding credible intervals for these behaviors were most precise for the HMVNN model, with intervals generally less than 20 seconds in width. Similar trends were observed between models for the WT mice.

At hour 12, the HDIR model estimated sniffing time for mice of both genotypes to be approximately 40-50% of the estimates from the ZIHGDM and HMVNN models. For all models, interval estimates for sniffing were generally wider than all other behaviors with the exception of resting. The HMVNN model estimated the average time spent resting was approximately 50% greater than estimated by the ZIHGDM and HDIR models, and more than twice as large for WT mice compared to the ZIHGDM model.

The mean time spent hanging was estimated to be approximately three times larger for the HDIR model compared to the ZIHGDM and HMVNN models. For this behavior, the HVMNN model had the smallest credible intervals (interval width = 2 seconds) and HDIR the largest intervals for this behavior (interval width = 30 seconds) which was observed for both genotypes.

The ZIHGDM and HDIR models satisfy the constraints that sum of estimated mean times equals 3600 for each hour. However, the HVMNN violate this constraint such that the sum of behavior estimates at each hour is approximately 50% greater than 3600.

Across the 24 hours, results of the three models are generally similar to the patterns described for hour 12 (Appendix 4.11.8), with the exception of resting and sniffing. Resting estimates were larger for ZIHGDM model compared to HDIR estimates for TDP-43 mice during early and late circadian hours. Additionally, for some early and late circadian hours, HDIR sniffing estimates are larger than ZIHGDM model estimates. In most behaviors, the ZIHGDM model reported smaller mean estimates for both genotypes. The HMVNN model violates the data constraints, generally overestimating the time spent performing many behaviors, and has credible intervals that are exceptionally small. The HDIR model has difficulty estimating smaller behavior times evidenced by larger mean estimates and larger lower credible interval bounds compared to the ZIHGDM model.

Table 4.1: Mean time in seconds per hour performing each behavior at hour 12 for TDP-43 and WT mice from the ZIHGDM, hierarchical dirichlet (HDIR), and hierarchical multivariate normal-normal (HMVNN) models, each with cyclic splines. 95% credible intervals are reported in parenthesis.

		ZIHGDM		HDIR		HMVNN	
Hour	Behavior	TDP-43	WT	TDP-43	WT	TDP-43	WT
12	Drink	15 (11,19)	25 (20,31)	104 (92,120)	117 (98,134)	38 (37,39)	39 (38,40)
12	Eat	148 (120,178)	101 (80,125)	169 (150,190)	136 (119,155)	505 (493,517)	472 (459,486)
12	EBH	117 (100,136)	93 (78,110)	186 (167,207)	162 (138,188)	322 (317,328)	306 (299,312)
12	Groom	493 (435,555)	464 (421,511)	486 (444,530)	455 (410,506)	744 (738,749)	740 (736,745)
12	Hang	26 (17,38)	37 (24,53)	105 (93,117)	106 (91,123)	25 (24,26)	24 (23,25)
12	Rear	116 (95,138)	150 (127,175)	149 (132,165)	171 (147,196)	381 (374,387)	399 (392,406)
12	Rest	1769 (1644,1890)	1216 (1161,1274)	1838 (1772,1904)	1848 (1781,1910)	2590 (2536,2642)	2603 (2558,2652)
12	Sniff	735 (609,863)	1288 (1210,1367)	454 (406,503)	474 (436,513)	977 (970,984)	982 (976,988)
12	Walk	182 (141,226)	225 (188,269)	109 (97,122)	132 (117,147)	86 (84,87)	90 (88,92)

Table 4.2 presents the posterior mean difference in behavior times and 95% credible intervals between TDP-43 and WT mice at hour 12. Credible intervals that do not contain zero-valued responses can be considered significant mean differences between the genotypes. A positive mean difference estimate indicates that TDP-43 mice spend more time performing the behavior than WT mice, and less time when estimates are negative.

The ZIHGDM model results from Table 4.2 indicate a significant difference between genotypes in the time spent drinking, eating, eating by hand, rearing, resting, sniffing, and walking at hour 12. According to ZIHGDM, TDP-43 mice spend significantly more time eating, eating by hand, and resting while they spend significantly less time drinking, rearing, sniffing, and walking. The HDIR and HMVNN models result in significant mean differences for eating, rearing, and walking, with overlapping credible intervals for these three behaviors with ZIHGDM. The HMVNN model also identifies significant mean differences for eating by hand, again with an overlapping credible interval to the ZIHGDM model. The HMVNN model reports the smallest credible intervals for each of these significant mean differences.

Across all hours of the day, the HMVNN model generally results in posterior mean behavior time differences less than 10 seconds in magnitude, with the exception of eating, rearing, and resting (Appendix 4.11.8). The HDIR model results in smaller absolute valued posterior mean differences at nearly every hour for all behaviors compared to the ZIHGDM model. The ZIHGDM model reports significant mean differences for each behavior at some hours across the day. The largest mean differences are shown for resting and sniffing whose differences surpass 600 seconds in magnitude at their peaks.

Table 4.2: Mean difference in time (seconds/hour) performing each behavior at hour 12 between TDP-43 and WT mice from the ZIHGDM, hierarchical dirichlet (HDIR), and hierarchical multivariate normal-normal (HMVNN) models, each with cyclic splines. 95% credible intervals are reported in parenthesis.

Hour	Behavior	ZIHGDM	HDIR	HMVNN
12	Drink	-11 (-16,-5)	-13 (-28,5)	-1 (-2,0)
12	Eat	47 (22,73)	33 (16,51)	33 (20,45)
12	EBH	24 (6,42)	24 (-4,56)	17 (11,23)
12	Groom	29 (-25,79)	31 (-12,67)	3 (-2,8)
12	Hang	-11 (-23,1)	0 (-14,13)	1 (0,2)
12	Rear	-34 (-57,-13)	-22 (-45,-1)	-18 (-25,-11)
12	Rest	552 (413,688)	-10 (-73,64)	-12 (-62,32)
12	Sniff	-553 (-699,-410)	-20 (-64,25)	-5 (-11,2)
12	Walk	-43 (-87,-4)	-23 (-34,-11)	-4 (-6,-2)

4.10 Discussion

We developed a model to appropriately and accurately model the time mice spend performing a set of 9 behaviors. The zero-inflated hierarchical generalized Dirichlet multinomial regression model with cyclic splines captures the relationships and characteristics of the data that previous models were unable to do. The fully Bayesian ZIHGDM model accounts for the covariate effects of genotype and gender, correlated observations within the same mouse and day, and a non-linear effect of time. Importantly, this model adheres to the constraints of the data such that hourly sums of time must be constant and no single behavior time can exceed the hourly sum. The ZIHGDM model takes into account the large proportion of zero-valued observations within our data, which the previous models did not consider.

The ZIHGDM model performed better than all previous models through the posterior predictive checks. The posterior predictive plots indicated that the ZIHGDM model was able to accurately predict observed values at nearly every hour of the day, for every behavior, across all 5 test statistics, for both TDP-43 and WT mice. None of the earlier models were able to accurately predict observed values across hours for a single test statistic for most behaviors. In addition, the ZIHGDM model was able to predict observed values with noticeably improved precision, such that credible intervals between the genotypes at the same hour did not overlap for many hours, across behaviors and test statistics. Again, this was not observed for any previous model.

Subsequently, we used the ZIHGDM model determine posterior mean behavior times using the fitted parameters. We observed significant differences between the genotypes across all behaviors at multiple hours. These differences ranged from several seconds to 600 seconds in magnitude. Similar results were not observed for the posterior mean differences of the hierarchical Dirichlet (HDIR) and hierarchical multivariate normal-normal (HMVNN) models both with cyclic splines. Although significant mean differences were still reported for the HDIR and HMVNN models, the magnitude of the differences were generally smaller, with too small of variance under the HMVNN and worse precision under the HDIR, both compared to the ZIHGDM model.

The posterior mean behavior time differences produced by ZIHGDM identify important phenotypes of TDP-43 mice compared to WT mice. In particular, TDP-43 have significant differences in sleep behavior with WT mice, such that TDP-43 mice are more active throughout the dark hours and rest more during light hours of the 24 hour cycle. Similarly, TDP-43 mice are sniffing more than WT mice during dark hours and much less during light hours. All other mean differences in behavior are only positive or negative across the 24 hours. Thus through resting and sniffing, we identify important differences in sleep behavior between these two genotypes.

The ZIHGDM model remains flexible for future application to alternative datasets. Additional covariates are able to be added to the model as needed, as well as additional hierarchical effects. The tradeoff may be the need to run the model for more iterations than the 12,000 in Section 4.8. Furthermore, the ZIHGDM model is not limited to a single non-linear, cyclic effect, as subsequent non-linear effects of different kinds may be added. However, the application of this model is best suited for data under a constant constraint on the sum of the multivariate outcomes, which may hinder its ability for widespread use. However, in these settings, the ZIHGDM model has shown promising ability to achieve desirable results.

Chapter 4 References

J. A. Alfieri, N. S. Pino, and L. M. Igaz. Reversible behavioral phenotypes in a conditional mouse model of tdp-43 proteinopathies. *Journal of Neuroscience*, 34(46):15244–15259, 2014. doi: 10.1523/jneurosci.1918-14.2014.

P.-C. Bürkner. Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1), Aug 2017. doi: 10.18637/jss.v080.i01.

R. Caron, D. Sinha, D. Dey, and A. Polpo. Categorical data analysis using a skewed weibull regression model. *Entropy*, 20(3):176, 2018. doi: 10.3390/e20030176.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

J. C. Douma and J. T. Weedon. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and dirichlet regression. *Methods in Ecology and Evolution*, 10(9):1412–1430, 2019. doi: 10.1111/2041-210x.13234.

A. Doyle, M. P. McGarry, N. A. Lee, and J. J. Lee. The construction of transgenic and gene knockout/knockin mouse models of human disease. *Transgenic Research*, 21(2):327–349, Jul 2011. doi: 10.1007/s11248-011-9537-3.

E. M. Fisher, L. Greensmith, A. Malaspina, P. Fratta, M. G. Hanna, G. Schiavo, A. M. Isaacs, R. W. Orrell, T. J. Cunningham, and A. A. Arozena. Opinion: More mouse models and more translation needed for als. *Molecular Neurodegeneration*, 18(1), May 2023. doi: 10.1186/s13024-023-00619-2.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, 3 edition, 2014.

C. J. Heath, T. J. Bussey, and L. M. Saksida. Motivational assessment of mice using the touchscreen operant testing system: Effects of dopaminergic drugs. *Psychopharmacology*, 232(21-22):4043–4057, 2015. doi: 10.1007/s00213-015-4009-8.

M. D. Hoffman and A. Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1351–1381, Apr 2014.

M. M. Hutmacher, J. L. French, S. Krishnaswami, and S. Menon. Estimating transformations for repeated measures modeling of continuous bounded outcome data. *Statistics in Medicine*, 30(9):935–949, Jan 2011. doi: 10.1002/sim.4155.

R. Janicki. Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates. *Communications in Statistics - Theory and Methods*, 49(9):2264–2284, Feb 2019. doi: 10.1080/03610926.2019.1570266.

H. Jhuang, E. Garrote, X. Yu, V. Khilnani, T. Poggio, A. D. Steele, and T. Serre. Automated home-cage behavioural phenotyping of mice. *Nature Communications*, 1(1), 2010. doi: 10.1038/ncomms1064.

A. C. Mar, A. E. Horner, S. R. Nilsson, J. Alsiö, B. A. Kent, C. H. Kim, A. Holmes, L. M. Saksida, and T. J. Bussey. The touchscreen operant platform for assessing executive function in rats and mice. *Nature Protocols*, 8(10):1985–2005, 2013. doi: 10.1038/nprot.2013.123.

R. Mejzini, L. L. Flynn, I. L. Pitout, S. Fletcher, S. D. Wilton, and P. A. Akkari. Als genetics, mechanisms, and therapeutics: Where are we now? *Frontiers in Neuroscience*, 13, Dec 2019. doi: 10.3389/fnins.2019.01310.

J. F. Monohan. *Primer on Linear Models*. CRC Press, 2008.

A. I. Mundo, J. R. Tipton, and T. J. Muldoon. Generalized additive models to analyze nonlinear trends in biomedical longitudinal data using r: Beyond repeated measures anova and linear mixed models. *Statistics in Medicine*, 41(21):4266–4283, Jul 2022. doi: 10.1002/sim.9505.

K. M. Nilles and B. London. Knockin animal models of inherited arrhythmogenic diseases: What have we learned from them? *Journal of Cardiovascular Electrophysiology*, 18(10): 1117–1125, Jun 2007. doi: 10.1111/j.1540-8167.2007.00884.x.

R. L. Perlman. Mouse models of human disease: An evolutionary perspective. *Evolution, Medicine, and Public Health*, Apr 2016. doi: 10.1093/emph/eow014.

J. A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, 1995.

R. Rusina, R. Vandenberghe, and R. Bruffaerts. Cognitive and behavioral manifestations in als: Beyond motor system involvement. *Diagnostics*, 11(4):624, Mar 2021. doi: 10.3390/diagnostics11040624.

H. Sennhenn-Reulen. Bayesian regression for a dirichlet distributed response using stan, Aug 2018. URL `https://arxiv.org/pdf/1808.06399.pdf`.

Z.-Z. Tang and G. Chen. Zero-inflated generalized dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20(4):698–713, 2018. doi: 10.1093/biostatistics/kxy025.

A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. Rank-normalization, folding, and localization: An improved rˆ for assessing convergence of mcmc (with discussion). *Bayesian Analysis*, 16(2), 2021. doi: 10.1214/20-ba1221.

M. A. White, E. Kim, A. Duffy, R. Adalbert, B. U. Phillips, O. M. Peters, J. Stephenson, S. Yang, F. Massenzio, Z. Lin, and et al. Tdp-43 gains function due to perturbed autoregulation in a tardbp knock-in mouse model of als-ftd. *Nature Neuroscience*, 21(4):552–563, 2018. doi: 10.1038/s41593-018-0113-5.

S. Zarei, K. Carr, L. Reiley, K. Diaz, O. Guerra, P. Altamirano, W. Pagani, D. Lodin, G. Orozco, and A. Chinea. A comprehensive review of amyotrophic lateral sclerosis. *Surgical Neurology International*, 6(1):171, 2015. doi: 10.4103/2152-7806.169561.

4.11 CHAPTER 4 Appendix

4.11.1 ANOVA Posterior Predictive Plots

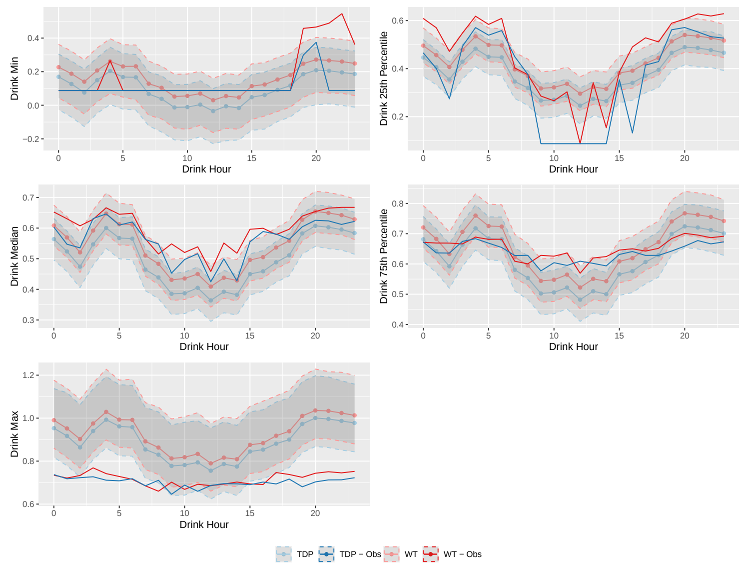

Figure 4.2: ANOVA Drink behavior posterior predictive plots.

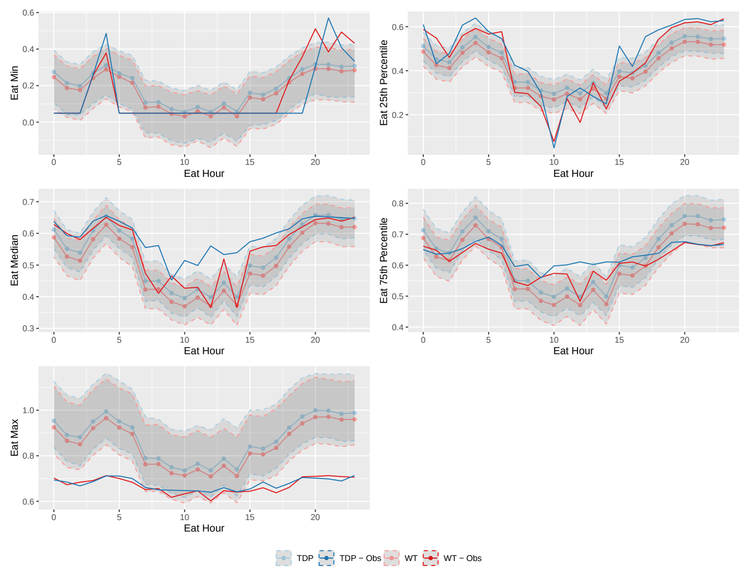

Figure 4.3: ANOVA Eat behavior posterior predictive plots.

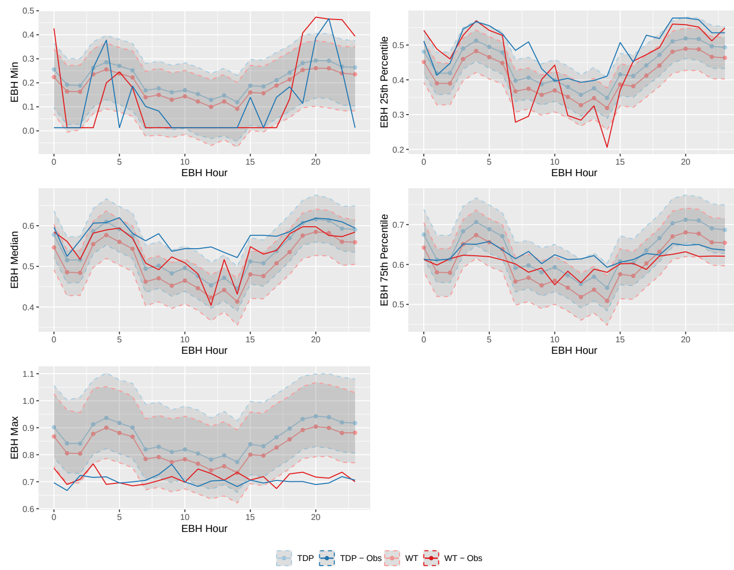

Figure 4.4: ANOVA EBH behavior posterior predictive plots.

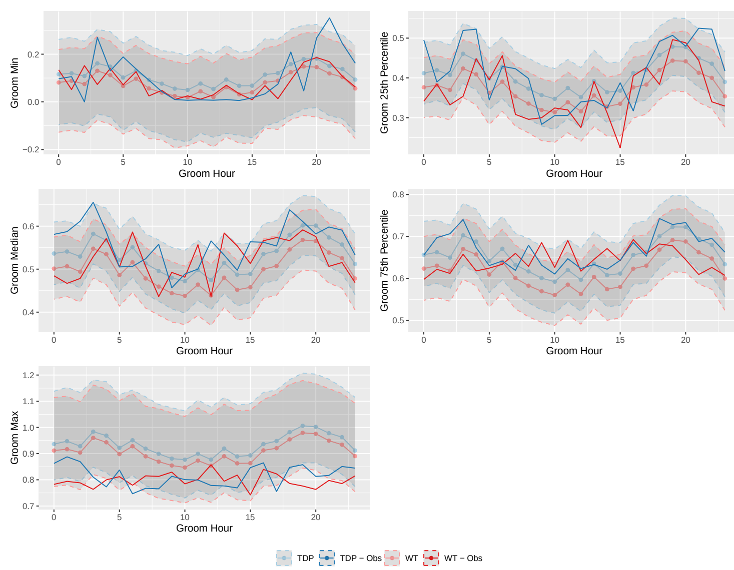

Figure 4.5: ANOVA Groom behavior posterior predictive plots.

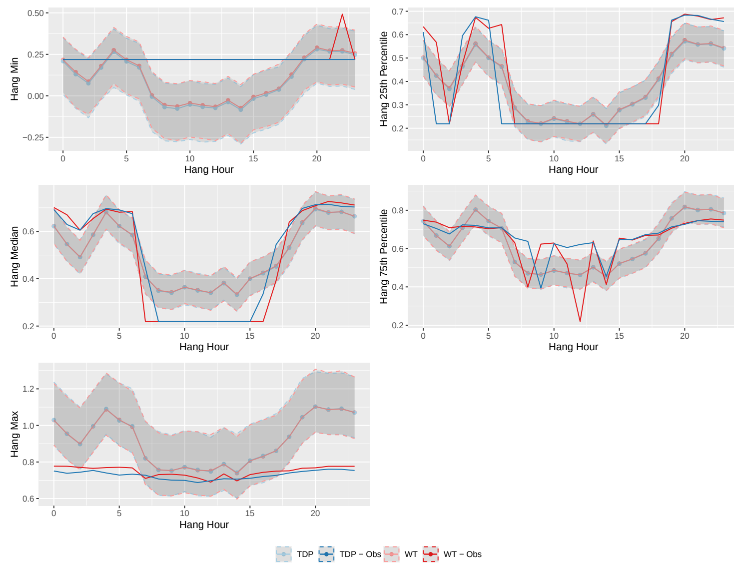

Figure 4.6: ANOVA Hang behavior posterior predictive plots.

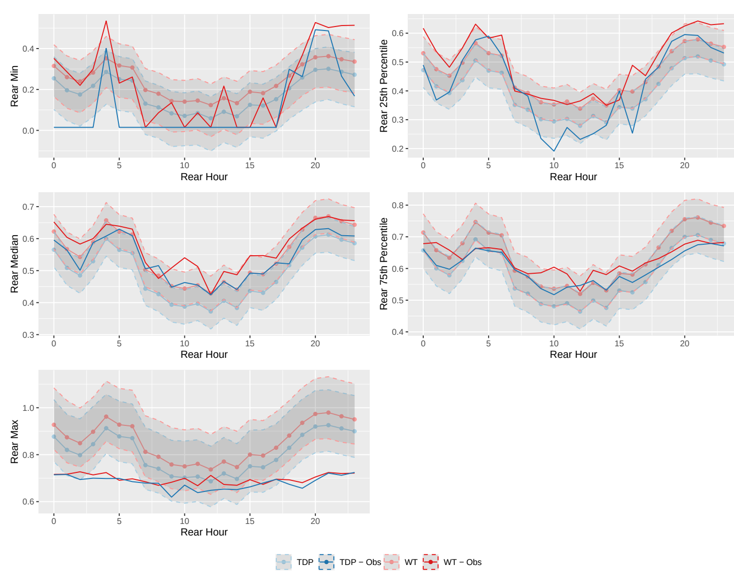

Figure 4.7: ANOVA Rear behavior posterior predictive plots.

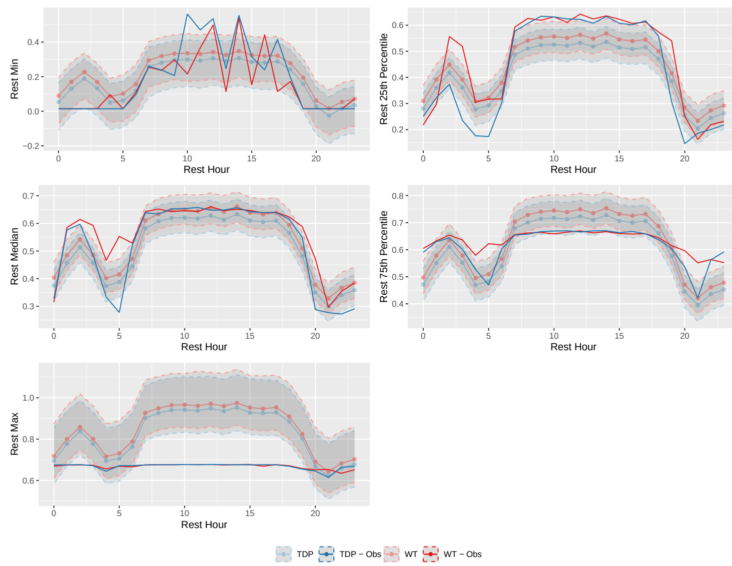

Figure 4.8: ANOVA Rest behavior posterior predictive plots.

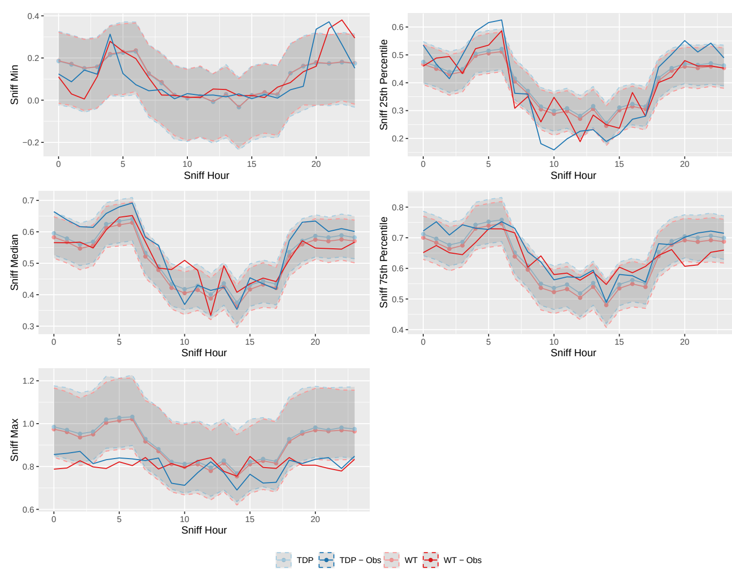

Figure 4.9: ANOVA Sniff behavior posterior predictive plots.

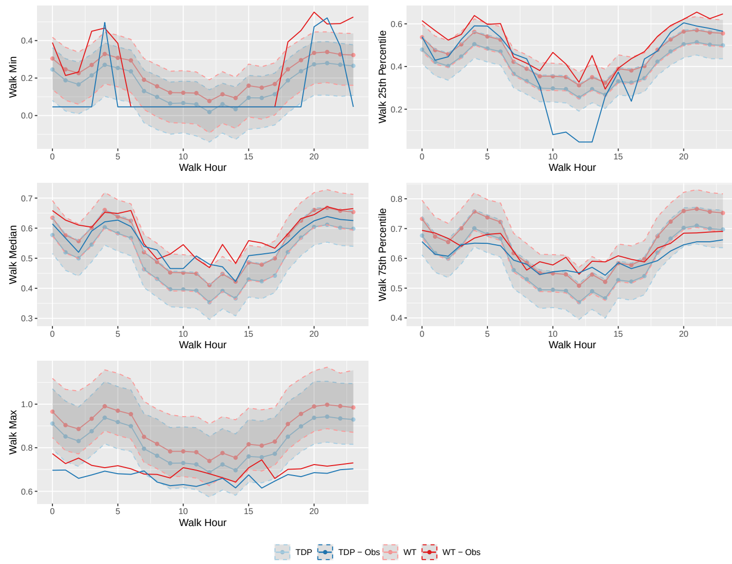

Figure 4.10: ANOVA Walk behavior posterior predictive plots.

4.11.2 Multivariate Normal-Normal Posterior Predictive Plots

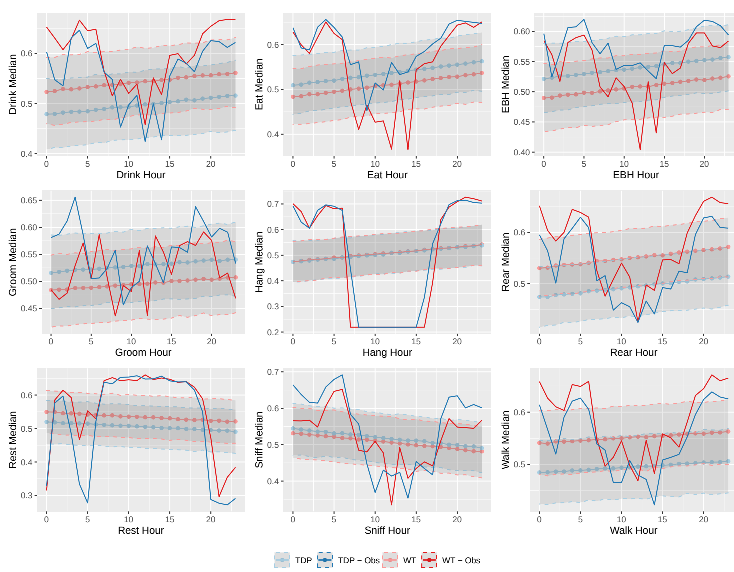

Figure 4.11: Median test statistic posterior predictive plot for the multivariate normal-normal model.

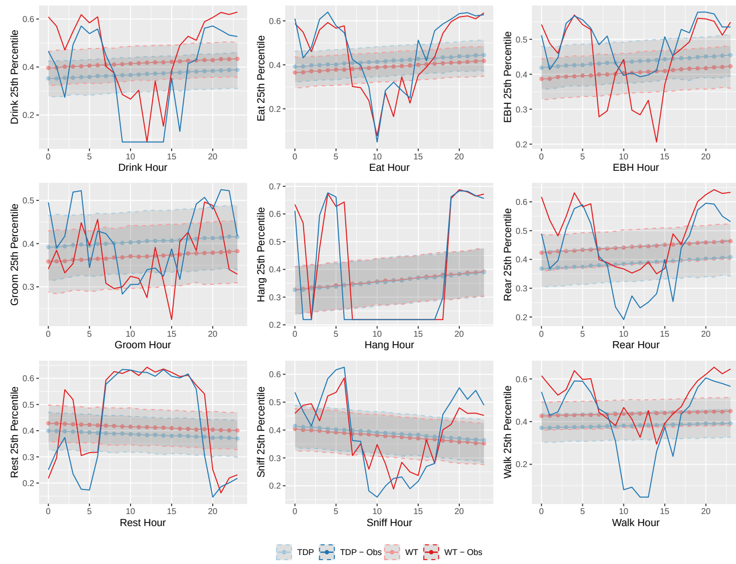

Figure 4.12: 25^{th}-percentile test statistic posterior predictive plot for the multivariate normal-normal model.

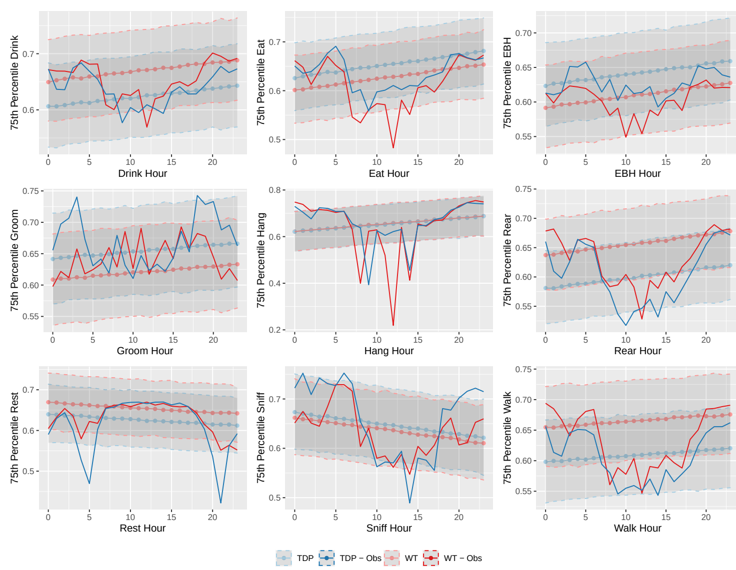

Figure 4.13: 75^{th}-percentile test statistic posterior predictive plot for the multivariate normal-normal model.

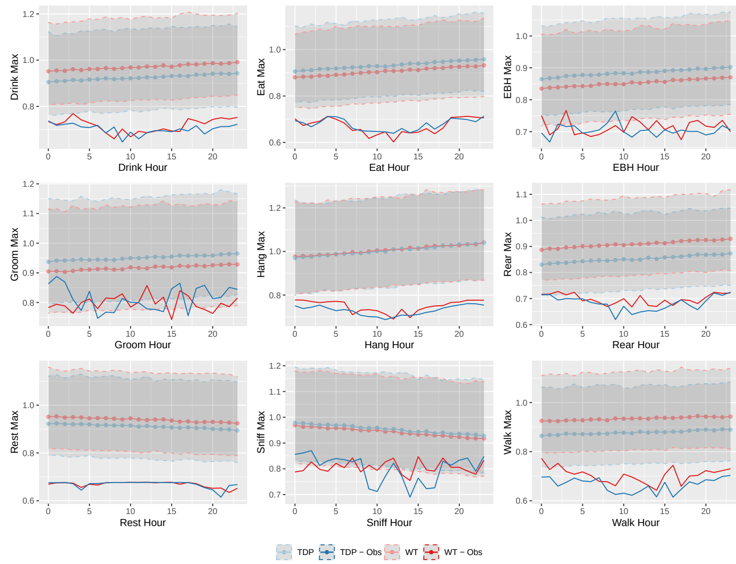

Figure 4.14: Maximum test statistic posterior predictive plot for the multivariate normal-normal model.

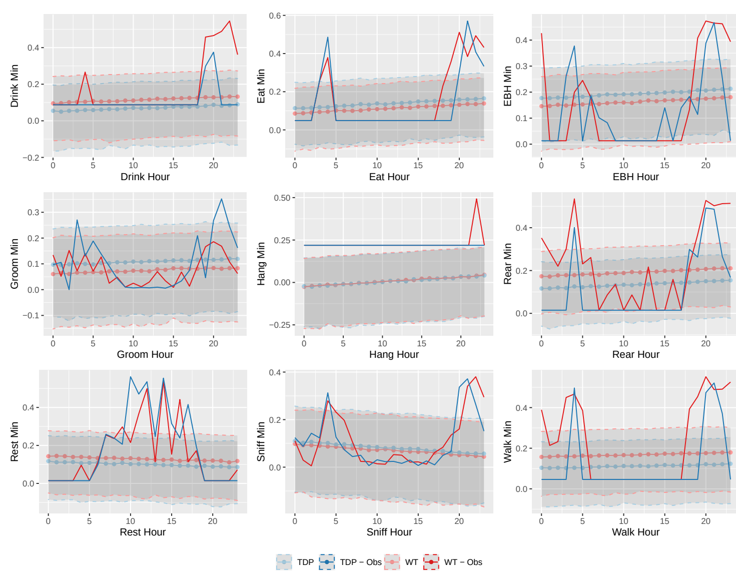

Figure 4.15: Minimum test statistic posterior predictive plot for the multivariate normal-normal model.

4.11.3 Hierarchical Multivariate Normal-Normal Posterior Predictive Plots

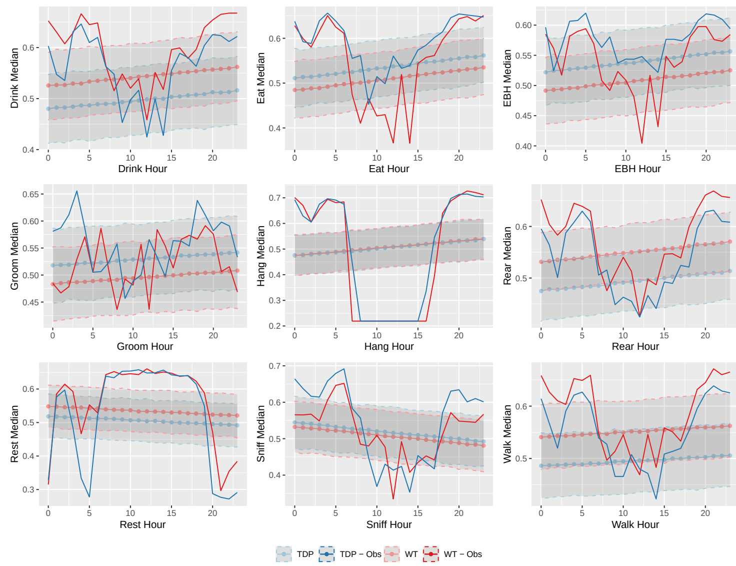

Figure 4.16: Median test statistic posterior predictive plot for the hierarchical multivariate normal-normal model.

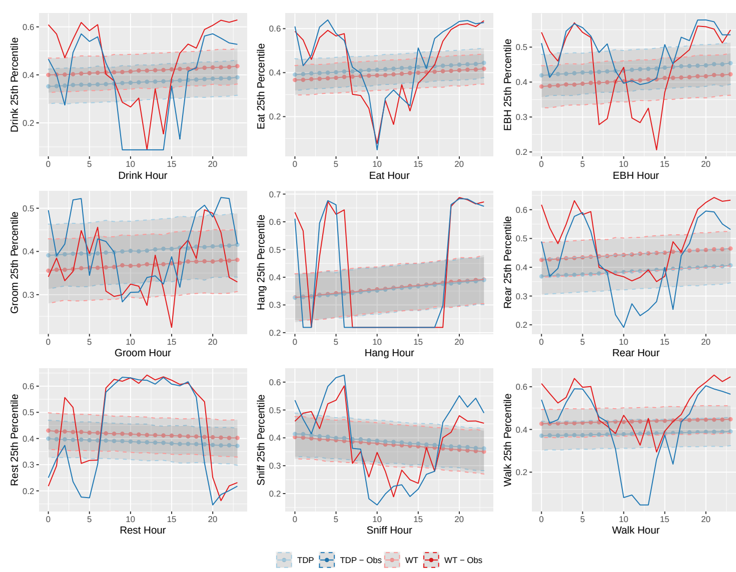

Figure 4.17: 25^{th}-percentile test statistic posterior predictive plot for the hierarchical multivariate normal-normal model.

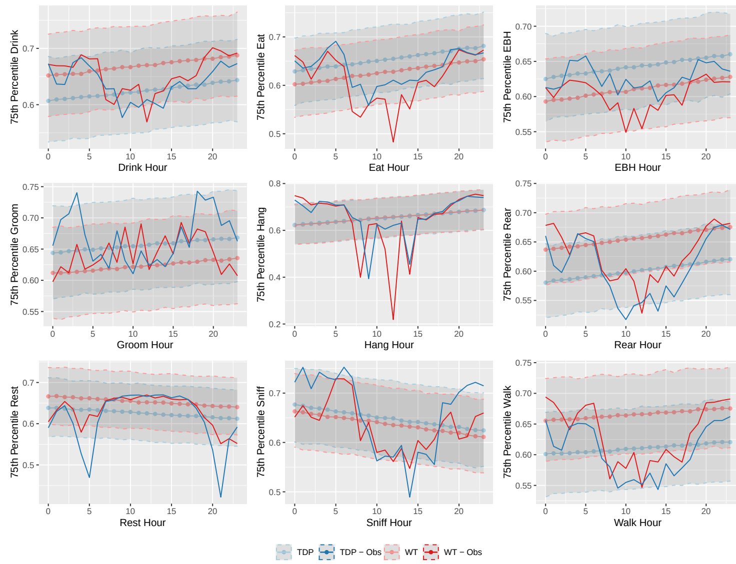

Figure 4.18: 75^{th}-percentile test statistic posterior predictive plot for the hierarchical multivariate normal-normal model.

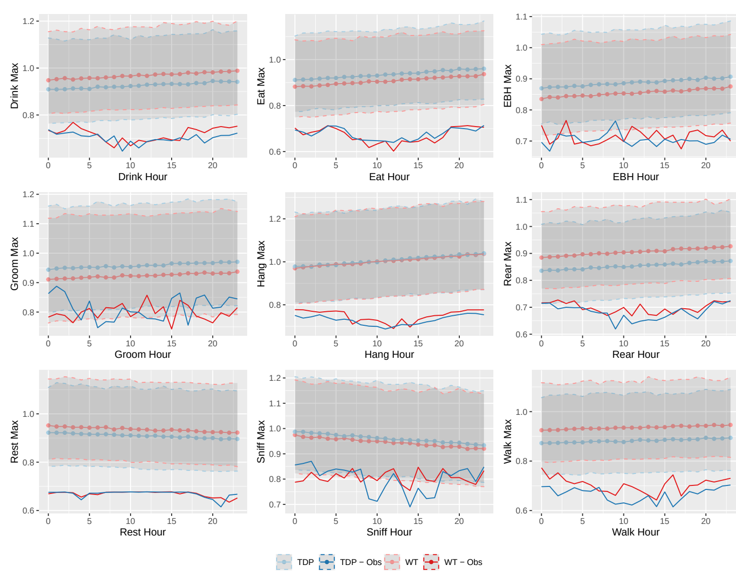

Figure 4.19: Maximum test statistic posterior predictive plot for the hierarchical multivariate normal-normal model.

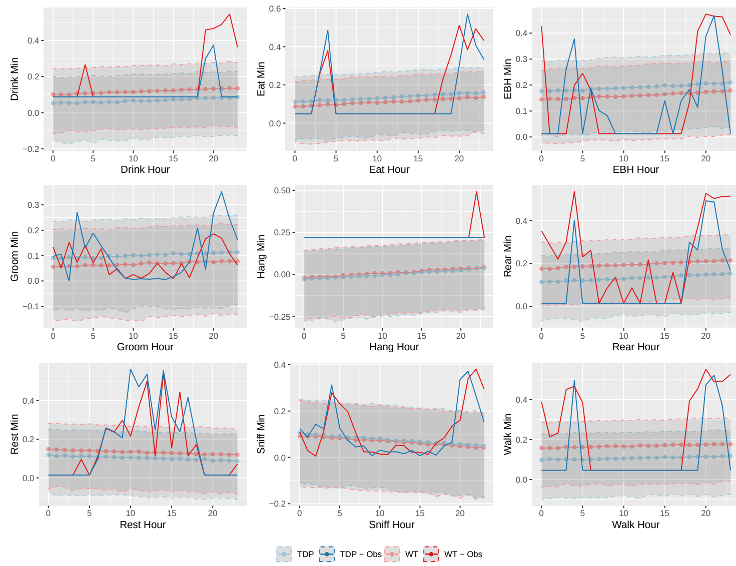

Figure 4.20: Minimum test statistic posterior predictive plot for the hierarchical multivariate normal-normal model.

4.11.4 Hierarchical Multivariate Normal-Normal with Cyclic Splines Posterior Predictive Plots

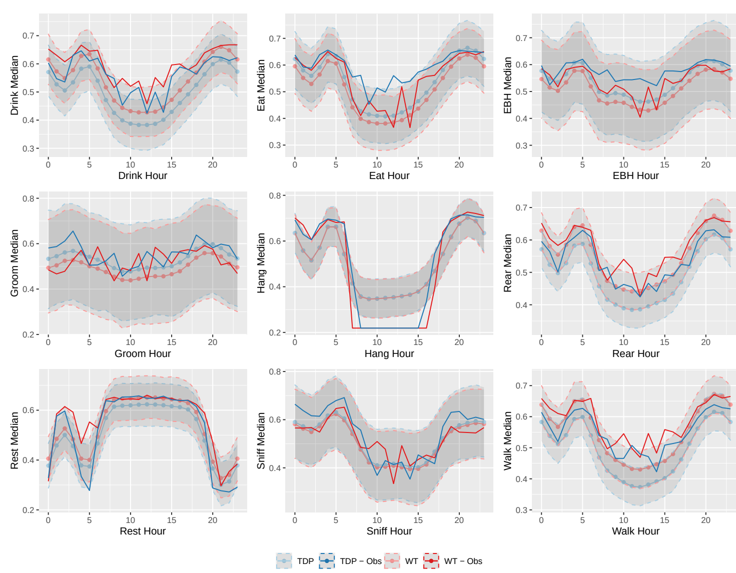

Figure 4.21: Median test statistic posterior predictive plot for the hierarchical multivariate normal-normal model with cyclic splines.

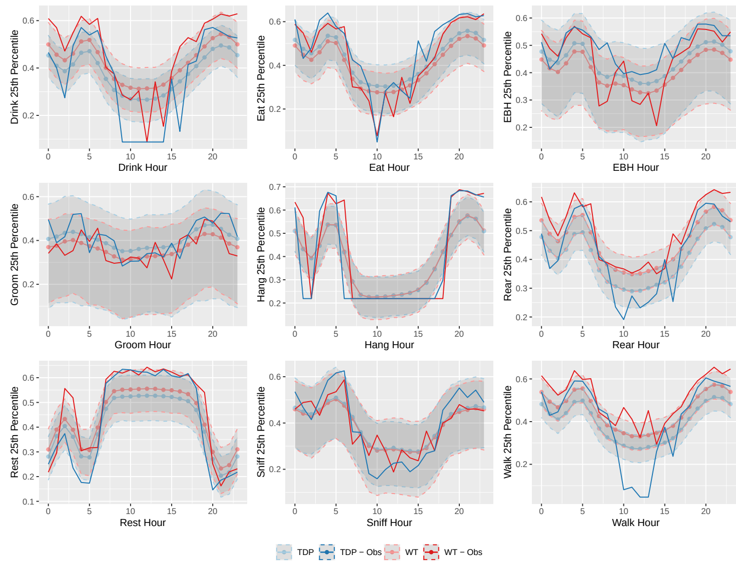

Figure 4.22: 25^{th}-percentile test statistic posterior predictive plot for the hierarchical multivariate normal-normal model with cyclic splines.

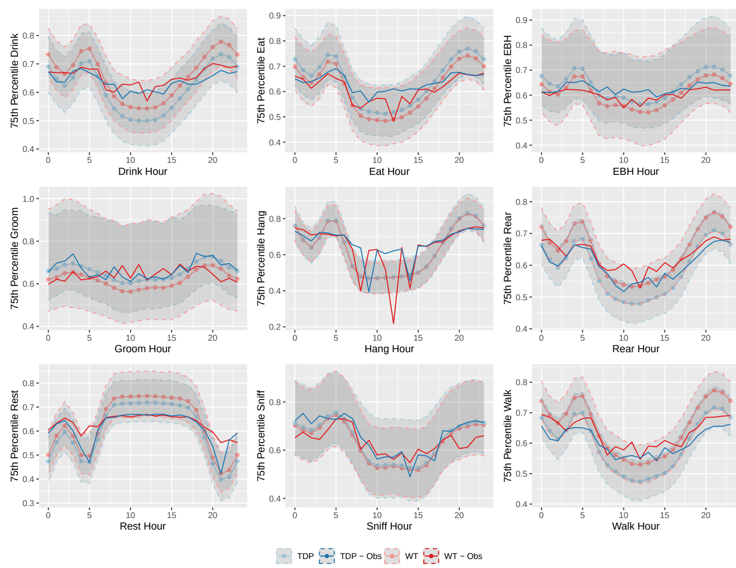

Figure 4.23: 75^{th}-percentile test statistic posterior predictive plot for the hierarchical multivariate normal-normal model with cyclic splines.

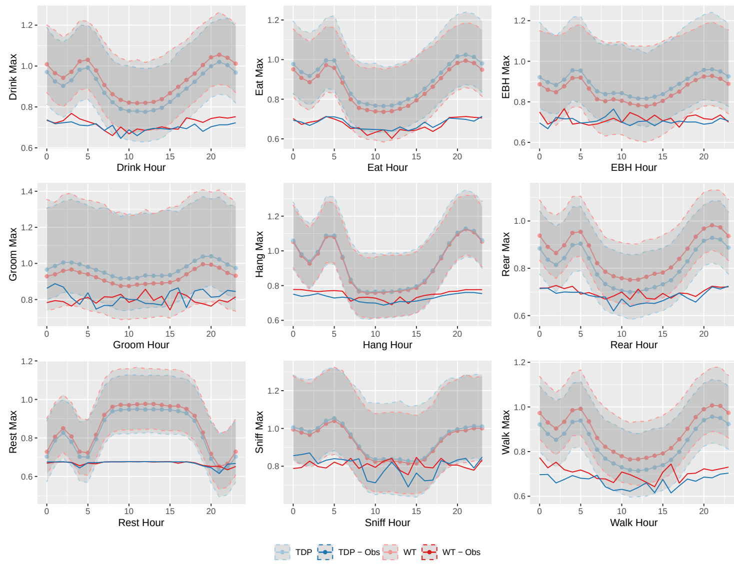

Figure 4.24: Maximum test statistic posterior predictive plot for the hierarchical multivariate normal-normal model with cyclic splines.

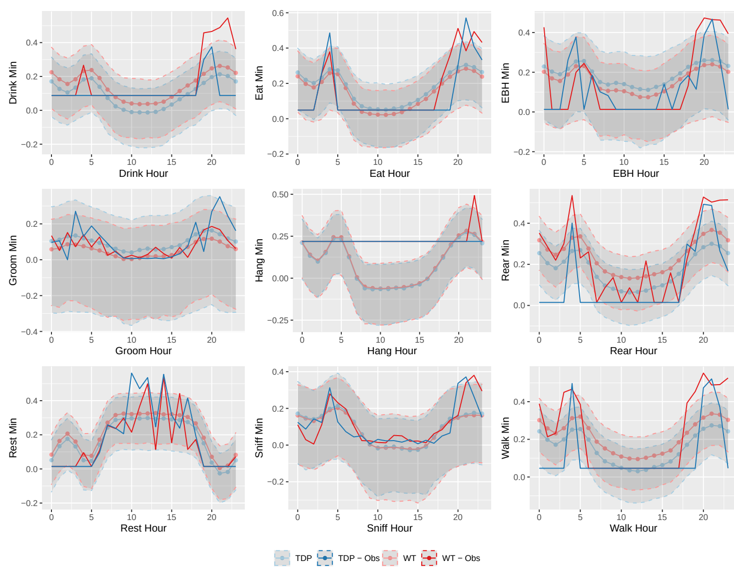

Figure 4.25: Minimum test statistic posterior predictive plot for the hierarchical multivariate normal-normal model with cyclic splines.

4.11.5 Hierarchical Dirichlet with Cyclic Splines Posterior Predictive Plots

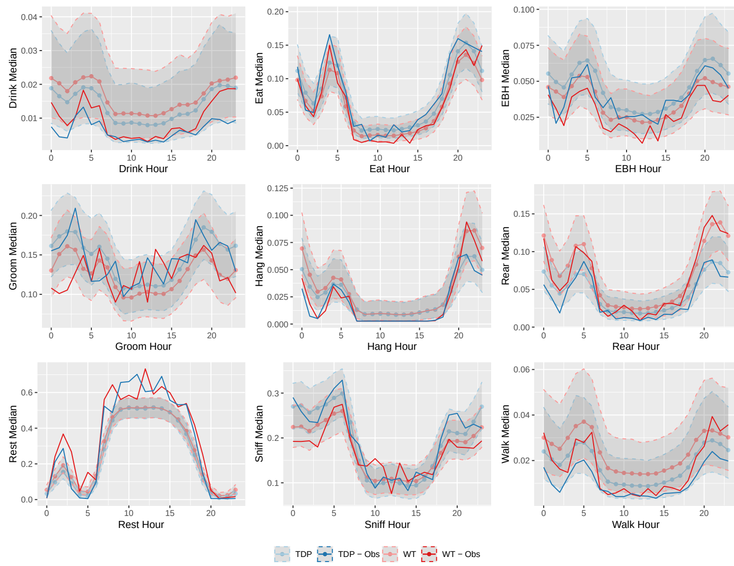

Figure 4.26: Median test statistic posterior predictive plot for the hierarchical Dirichlet model with cyclic splines.

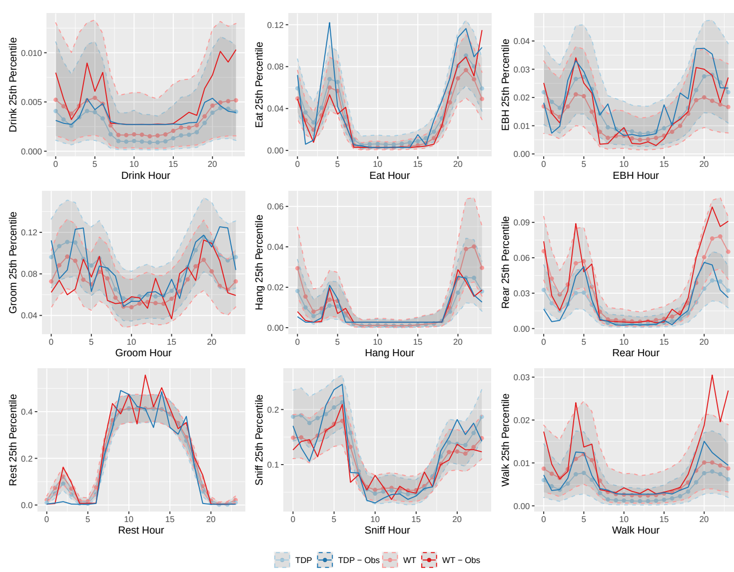

Figure 4.27: 25^{th}-percentile test statistic posterior predictive plot for the hierarchical Dirichlet model with cyclic splines.

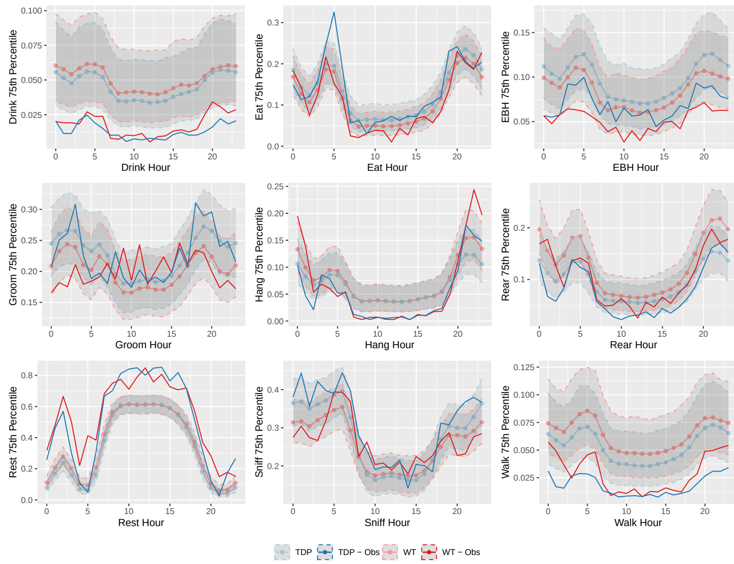

Figure 4.28: 75^{th}-percentile test statistic posterior predictive plot for the hierarchical Dirichlet model with cyclic splines.

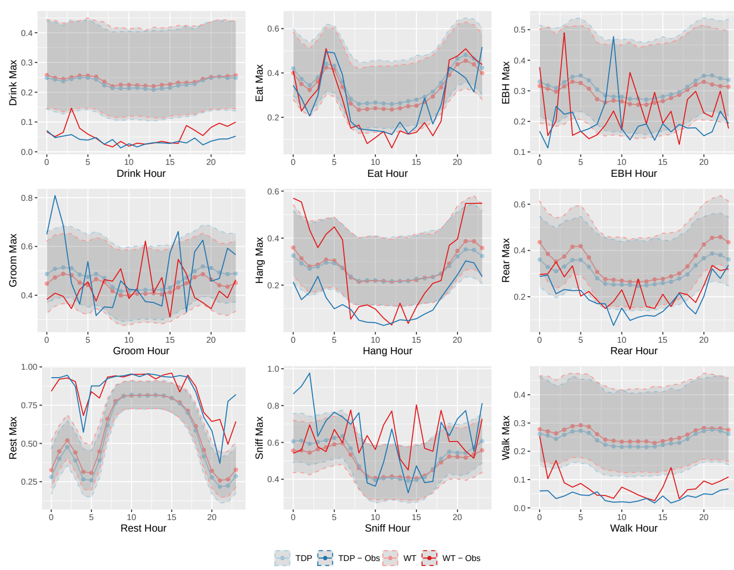

Figure 4.29: Maximum test statistic posterior predictive plot for the hierarchical Dirichlet model with cyclic splines.

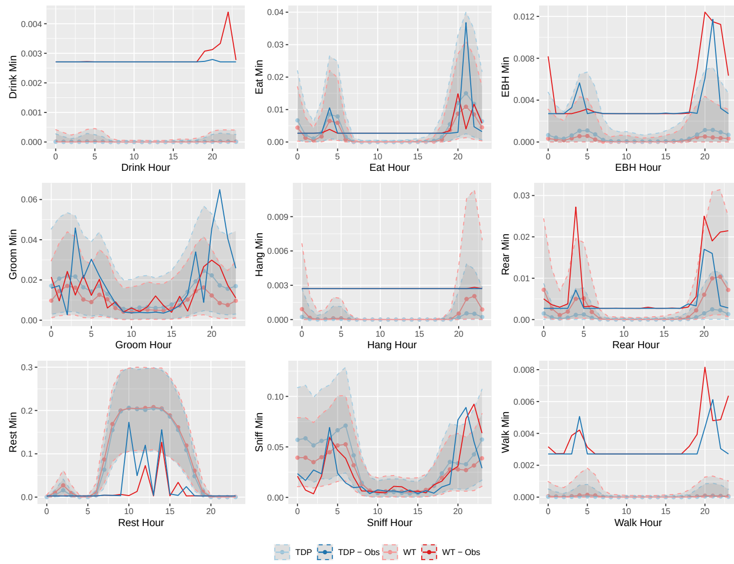

Figure 4.30: Minimum test statistic posterior predictive plot for the hierarchical Dirichlet model with cyclic splines.

4.11.6 Hierarchical Dirichlet with Skewed Link and Cyclic Splines Posterior Predictive Plots

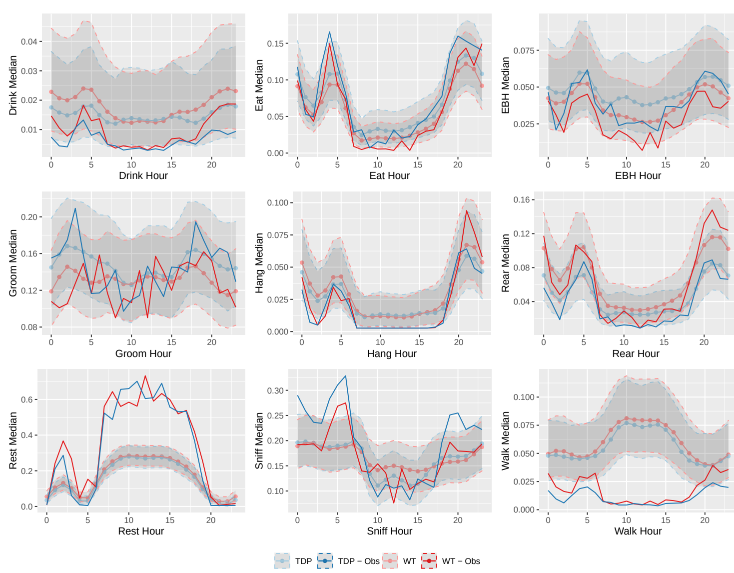

Figure 4.31: Median test statistic posterior predictive plot for the hierarchical Dirichlet model with skewed link and cyclic splines.

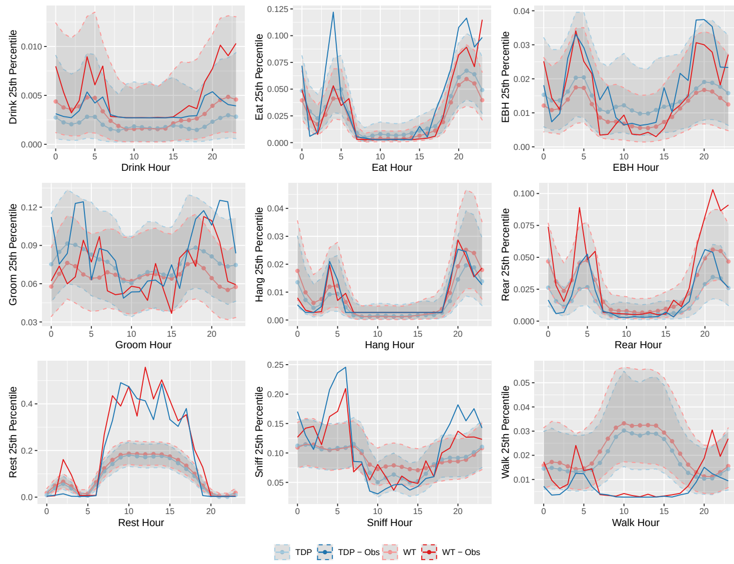

Figure 4.32: 25^{th}-percentile test statistic posterior predictive plot for the hierarchical Dirichlet model with skewed link and cyclic splines.

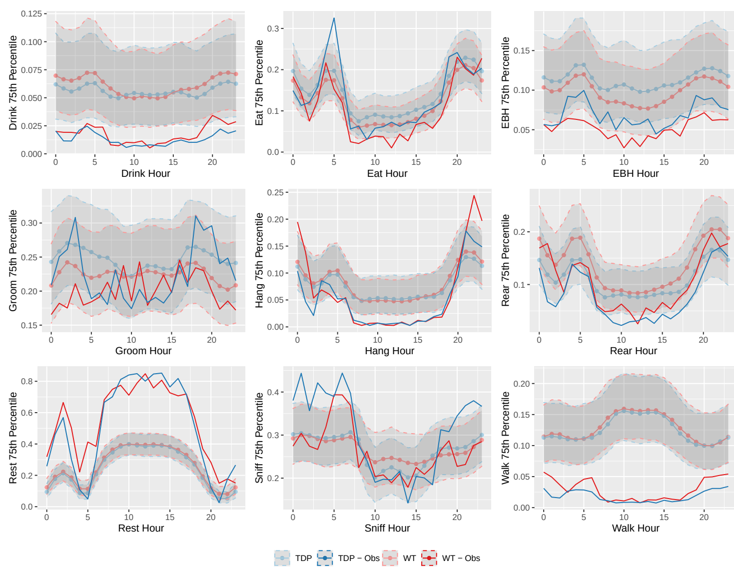

Figure 4.33: 75^{th}-percentile test statistic posterior predictive plot for the hierarchical Dirichlet model with skewed link and cyclic splines.

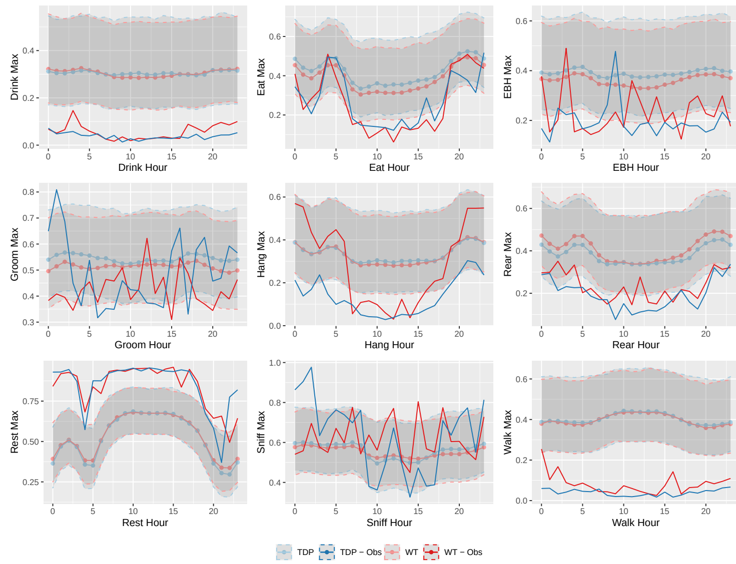

Figure 4.34: Maximum test statistic posterior predictive plot for the hierarchical Dirichlet model with skewed link and cyclic splines.

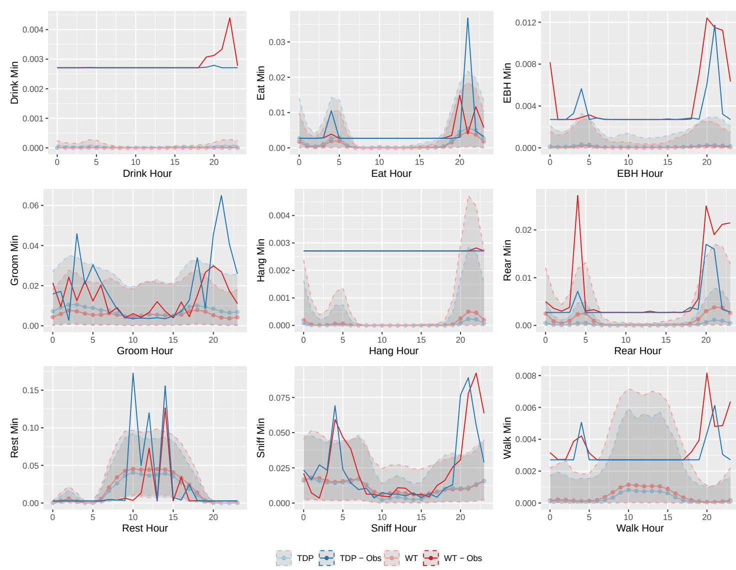

Figure 4.35: Minimum test statistic posterior predictive plot for the hierarchical Dirichlet model with skewed link and cyclic splines.

4.11.7 Zero-Inflated Hierarchical Generalized Dirichlet Multinomial Model with Cyclic Splines Posterior Predictive Plots

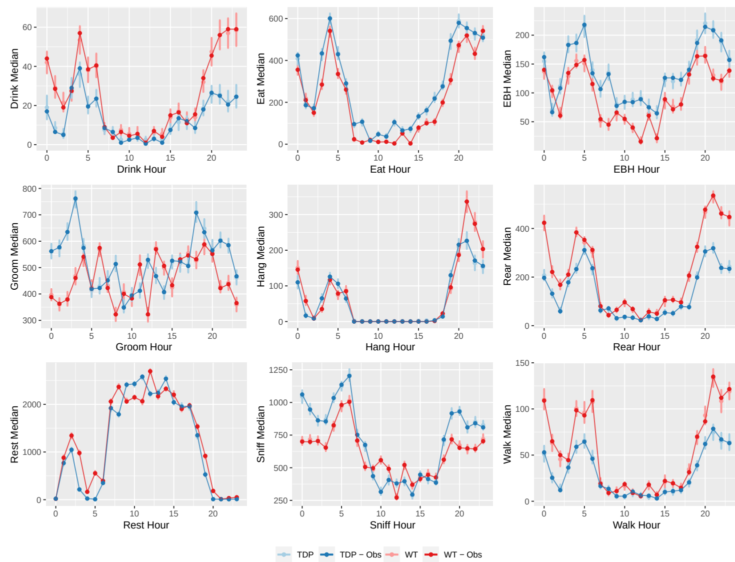

Figure 4.36: Median test statistic posterior predictive plot for the ZIHGDM model with cyclic splines.

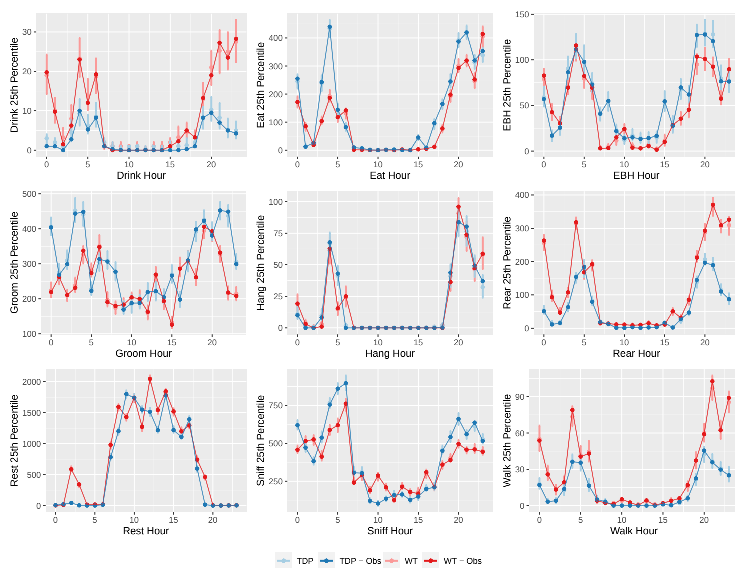

Figure 4.37: 25^{th}-percentile test statistic posterior predictive plot for the ZIHGDM model with cyclic splines.

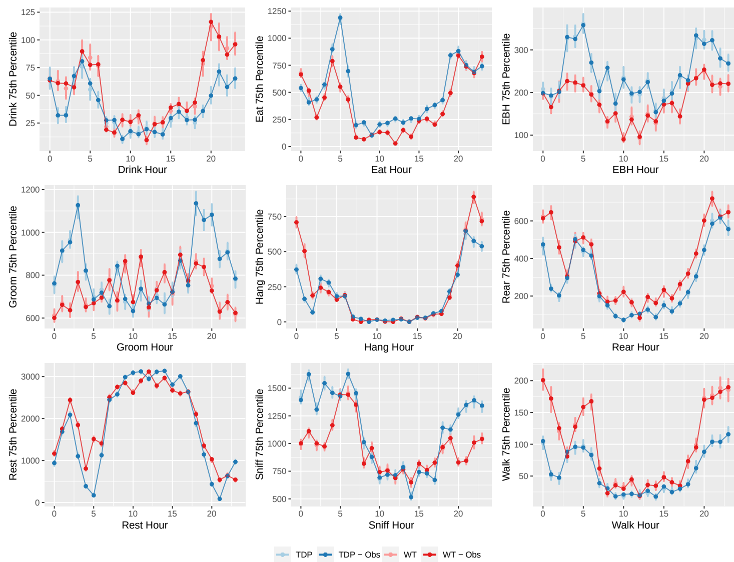

Figure 4.38: 75^{th}-percentile test statistic posterior predictive plot for the ZIHGDM model with cyclic splines.

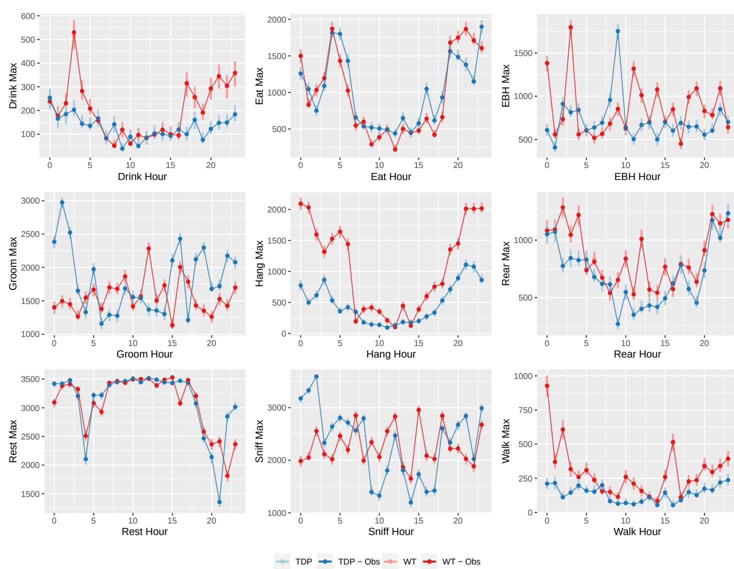

Figure 4.39: Maximum test statistic posterior predictive plot for the ZIHGDM model with cyclic splines.

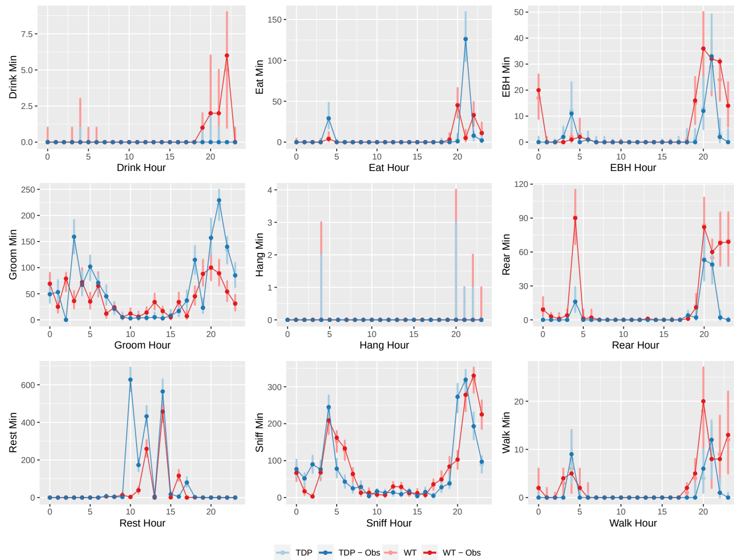

Figure 4.40: Minimum test statistic posterior predictive plot for the ZIHGDM model with cyclic splines.

4.11.8 Model Comparison Results

Table 4.3: Mean time in seconds per hour performing Drink, Eat, and EBH at all hours for TDP-43 and WT mice from the ZIHGDM, hierarchical dirichlet (HDIR), and hierarchical multivariate normal-normal (HMVNN) models, each with cyclic splines. 95% credible intervals are reported in parenthesis.

		ZIHGDM		HDIR		HMVNN	
Hour	Behavior	TDP-43	WT	TDP-43	WT	TDP-43	WT
0	Drink	34 (29,39)	49 (42,56)	153 (136,170)	164 (145,186)	45 (44,46)	48 (47,49)
1	Drink	26 (22,31)	40 (34,45)	143 (128,159)	158 (139,175)	44 (43,44)	46 (45,47)
2	Drink	23 (18,28)	36 (31,42)	135 (118,154)	149 (127,170)	43 (42,44)	45 (43,46)
3	Drink	28 (23,33)	42 (36,49)	145 (129,162)	158 (138,178)	44 (43,45)	46 (45,47)
4	Drink	37 (32,42)	53 (45,61)	153 (134,171)	166 (146,190)	46 (45,47)	49 (48,50)
5	Drink	37 (32,43)	51 (43,60)	153 (132,174)	167 (144,194)	46 (45,47)	49 (48,51)
6	Drink	29 (25,33)	37 (31,43)	144 (126,163)	160 (140,180)	44 (43,45)	46 (45,48)
7	Drink	21 (17,25)	26 (22,31)	121 (103,139)	134 (113,155)	41 (40,42)	43 (42,44)
8	Drink	17 (14,21)	23 (20,27)	107 (94,121)	119 (103,135)	40 (39,40)	41 (40,42)
9	Drink	16 (12,20)	24 (19,30)	106 (93,123)	119 (102,137)	39 (37,40)	40 (39,41)
10	Drink	16 (12,20)	25 (21,31)	107 (94,122)	120 (105,136)	38 (37,39)	39 (38,40)
11	Drink	15 (12,19)	26 (21,32)	106 (92,122)	119 (101,136)	38 (37,39)	39 (38,40)
12	Drink	15 (11,19)	25 (20,31)	104 (92,120)	117 (98,134)	38 (37,39)	39 (38,40)
13	Drink	14 (11,18)	25 (21,30)	104 (91,119)	117 (101,134)	38 (37,39)	39 (38,40)
14	Drink	15 (11,19)	26 (21,32)	107 (92,125)	120 (102,139)	39 (38,40)	40 (39,41)
15	Drink	18 (14,22)	30 (25,36)	113 (100,130)	126 (110,143)	40 (39,41)	41 (40,42)
16	Drink	21 (16,25)	33 (27,40)	118 (104,138)	131 (111,151)	41 (40,42)	43 (41,44)
17	Drink	22 (18,26)	31 (26,36)	120 (108,135)	131 (114,147)	42 (41,43)	44 (43,45)
18	Drink	24 (20,28)	29 (25,35)	125 (110,141)	134 (116,153)	43 (42,45)	46 (44,47)
19	Drink	29 (25,32)	34 (29,40)	139 (120,157)	147 (129,170)	45 (44,46)	48 (46,49)
20	Drink	36 (32,41)	47 (40,54)	152 (131,170)	158 (137,187)	46 (45,48)	50 (48,51)
21	Drink	41 (36,47)	58 (49,67)	157 (136,177)	162 (138,198)	47 (46,48)	51 (49,52)
22	Drink	40 (35,45)	57 (49,66)	157 (140,174)	164 (144,191)	47 (46,48)	50 (49,51)
23	Drink	34 (29,39)	49 (42,56)	153 (136,170)	164 (145,186)	45 (44,46)	48 (47,49)
0	Eat	477 (431,524)	413 (368,459)	484 (450,523)	435 (394,480)	604 (592,616)	605 (590,619)
1	Eat	348 (315,383)	282 (250,316)	370 (345,398)	326 (295,356)	583 (571,594)	576 (563,589)
2	Eat	296 (260,334)	232 (201,267)	315 (286,345)	274 (241,307)	572 (560,585)	562 (547,577)
3	Eat	366 (330,407)	300 (265,339)	393 (360,428)	350 (314,387)	588 (576,600)	583 (570,597)
4	Eat	506 (456,561)	443 (394,495)	525 (480,573)	486 (438,532)	614 (600,627)	618 (602,633)
5	Eat	489 (440,547)	425 (376,479)	511 (467,557)	473 (425,523)	610 (596,624)	613 (596,629)
6	Eat	300 (264,339)	236 (205,271)	345 (315,379)	300 (269,336)	571 (559,582)	559 (546,573)
7	Eat	175 (147,207)	123 (101,148)	208 (182,234)	167 (146,191)	531 (519,543)	506 (492,520)
8	Eat	143 (120,168)	96 (80,115)	167 (148,187)	131 (116,149)	512 (500,523)	481 (469,493)
9	Eat	146 (120,176)	99 (79,120)	172 (151,194)	137 (116,157)	506 (494,519)	474 (461,487)
10	Eat	149 (122,177)	101 (82,123)	175 (156,196)	140 (122,159)	504 (492,515)	471 (460,482)
11	Eat	147 (118,177)	100 (80,123)	170 (151,191)	136 (119,155)	503 (491,516)	470 (457,483)
12	Eat	148 (120,178)	101 (80,125)	169 (150,190)	136 (119,155)	505 (493,517)	472 (459,486)
13	Eat	156 (128,184)	107 (87,132)	176 (157,195)	143 (125,161)	510 (498,521)	479 (465,490)
14	Eat	168 (138,200)	118 (94,144)	186 (165,207)	152 (129,173)	517 (506,529)	489 (475,502)
15	Eat	184 (157,213)	131 (109,155)	198 (179,218)	161 (141,181)	528 (517,539)	503 (490,514)
16	Eat	209 (177,243)	152 (127,183)	219 (195,244)	179 (155,204)	543 (531,555)	522 (507,537)
17	Eat	253 (223,287)	192 (167,223)	260 (237,284)	218 (195,244)	562 (550,573)	547 (535,561)
18	Eat	327 (291,365)	262 (229,299)	338 (305,370)	295 (263,329)	583 (570,595)	576 (561,591)
19	Eat	444 (397,492)	379 (337,428)	462 (421,506)	416 (374,460)	603 (590,616)	604 (589,619)
20	Eat	582 (527,644)	526 (471,579)	582 (532,632)	529 (480,578)	619 (606,634)	625 (610,641)
21	Eat	657 (594,730)	607 (540,671)	628 (578,684)	567 (515,628)	625 (612,640)	634 (617,652)
22	Eat	607 (552,668)	551 (499,603)	585 (545,627)	526 (481,576)	620 (608,633)	627 (613,641)
23	Eat	477 (431,524)	413 (368,459)	484 (450,523)	435 (394,480)	604 (592,616)	605 (590,619)
0	EBH	191 (171,212)	166 (147,185)	289 (259,314)	255 (220,295)	342 (337,346)	339 (333,345)
1	EBH	169 (151,186)	144 (129,161)	266 (240,289)	240 (212,271)	336 (332,340)	330 (325,336)
2	EBH	167 (146,190)	142 (124,162)	253 (226,280)	229 (197,262)	334 (330,339)	327 (320,333)
3	EBH	198 (177,224)	172 (152,193)	284 (260,310)	257 (225,291)	339 (335,344)	335 (330,342)
4	EBH	239 (211,267)	211 (188,236)	316 (287,345)	284 (249,321)	347 (342,352)	349 (342,355)
5	EBH	238 (210,267)	210 (184,235)	321 (289,354)	279 (244,317)	347 (341,352)	348 (342,355)
6	EBH	188 (167,210)	162 (143,182)	295 (266,331)	242 (208,277)	337 (333,342)	331 (326,337)
7	EBH	145 (126,166)	120 (102,141)	238 (209,274)	188 (155,221)	328 (324,333)	316 (309,322)
8	EBH	136 (117,154)	110 (94,128)	203 (183,230)	169 (144,194)	326 (322,331)	312 (307,318)
9	EBH	140 (119,161)	113 (95,134)	197 (177,220)	175 (149,204)	327 (323,332)	314 (308,320)
10	EBH	134 (115,152)	108 (91,126)	195 (175,215)	176 (153,202)	327 (322,331)	313 (308,319)
11	EBH	122 (104,142)	98 (82,115)	190 (170,211)	168 (146,192)	324 (319,329)	309 (303,315)
12	EBH	117 (100,136)	93 (78,110)	186 (167,207)	162 (138,188)	322 (317,328)	306 (299,312)
13	EBH	121 (103,138)	95 (81,112)	186 (169,205)	163 (140,187)	322 (317,327)	305 (299,311)
14	EBH	130 (109,148)	102 (84,121)	190 (170,212)	168 (142,193)	324 (319,329)	308 (301,314)
15	EBH	140 (121,159)	112 (95,129)	200 (182,219)	176 (153,200)	327 (322,331)	313 (308,318)
16	EBH	153 (131,176)	126 (107,146)	212 (190,237)	188 (158,220)	331 (326,335)	320 (314,326)
17	EBH	172 (152,194)	145 (128,164)	227 (206,254)	207 (178,238)	336 (331,340)	329 (323,334)
18	EBH	197 (175,220)	171 (150,193)	253 (224,285)	236 (204,272)	340 (336,345)	337 (331,344)
19	EBH	228 (203,255)	201 (178,227)	293 (262,328)	268 (234,304)	345 (340,350)	345 (339,351)
20	EBH	254 (226,284)	226 (202,252)	322 (292,352)	276 (242,311)	348 (343,353)	350 (343,356)
21	EBH	253 (224,284)	224 (198,252)	324 (291,356)	267 (232,307)	348 (343,354)	351 (344,357)
22	EBH	224 (203,247)	197 (178,218)	309 (282,335)	261 (229,293)	346 (341,350)	347 (341,352)
23	EBH	191 (171,212)	166 (147,185)	289 (259,314)	255 (220,295)	342 (337,346)	339 (333,345)

Table 4.4: Mean time in seconds per hour performing Groom, Hang, and Rear at all hours for TDP-43 and WT mice from the ZIHGDM, hierarchical dirichlet (HDIR), and hierarchical multivariate normal-normal (HMVNN) models, each with cyclic splines. 95% credible intervals are reported in parenthesis.

		ZIHGDM		HDIR		HMVNN	
Hour	Behavior	TDP-43	WT	TDP-43	WT	TDP-43	WT
0	Groom	605 (550,664)	514 (475,555)	654 (607,701)	548 (501,593)	752 (748,757)	746 (742,750)
1	Groom	629 (575,687)	537 (498,578)	698 (647,759)	617 (568,661)	754 (749,759)	748 (744,751)
2	Groom	684 (611,760)	572 (523,622)	718 (650,800)	652 (597,706)	758 (752,764)	750 (746,754)
3	Groom	713 (646,782)	583 (540,632)	714 (662,776)	638 (585,684)	759 (753,764)	751 (747,755)
4	Groom	689 (628,749)	558 (515,603)	643 (586,702)	554 (503,609)	756 (751,761)	749 (745,754)
5	Groom	657 (595,722)	541 (496,588)	615 (553,682)	533 (483,589)	753 (748,759)	747 (742,751)
6	Groom	622 (566,681)	535 (492,577)	649 (599,700)	591 (544,639)	751 (746,755)	745 (741,749)
7	Groom	562 (499,631)	506 (460,554)	600 (552,654)	559 (508,611)	747 (742,753)	743 (739,747)
8	Groom	506 (451,565)	473 (432,516)	511 (469,553)	475 (437,514)	744 (739,749)	740 (736,744)
9	Groom	474 (412,537)	452 (404,500)	469 (422,513)	431 (389,473)	741 (735,747)	738 (733,743)
10	Groom	471 (418,529)	450 (405,493)	467 (426,505)	430 (392,471)	741 (735,746)	738 (734,742)
11	Groom	484 (424,543)	459 (414,504)	481 (441,523)	448 (406,498)	742 (737,748)	739 (735,744)
12	Groom	493 (435,555)	464 (421,511)	486 (444,530)	455 (410,506)	744 (738,749)	740 (736,745)
13	Groom	493 (439,549)	464 (421,503)	479 (439,519)	446 (405,488)	744 (739,749)	741 (736,745)
14	Groom	494 (435,555)	464 (422,507)	481 (438,524)	444 (399,492)	744 (739,750)	741 (736,745)
15	Groom	510 (455,565)	473 (436,514)	505 (465,546)	464 (425,507)	746 (740,750)	742 (738,745)
16	Groom	553 (489,622)	498 (454,550)	550 (500,601)	503 (454,559)	749 (742,754)	744 (739,748)
17	Groom	634 (573,701)	545 (504,593)	609 (562,660)	554 (509,610)	754 (749,759)	748 (744,752)
18	Groom	730 (664,800)	597 (548,646)	682 (627,743)	612 (560,673)	760 (755,765)	752 (748,757)
19	Groom	793 (727,864)	624 (572,681)	736 (684,793)	641 (590,694)	764 (759,769)	755 (750,760)
20	Groom	790 (723,857)	610 (560,664)	715 (662,771)	594 (548,643)	764 (759,770)	755 (750,761)
21	Groom	729 (662,799)	568 (523,622)	662 (602,723)	529 (477,586)	761 (755,767)	753 (748,757)
22	Groom	649 (597,705)	528 (492,567)	637 (589,687)	511 (467,554)	756 (751,760)	749 (745,752)
23	Groom	605 (550,664)	514 (475,555)	654 (607,701)	548 (501,593)	752 (748,757)	746 (742,750)
0	Hang	185 (160,210)	248 (214,284)	265 (240,291)	334 (298,373)	38 (36,39)	39 (38,41)
1	Hang	122 (99,141)	166 (138,197)	209 (192,227)	248 (222,275)	34 (33,35)	35 (33,36)
2	Hang	94 (74,116)	132 (103,163)	173 (155,191)	192 (166,219)	32 (31,33)	32 (31,34)
3	Hang	121 (102,142)	169 (142,200)	189 (171,206)	204 (180,230)	35 (33,36)	35 (34,37)
4	Hang	162 (141,186)	225 (195,257)	218 (197,240)	238 (209,269)	39 (38,41)	41 (39,43)
5	Hang	165 (143,190)	227 (196,261)	216 (194,237)	234 (204,264)	39 (38,41)	41 (39,43)
6	Hang	112 (90,134)	154 (123,185)	177 (160,194)	183 (161,207)	33 (32,35)	34 (32,35)
7	Hang	50 (35,69)	70 (48,94)	128 (112,145)	126 (107,147)	28 (26,29)	27 (26,28)
8	Hang	29 (19,42)	41 (28,57)	108 (97,120)	106 (91,120)	25 (24,26)	24 (23,26)
9	Hang	25 (15,38)	35 (22,51)	108 (96,122)	108 (91,126)	25 (24,26)	24 (23,25)
10	Hang	24 (16,36)	34 (22,48)	110 (99,122)	110 (95,127)	25 (24,26)	24 (23,25)
11	Hang	25 (16,36)	35 (23,50)	108 (96,120)	108 (92,126)	25 (24,26)	24 (23,25)
12	Hang	26 (17,38)	37 (24,53)	105 (93,117)	106 (91,123)	25 (24,26)	24 (23,25)
13	Hang	29 (18,42)	41 (27,58)	105 (94,115)	105 (90,121)	25 (24,26)	24 (23,26)
14	Hang	33 (20,48)	46 (28,67)	108 (95,120)	108 (91,125)	26 (25,27)	25 (24,26)
15	Hang	36 (24,50)	50 (33,69)	114 (103,126)	115 (100,131)	26 (25,27)	25 (24,26)
16	Hang	42 (29,58)	60 (41,81)	122 (108,137)	122 (105,141)	28 (26,29)	27 (26,28)
17	Hang	57 (43,74)	81 (61,103)	128 (116,142)	126 (111,142)	30 (29,31)	30 (29,31)
18	Hang	86 (68,105)	123 (98,149)	146 (131,161)	145 (125,168)	33 (32,35)	34 (32,35)
19	Hang	137 (116,160)	196 (167,229)	194 (175,212)	205 (178,233)	37 (36,38)	38 (37,40)
20	Hang	205 (182,231)	286 (255,324)	262 (237,288)	305 (269,342)	40 (39,42)	42 (40,44)
21	Hang	250 (220,283)	340 (300,387)	309 (275,346)	386 (337,436)	42 (40,43)	44 (42,46)
22	Hang	240 (216,267)	321 (284,365)	306 (276,338)	392 (354,435)	41 (39,42)	43 (41,45)
23	Hang	185 (160,210)	248 (214,284)	265 (240,291)	334 (298,373)	38 (36,39)	39 (38,41)
0	Rear	319 (287,353)	456 (416,503)	349 (320,379)	513 (464,565)	438 (432,445)	456 (449,463)
1	Rear	234 (209,258)	320 (289,351)	287 (265,308)	402 (366,438)	423 (417,428)	441 (434,447)
2	Rear	193 (167,217)	259 (228,293)	247 (225,269)	329 (291,369)	415 (408,421)	433 (425,439)
3	Rear	229 (203,254)	322 (287,358)	282 (259,304)	376 (336,417)	424 (418,431)	442 (435,449)
4	Rear	308 (275,339)	454 (412,503)	336 (308,366)	467 (419,517)	441 (434,448)	459 (452,467)
5	Rear	322 (287,357)	472 (426,521)	339 (309,370)	472 (417,527)	443 (435,451)	461 (454,469)
6	Rear	238 (210,266)	325 (289,360)	278 (253,300)	363 (323,408)	424 (417,430)	442 (436,449)
7	Rear	158 (132,182)	198 (170,230)	198 (178,220)	239 (207,277)	402 (395,409)	420 (413,427)
8	Rear	130 (110,151)	163 (142,187)	162 (146,179)	189 (166,213)	390 (384,397)	408 (402,414)
9	Rear	125 (101,147)	164 (138,191)	157 (141,175)	183 (155,212)	385 (378,392)	403 (396,410)
10	Rear	121 (99,142)	161 (138,185)	155 (140,171)	180 (157,205)	382 (375,388)	400 (393,406)
11	Rear	116 (95,138)	153 (130,178)	151 (134,168)	173 (148,199)	380 (374,387)	398 (391,405)
12	Rear	116 (95,138)	150 (127,175)	149 (132,165)	171 (147,196)	381 (374,387)	399 (392,406)
13	Rear	119 (98,139)	155 (132,179)	151 (137,166)	175 (153,199)	383 (377,390)	401 (395,407)
14	Rear	124 (102,146)	163 (136,190)	156 (140,173)	184 (156,213)	387 (380,393)	404 (397,411)
15	Rear	128 (108,149)	171 (148,194)	165 (150,180)	193 (169,220)	390 (383,396)	407 (401,414)
16	Rear	141 (118,164)	185 (157,214)	177 (157,195)	208 (177,238)	395 (388,402)	413 (406,421)
17	Rear	164 (143,184)	210 (184,238)	194 (176,211)	233 (205,260)	406 (400,413)	424 (418,431)
18	Rear	197 (174,221)	263 (233,298)	229 (210,251)	285 (249,321)	421 (414,428)	439 (432,446)
19	Rear	260 (232,286)	375 (336,416)	292 (268,319)	383 (339,430)	436 (429,443)	454 (447,461)
20	Rear	341 (309,375)	518 (474,564)	359 (328,392)	495 (448,544)	448 (440,455)	466 (458,473)
21	Rear	393 (353,433)	596 (544,649)	395 (358,437)	566 (508,628)	453 (445,461)	471 (463,479)
22	Rear	386 (353,420)	570 (526,615)	390 (356,427)	573 (523,624)	449 (443,456)	467 (461,474)
23	Rear	319 (287,353)	456 (416,503)	349 (320,379)	513 (464,565)	438 (432,445)	456 (449,463)

Table 4.5: Mean time in seconds per hour performing Rest, Sniff, and Walk at all hours for TDP-43 and WT mice from the ZIHGDM, hierarchical dirichlet (HDIR), and hierarchical multivariate normal-normal (HMVNN) models, each with cyclic splines. 95% credible intervals are reported in parenthesis.

		ZIHGDM		HDIR		HMVNN	
Hour	Behavior	TDP-43	WT	TDP-43	WT	TDP-43	WT
0	Rest	406 (341,481)	789 (743,836)	209 (184,233)	287 (258,315)	2077 (2036,2117)	2171 (2131,2212)
1	Rest	682 (596,770)	947 (897,996)	435 (397,471)	550 (510,588)	2237 (2197,2278)	2307 (2269,2346)
2	Rest	807 (689,928)	998 (942,1050)	632 (573,691)	761 (705,822)	2325 (2277,2376)	2381 (2340,2428)
3	Rest	559 (472,654)	908 (855,957)	414 (375,453)	534 (490,580)	2233 (2189,2281)	2304 (2265,2347)
4	Rest	282 (222,342)	750 (695,796)	182 (160,204)	265 (235,294)	2076 (2031,2120)	2171 (2131,2213)
5	Rest	286 (229,350)	768 (715,814)	166 (146,187)	249 (220,280)	2068 (2021,2112)	2163 (2121,2207)
6	Rest	619 (534,707)	996 (944,1048)	417 (383,454)	558 (517,601)	2253 (2206,2296)	2321 (2283,2361)
7	Rest	1187 (1049,1330)	1167 (1113,1227)	1064 (988,1137)	1213 (1144,1282)	2470 (2416,2522)	2503 (2460,2547)
8	Rest	1602 (1486,1724)	1214 (1157,1268)	1608 (1542,1682)	1676 (1613,1739)	2569 (2520,2613)	2585 (2541,2628)
9	Rest	1782 (1650,1906)	1211 (1152,1267)	1814 (1735,1887)	1826 (1757,1890)	2582 (2528,2636)	2596 (2549,2644)
10	Rest	1825 (1715,1928)	1212 (1157,1269)	1853 (1783,1922)	1852 (1790,1912)	2582 (2533,2632)	2596 (2554,2642)
11	Rest	1802 (1677,1917)	1220 (1160,1277)	1842 (1773,1909)	1848 (1782,1908)	2587 (2534,2638)	2600 (2556,2650)
12	Rest	1769 (1644,1890)	1216 (1161,1274)	1838 (1772,1904)	1848 (1781,1910)	2590 (2536,2642)	2603 (2558,2652)
13	Rest	1750 (1637,1860)	1200 (1143,1257)	1851 (1788,1921)	1856 (1790,1914)	2589 (2540,2638)	2602 (2559,2648)
14	Rest	1735 (1614,1858)	1182 (1125,1238)	1839 (1767,1918)	1839 (1772,1906)	2584 (2529,2638)	2597 (2552,2648)
15	Rest	1704 (1589,1808)	1168 (1117,1221)	1766 (1702,1835)	1771 (1712,1833)	2575 (2528,2626)	2590 (2551,2635)
16	Rest	1600 (1459,1717)	1142 (1089,1200)	1615 (1538,1695)	1636 (1566,1705)	2564 (2513,2623)	2581 (2536,2630)
17	Rest	1375 (1260,1483)	1086 (1035,1140)	1362 (1296,1425)	1416 (1356,1475)	2542 (2495,2590)	2562 (2520,2607)
18	Rest	1015 (890,1136)	982 (930,1030)	955 (887,1019)	1052 (988,1116)	2461 (2413,2511)	2496 (2449,2539)
19	Rest	569 (474,661)	813 (763,859)	479 (437,519)	585 (542,632)	2276 (2233,2320)	2340 (2299,2380)
20	Rest	260 (204,317)	623 (573,667)	198 (175,225)	272 (244,302)	2056 (2014,2097)	2153 (2115,2190)
21	Rest	169 (126,218)	538 (485,585)	113 (97,131)	165 (144,187)	1933 (1887,1976)	2047 (2007,2090)
22	Rest	224 (180,271)	616 (567,661)	121 (107,137)	175 (156,194)	1955 (1917,1992)	2067 (2028,2106)
23	Rest	406 (341,481)	789 (743,836)	209 (184,233)	287 (258,315)	2077 (2036,2117)	2171 (2131,2212)
0	Sniff	1179 (1077,1277)	723 (665,779)	1023 (957,1091)	868 (812,922)	1024 (1017,1031)	1016 (1011,1022)
1	Sniff	1191 (1083,1302)	918 (854,982)	1034 (972,1101)	875 (826,928)	1019 (1013,1025)	1012 (1007,1017)
2	Sniff	1143 (1010,1275)	996 (923,1066)	979 (898,1059)	839 (784,896)	1015 (1008,1022)	1010 (1004,1016)
3	Sniff	1195 (1084,1300)	885 (817,952)	1013 (943,1083)	887 (836,938)	1021 (1014,1028)	1014 (1009,1020)
4	Sniff	1189 (1097,1282)	693 (630,752)	1042 (973,1114)	926 (873,979)	1031 (1025,1039)	1021 (1015,1028)
5	Sniff	1236 (1133,1331)	695 (636,754)	1090 (1014,1176)	972 (912,1035)	1035 (1028,1043)	1024 (1018,1031)
6	Sniff	1337 (1218,1446)	945 (880,1013)	1122 (1049,1196)	992 (933,1056)	1028 (1021,1035)	1019 (1013,1025)
7	Sniff	1145 (995,1286)	1173 (1100,1247)	902 (826,984)	803 (745,866)	1013 (1006,1020)	1008 (1002,1014)
8	Sniff	874 (752,996)	1263 (1188,1338)	616 (565,667)	592 (555,633)	995 (989,1001)	995 (990,1000)
9	Sniff	718 (597,849)	1279 (1201,1354)	463 (413,514)	485 (443,524)	981 (973,988)	985 (978,991)
10	Sniff	683 (583,793)	1286 (1209,1362)	426 (383,470)	457 (418,493)	975 (969,982)	981 (975,987)
11	Sniff	713 (595,834)	1293 (1214,1371)	443 (395,491)	468 (427,507)	976 (969,983)	981 (975,987)
12	Sniff	735 (609,863)	1288 (1210,1367)	454 (406,503)	474 (436,513)	977 (970,984)	982 (976,988)
13	Sniff	723 (618,830)	1271 (1194,1349)	439 (394,485)	464 (426,498)	975 (969,982)	981 (975,986)
14	Sniff	703 (587,826)	1250 (1170,1328)	422 (372,475)	453 (413,491)	973 (966,981)	979 (973,985)
15	Sniff	697 (593,807)	1229 (1152,1302)	424 (381,468)	457 (419,491)	973 (967,980)	979 (974,985)
16	Sniff	724 (602,858)	1189 (1112,1263)	466 (415,518)	489 (446,532)	978 (971,986)	983 (977,989)
17	Sniff	796 (687,908)	1111 (1038,1180)	572 (524,625)	563 (525,605)	991 (985,997)	992 (987,997)
18	Sniff	902 (767,1030)	979 (914,1047)	729 (666,799)	674 (631,722)	1006 (998,1013)	1003 (997,1009)
19	Sniff	984 (868,1090)	779 (717,840)	837 (769,910)	764 (717,813)	1016 (1009,1023)	1010 (1004,1016)
20	Sniff	957 (861,1043)	568 (515,623)	825 (763,892)	765 (717,816)	1020 (1013,1027)	1013 (1008,1020)
21	Sniff	931 (838,1023)	471 (420,520)	819 (751,896)	751 (693,807)	1023 (1015,1030)	1015 (1010,1021)
22	Sniff	1041 (952,1129)	541 (491,591)	910 (855,977)	798 (749,847)	1025 (1018,1031)	1017 (1012,1022)
23	Sniff	1179 (1077,1277)	723 (665,779)	1023 (957,1091)	868 (812,922)	1024 (1017,1031)	1016 (1011,1022)
0	Walk	204 (160,259)	244 (199,300)	174 (157,194)	196 (175,221)	103 (101,105)	108 (106,110)
1	Walk	201 (156,260)	246 (201,302)	158 (142,176)	185 (166,210)	99 (97,101)	104 (102,106)
2	Walk	193 (140,264)	232 (185,288)	149 (131,166)	177 (156,200)	97 (95,99)	102 (99,104)
3	Walk	191 (143,255)	219 (177,272)	166 (149,183)	196 (175,217)	99 (97,101)	104 (102,106)
4	Walk	188 (146,242)	214 (174,264)	185 (167,206)	215 (194,238)	104 (102,106)	109 (107,111)
5	Walk	170 (134,215)	210 (171,261)	189 (168,212)	222 (199,247)	104 (102,106)	110 (107,112)
6	Walk	154 (122,192)	211 (173,258)	174 (157,192)	211 (189,234)	99 (97,101)	104 (102,106)
7	Walk	157 (125,195)	216 (178,259)	139 (124,156)	171 (151,192)	93 (91,95)	98 (96,100)
8	Walk	163 (131,204)	217 (183,256)	118 (106,130)	143 (129,159)	90 (88,92)	94 (92,96)
9	Walk	175 (135,218)	223 (185,266)	113 (101,126)	136 (120,152)	88 (86,90)	92 (90,95)
10	Walk	178 (137,221)	221 (182,266)	112 (100,124)	134 (120,149)	87 (85,89)	91 (89,93)
11	Walk	176 (137,218)	218 (181,262)	110 (98,123)	133 (119,147)	86 (84,88)	90 (88,92)
12	Walk	182 (141,226)	225 (188,269)	109 (97,122)	132 (117,147)	86 (84,87)	90 (88,92)
13	Walk	194 (153,239)	241 (200,287)	109 (98,121)	131 (117,146)	86 (84,88)	90 (88,92)
14	Walk	199 (155,248)	249 (204,299)	111 (98,124)	133 (117,148)	87 (85,89)	91 (89,93)
15	Walk	184 (145,230)	236 (196,282)	115 (104,128)	137 (122,152)	88 (86,89)	92 (90,94)
16	Walk	157 (121,199)	216 (177,261)	121 (107,135)	143 (127,160)	89 (87,91)	94 (92,96)
17	Walk	128 (100,165)	199 (164,241)	129 (116,142)	151 (135,167)	93 (91,94)	97 (95,99)
18	Walk	122 (92,161)	194 (158,238)	144 (128,160)	166 (149,185)	97 (95,99)	102 (100,104)
19	Walk	156 (112,217)	198 (161,243)	168 (152,186)	191 (171,212)	101 (99,103)	106 (104,108)
20	Walk	177 (137,228)	198 (163,247)	187 (167,207)	206 (184,231)	104 (102,106)	110 (108,112)
21	Walk	176 (139,224)	199 (161,257)	191 (171,215)	206 (183,235)	106 (103,108)	111 (109,114)
22	Walk	189 (153,237)	219 (181,272)	185 (167,207)	201 (182,227)	105 (103,107)	111 (109,113)
23	Walk	204 (160,259)	244 (199,300)	174 (157,194)	196 (175,221)	103 (101,105)	108 (106,110)

Table 4.6: Mean difference in time (seconds/hour) performing Drink, Eat, and EBH at all hours between TDP-43 and WT mice from the ZIHGDM, hierarchical dirichlet (HDIR), and hierarchical multivariate normal-normal (HMVNN) models, each with cyclic splines. 95% credible intervals are reported in parenthesis.

Hour	Behavior	ZIHGDM	HDIR	HMVNN
0	Drink	-15 (-22,-8)	-11 (-38,7)	-3 (-4,-2)
1	Drink	-13 (-20,-8)	-14 (-32,3)	-2 (-3,-1)
2	Drink	-13 (-20,-8)	-14 (-32,7)	-2 (-3,-1)
3	Drink	-14 (-20,-8)	-13 (-32,6)	-2 (-3,-1)
4	Drink	-16 (-24,-8)	-12 (-37,10)	-3 (-4,-2)
5	Drink	-14 (-22,-6)	-14 (-37,10)	-3 (-4,-2)
6	Drink	-8 (-13,-2)	-16 (-34,8)	-2 (-4,-1)
7	Drink	-5 (-11,0)	-13 (-30,12)	-2 (-3,-1)
8	Drink	-6 (-11,-1)	-13 (-27,5)	-1 (-2,0)
9	Drink	-8 (-14,-3)	-13 (-30,3)	-1 (-2,0)
10	Drink	-10 (-15,-4)	-13 (-28,1)	-1 (-2,0)
11	Drink	-10 (-16,-5)	-13 (-29,4)	-1 (-2,0)
12	Drink	-11 (-16,-5)	-13 (-28,5)	-1 (-2,0)
13	Drink	-11 (-16,-6)	-13 (-30,2)	-1 (-2,0)
14	Drink	-11 (-17,-6)	-13 (-32,2)	-1 (-2,0)
15	Drink	-13 (-19,-7)	-13 (-32,2)	-1 (-2,0)
16	Drink	-12 (-19,-6)	-13 (-32,5)	-2 (-3,-1)
17	Drink	-9 (-14,-3)	-11 (-25,6)	-2 (-3,-1)
18	Drink	-5 (-11,0)	-9 (-23,10)	-2 (-3,-1)
19	Drink	-6 (-11,0)	-8 (-27,10)	-3 (-4,-2)
20	Drink	-11 (-17,-4)	-6 (-31,14)	-3 (-5,-2)
21	Drink	-16 (-26,-7)	-5 (-40,15)	-3 (-5,-2)
22	Drink	-17 (-26,-8)	-7 (-37,11)	-3 (-5,-2)
23	Drink	-15 (-22,-8)	-11 (-38,7)	-3 (-4,-2)
0	Eat	64 (21,113)	49 (8,86)	-1 (-16,14)
1	Eat	66 (34,98)	44 (19,67)	7 (-6,20)
2	Eat	64 (35,92)	41 (18,64)	11 (-2,23)
3	Eat	66 (33,99)	43 (16,69)	5 (-9,18)
4	Eat	63 (15,115)	39 (-2,80)	-4 (-21,12)
5	Eat	64 (19,113)	38 (-1,77)	-3 (-18,13)
6	Eat	64 (35,93)	45 (19,70)	11 (-2,23)
7	Eat	52 (28,76)	41 (18,65)	25 (13,36)
8	Eat	46 (24,70)	36 (17,56)	31 (18,42)
9	Eat	47 (24,71)	35 (17,54)	32 (19,44)
10	Eat	47 (24,73)	35 (18,53)	33 (20,45)
11	Eat	47 (22,73)	34 (17,52)	33 (20,45)
12	Eat	47 (22,73)	33 (16,51)	33 (20,45)
13	Eat	49 (23,74)	33 (16,51)	31 (19,44)
14	Eat	51 (25,76)	34 (17,52)	29 (17,41)
15	Eat	53 (29,79)	37 (20,54)	25 (13,37)
16	Eat	57 (31,83)	40 (22,58)	21 (9,33)
17	Eat	61 (35,88)	41 (24,60)	14 (3,26)
18	Eat	65 (35,96)	43 (21,65)	7 (-6,20)
19	Eat	65 (25,107)	46 (12,81)	0 (-15,14)
20	Eat	57 (-3,122)	53 (0,104)	-6 (-22,10)
21	Eat	50 (-24,130)	61 (-6,120)	-9 (-25,8)
22	Eat	55 (-9,125)	58 (3,109)	-7 (-22,9)
23	Eat	64 (21,113)	49 (8,86)	-1 (-16,14)
0	EBH	25 (7,43)	33 (-22,72)	2 (-3,8)
1	EBH	24 (7,42)	25 (-15,57)	6 (1,12)
2	EBH	25 (8,42)	25 (-20,64)	8 (2,13)
3	EBH	26 (7,46)	28 (-14,64)	4 (-2,10)
4	EBH	28 (3,54)	32 (-14,73)	-2 (-8,4)
5	EBH	28 (3,54)	42 (-2,91)	-2 (-8,4)
6	EBH	26 (8,45)	53 (13,110)	6 (1,11)
7	EBH	25 (7,42)	50 (12,103)	13 (7,18)
8	EBH	25 (7,43)	35 (8,75)	14 (8,20)
9	EBH	26 (8,45)	22 (-9,58)	13 (8,19)
10	EBH	25 (8,44)	19 (-11,46)	14 (8,19)
11	EBH	24 (7,42)	22 (-6,52)	16 (9,21)
12	EBH	24 (6,42)	24 (-4,56)	17 (11,23)
13	EBH	26 (7,44)	23 (-4,53)	17 (11,23)
14	EBH	28 (8,47)	22 (-10,57)	16 (11,22)
15	EBH	28 (9,46)	24 (-3,56)	14 (8,19)
16	EBH	27 (9,44)	24 (-10,63)	11 (5,16)
17	EBH	26 (8,43)	20 (-16,57)	7 (2,12)
18	EBH	26 (7,45)	17 (-28,54)	3 (-2,9)
19	EBH	27 (4,49)	25 (-19,66)	0 (-6,6)
20	EBH	28 (-1,57)	45 (-1,88)	-2 (-8,4)
21	EBH	29 (-1,59)	57 (5,103)	-3 (-9,3)
22	EBH	27 (4,50)	49 (2,86)	-1 (-7,5)
23	EBH	25 (7,43)	33 (-22,72)	2 (-3,8)

Table 4.7: Mean difference in time (seconds/hour) performing Groom, Hang, and Rear at all hours between TDP-43 and WT mice from the ZIHGDM, hierarchical dirichlet (HDIR), and hierarchical multivariate normal-normal (HMVNN) models, each with cyclic splines. 95% credible intervals are reported in parenthesis.

Hour	Behavior	ZIHGDM	HDIR	HMVNN
0	Groom	92 (49,136)	106 (62,157)	6 (2,10)
1	Groom	92 (47,139)	81 (37,133)	7 (3,11)
2	Groom	112 (62,170)	66 (11,139)	8 (3,13)
3	Groom	129 (76,187)	77 (38,124)	8 (3,13)
4	Groom	131 (81,184)	89 (34,146)	7 (3,12)
5	Groom	116 (72,163)	82 (9,151)	6 (2,10)
6	Groom	87 (46,131)	58 (11,106)	5 (1,9)
7	Groom	56 (7,106)	41 (4,80)	4 (0,9)
8	Groom	33 (-20,83)	35 (5,69)	3 (-2,8)
9	Groom	22 (-32,76)	38 (5,73)	3 (-3,8)
10	Groom	21 (-32,73)	38 (4,70)	2 (-3,8)
11	Groom	25 (-28,76)	34 (-6,68)	3 (-3,8)
12	Groom	29 (-25,79)	31 (-12,67)	3 (-2,8)
13	Groom	29 (-22,80)	33 (-6,66)	3 (-2,8)
14	Groom	30 (-22,81)	36 (-2,69)	4 (-1,8)
15	Groom	37 (-13,87)	41 (2,76)	4 (-1,8)
16	Groom	55 (6,103)	48 (0,88)	5 (0,9)
17	Groom	89 (42,136)	55 (0,105)	6 (2,10)
18	Groom	133 (82,191)	69 (7,127)	8 (3,13)
19	Groom	169 (107,240)	95 (46,143)	9 (3,15)
20	Groom	180 (116,253)	121 (76,167)	9 (3,15)
21	Groom	161 (103,226)	133 (72,196)	8 (3,14)
22	Groom	121 (77,171)	126 (68,193)	7 (3,12)
23	Groom	92 (49,136)	106 (62,157)	6 (2,10)
0	Hang	-63 (-95,-34)	-69 (-109,-32)	-1 (-3,0)
1	Hang	-45 (-70,-21)	-39 (-65,-17)	-1 (-2,1)
2	Hang	-37 (-59,-15)	-19 (-40,0)	0 (-1,1)
3	Hang	-48 (-72,-26)	-16 (-35,3)	-1 (-2,1)
4	Hang	-63 (-90,-39)	-20 (-44,2)	-2 (-3,0)
5	Hang	-62 (-89,-38)	-18 (-43,6)	-2 (-3,0)
6	Hang	-42 (-67,-20)	-7 (-27,14)	0 (-2,1)
7	Hang	-20 (-37,-2)	2 (-15,20)	1 (0,2)
8	Hang	-12 (-25,1)	3 (-10,17)	1 (0,2)
9	Hang	-10 (-22,1)	1 (-13,15)	1 (0,2)
10	Hang	-10 (-22,1)	-1 (-15,13)	1 (0,2)
11	Hang	-10 (-22,1)	-1 (-15,13)	1 (0,2)
12	Hang	-11 (-23,1)	0 (-14,13)	1 (0,2)
13	Hang	-12 (-25,1)	0 (-13,14)	1 (0,2)
14	Hang	-13 (-26,1)	0 (-13,15)	1 (0,2)
15	Hang	-14 (-28,0)	0 (-13,13)	1 (0,2)
16	Hang	-17 (-32,-1)	0 (-14,14)	1 (0,2)
17	Hang	-24 (-41,-6)	2 (-12,15)	0 (-1,1)
18	Hang	-37 (-56,-16)	1 (-15,18)	0 (-2,1)
19	Hang	-59 (-85,-37)	-11 (-33,9)	-1 (-3,0)
20	Hang	-82 (-116,-49)	-43 (-76,-13)	-2 (-4,0)
21	Hang	-90 (-136,-46)	-77 (-129,-28)	-2 (-4,0)
22	Hang	-81 (-123,-43)	-85 (-139,-37)	-2 (-4,0)
23	Hang	-63 (-95,-34)	-69 (-109,-32)	-1 (-3,0)
0	Rear	-136 (-174,-103)	-164 (-217,-116)	-18 (-25,-11)
1	Rear	-86 (-112,-61)	-115 (-155,-83)	-18 (-24,-12)
2	Rear	-67 (-90,-43)	-82 (-117,-53)	-18 (-24,-12)
3	Rear	-93 (-119,-66)	-95 (-130,-66)	-18 (-25,-12)
4	Rear	-146 (-185,-112)	-130 (-176,-94)	-18 (-25,-11)
5	Rear	-150 (-188,-115)	-133 (-180,-92)	-18 (-26,-11)
6	Rear	-86 (-114,-61)	-86 (-118,-52)	-18 (-24,-11)
7	Rear	-40 (-66,-18)	-40 (-68,-10)	-18 (-24,-12)
8	Rear	-33 (-57,-12)	-27 (-48,-4)	-18 (-24,-11)
9	Rear	-39 (-63,-17)	-26 (-47,-2)	-18 (-25,-11)
10	Rear	-40 (-63,-19)	-24 (-45,-2)	-18 (-25,-11)
11	Rear	-37 (-60,-15)	-22 (-45,0)	-18 (-25,-11)
12	Rear	-34 (-57,-13)	-22 (-45,-1)	-18 (-25,-11)
13	Rear	-36 (-60,-14)	-24 (-45,-3)	-18 (-25,-11)
14	Rear	-40 (-65,-18)	-27 (-49,-3)	-18 (-25,-11)
15	Rear	-43 (-67,-22)	-29 (-50,-7)	-18 (-25,-11)
16	Rear	-44 (-69,-21)	-32 (-54,-7)	-18 (-24,-11)
17	Rear	-46 (-70,-24)	-39 (-62,-16)	-18 (-24,-12)
18	Rear	-66 (-90,-43)	-56 (-85,-27)	-18 (-24,-12)
19	Rear	-115 (-146,-87)	-91 (-128,-56)	-18 (-25,-11)
20	Rear	-177 (-219,-137)	-136 (-180,-99)	-18 (-26,-10)
21	Rear	-202 (-258,-150)	-170 (-230,-115)	-18 (-26,-9)
22	Rear	-184 (-234,-138)	-182 (-241,-127)	-18 (-26,-10)
23	Rear	-136 (-174,-103)	-164 (-217,-116)	-18 (-25,-11)

Table 4.8: Mean difference in time (seconds/hour) performing Rest, Sniff, and Walk at all hours between TDP-43 and WT mice from the ZIHGDM, hierarchical dirichlet (HDIR), and hierarchical multivariate normal-normal (HMVNN) models, each with cyclic splines. 95% credible intervals are reported in parenthesis.

Hour	Behavior	ZIHGDM	HDIR	HMVNN
0	Rest	-382 (-460,-300)	-78 (-102,-53)	-94 (-140,-56)
1	Rest	-265 (-357,-162)	-114 (-148,-80)	-70 (-113,-33)
2	Rest	-191 (-319,-62)	-129 (-170,-88)	-56 (-98,-19)
3	Rest	-348 (-441,-249)	-121 (-156,-87)	-71 (-111,-33)
4	Rest	-468 (-543,-394)	-83 (-109,-56)	-94 (-138,-54)
5	Rest	-482 (-555,-410)	-83 (-108,-57)	-96 (-139,-55)
6	Rest	-377 (-476,-275)	-141 (-176,-105)	-68 (-109,-31)
7	Rest	20 (-133,171)	-149 (-203,-93)	-33 (-78,6)
8	Rest	389 (265,514)	-67 (-128,-2)	-16 (-64,27)
9	Rest	571 (433,698)	-12 (-78,56)	-14 (-63,31)
10	Rest	612 (493,729)	1 (-62,70)	-14 (-63,30)
11	Rest	582 (446,718)	-6 (-68,65)	-13 (-63,31)
12	Rest	552 (413,688)	-10 (-73,64)	-12 (-62,32)
13	Rest	550 (421,675)	-5 (-71,65)	-13 (-62,33)
14	Rest	553 (416,685)	0 (-68,66)	-14 (-62,32)
15	Rest	536 (411,652)	-4 (-74,60)	-15 (-63,29)
16	Rest	458 (311,591)	-22 (-86,46)	-17 (-65,27)
17	Rest	289 (170,404)	-55 (-111,1)	-21 (-69,21)
18	Rest	33 (-101,157)	-97 (-141,-50)	-34 (-80,7)
19	Rest	-244 (-345,-144)	-107 (-142,-74)	-64 (-105,-28)
20	Rest	-364 (-433,-291)	-74 (-99,-47)	-97 (-143,-56)
21	Rest	-368 (-430,-306)	-52 (-72,-31)	-114 (-165,-68)
22	Rest	-392 (-447,-332)	-54 (-74,-35)	-111 (-160,-66)
23	Rest	-382 (-460,-300)	-78 (-102,-53)	-94 (-140,-56)
0	Sniff	456 (347,565)	155 (87,230)	8 (2,14)
1	Sniff	273 (152,398)	159 (93,236)	6 (0,12)
2	Sniff	148 (4,296)	140 (71,219)	6 (0,11)
3	Sniff	310 (194,430)	127 (67,192)	7 (1,13)
4	Sniff	496 (400,596)	115 (55,185)	10 (3,16)
5	Sniff	540 (447,642)	118 (47,195)	11 (4,18)
6	Sniff	392 (277,507)	130 (60,208)	9 (3,15)
7	Sniff	-28 (-183,135)	99 (32,172)	5 (-1,10)
8	Sniff	-389 (-522,-256)	23 (-19,67)	0 (-5,5)
9	Sniff	-561 (-701,-411)	-22 (-68,21)	-4 (-10,2)
10	Sniff	-603 (-728,-468)	-32 (-76,14)	-5 (-12,1)
11	Sniff	-580 (-724,-439)	-25 (-71,22)	-5 (-12,1)
12	Sniff	-553 (-699,-410)	-20 (-64,25)	-5 (-11,2)
13	Sniff	-548 (-686,-415)	-24 (-65,18)	-5 (-12,1)
14	Sniff	-547 (-691,-399)	-30 (-74,15)	-6 (-13,1)
15	Sniff	-532 (-655,-393)	-32 (-77,14)	-6 (-13,1)
16	Sniff	-465 (-609,-317)	-23 (-74,24)	-5 (-11,2)
17	Sniff	-315 (-441,-187)	9 (-36,53)	-1 (-7,5)
18	Sniff	-77 (-217,68)	55 (4,109)	3 (-2,8)
19	Sniff	205 (76,322)	73 (17,134)	6 (0,11)
20	Sniff	389 (280,490)	59 (-3,119)	7 (1,12)
21	Sniff	460 (363,559)	68 (-3,136)	8 (1,14)
22	Sniff	500 (405,602)	112 (46,179)	8 (2,14)
23	Sniff	456 (347,565)	155 (87,230)	8 (2,14)
0	Walk	-40 (-91,9)	-22 (-39,-4)	-6 (-8,-3)
1	Walk	-45 (-93,5)	-27 (-42,-12)	-5 (-7,-3)
2	Walk	-39 (-90,15)	-28 (-42,-14)	-5 (-7,-3)
3	Walk	-28 (-76,28)	-30 (-45,-14)	-5 (-7,-3)
4	Walk	-26 (-75,23)	-30 (-47,-12)	-6 (-8,-3)
5	Walk	-40 (-86,2)	-33 (-51,-14)	-6 (-8,-3)
6	Walk	-57 (-99,-20)	-37 (-54,-21)	-5 (-7,-3)
7	Walk	-59 (-97,-26)	-31 (-45,-18)	-5 (-7,-3)
8	Walk	-54 (-89,-21)	-26 (-37,-14)	-4 (-6,-3)
9	Walk	-48 (-87,-9)	-23 (-35,-11)	-4 (-6,-3)
10	Walk	-43 (-83,-5)	-22 (-34,-10)	-4 (-6,-2)
11	Walk	-42 (-83,-4)	-23 (-34,-11)	-4 (-6,-2)
12	Walk	-43 (-87,-4)	-23 (-34,-11)	-4 (-6,-2)
13	Walk	-47 (-95,-4)	-23 (-34,-11)	-4 (-6,-2)
14	Walk	-50 (-98,-6)	-22 (-34,-10)	-4 (-6,-2)
15	Walk	-52 (-94,-11)	-22 (-33,-10)	-4 (-6,-3)
16	Walk	-59 (-96,-21)	-22 (-33,-9)	-4 (-6,-3)
17	Walk	-71 (-107,-39)	-22 (-34,-9)	-5 (-6,-3)
18	Walk	-72 (-109,-36)	-23 (-35,-9)	-5 (-7,-3)
19	Walk	-42 (-86,9)	-23 (-38,-8)	-5 (-7,-3)
20	Walk	-21 (-68,27)	-19 (-37,-2)	-6 (-8,-3)
21	Walk	-23 (-69,24)	-15 (-34,5)	-6 (-8,-3)
22	Walk	-30 (-78,15)	-16 (-35,4)	-6 (-8,-3)
23	Walk	-40 (-91,9)	-22 (-39,-4)	-6 (-8,-3)