UNIVERSITY OF CALIFORNIA SAN DIEGO

Measurement Error and Causal Inference: Implications in the analysis of mobile-health data

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Biostatistics

by

Ruohui Chen

Committee in charge:

Professor Loki Natarajan, Co-Chair
Professor Lin Liu, Co-Chair
Professor Sheri Hartman
Professor Andrea LaCroix
Professor Xin Tu

2023

The dissertation of Ruohui Chen is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my parents: for all the unconditional love, understanding and support.

To my advisors: for all the guidance and support.

EPIGRAPH

Don't judge each day by the harvest you reap but by the seeds that you plant.
—Robert Louis Stevenson

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Words can't express my gratitude for all the support and love I got from my family, mentors, colleagues, and friends during the journey of acquiring a Ph.D. in biostatistics. It has been an incredible life experience packed with ups and downs, and I could not imagine where I am at today without that tremendous support. Over the past four and half years I have learned so much from working with many amazing people. Among all the people I had the privilege to work with at UCSD, there were dozens of individuals that not only helped me grow into a mature independent researcher, enlightened me with brilliant ideas and helpful feedback but more importantly shaped me into a better person.

First and foremost, it has been such a great pleasure to work closely with Dr. Loki Natarajan, Dr. Lin Liu, and Dr. Xin Tu over the past few years. Their enthusiasm towards mentoring students and research has had a huge impact on me and shaped my perspectives toward professional career development. My interaction with Dr. Loki Natarajan started on the first day I joined the Ph.D. program. Over the course of being a teaching assistant and graduate student researcher with her, she provided a tremendous amount of support for me to explore my research interests, build statistical analysis skills, and grow as an individual. I have learned so much from our countless meetings over the years, such as brainstorming the best solutions for the questions at hand, cracking complicated statistical inferences for ongoing projects, and strategies to be a great collaborator. Her immense knowledge of statistics and extensive experience in collaboration with other researchers in different fields have encouraged me all the time in my academic research and daily life. In my second year in the Ph.D. program, I started to work with Dr. Lin Liu on mobile-device-related projects. We spent so much time working together to come up with the best methods for the interested questions based on the data we have. Her wisdom, attention to detail, and insightful propositions have not only guided me to provide solid statistical analysis but also helped me to finish the dissertation. Then it is Dr. Xin Tu, who to me is not only a great academic advisor but also a lifetime mentor. I started to work with Dr. Xin Tu by the end of my second year in the Ph.D. program through a rotation, and after that our projects and collaborations never stop. Experience working with him shaped my perspectives toward academia and I started to think about my career as a researcher, collaborator, and educator. Besides all the statistical knowledge I learned from our regular weekly meetings, no matter how busy he is, he is always making sure my questions got answered and always eager to help, no matter whether it is research, work plan, or life in general. I am indebted and grateful for his indispensable guidance, constant encouragement, and tremendous amount of support.

Additionally, I would like to express my sincere gratitude to Dr. Sheri Hartman for her treasured support and feedback over the years while I am developing appropriate and efficient statistical methods to analyze the mobile device-based activity data. I would also like to extend my sincere thanks to Dr. Florin Vaida and Dr. Andrea LaCroix for their great support.

Last but not least, I would like to dedicate this dissertation to my parents for their unconditional love and tremendous support. Without their understanding and encouragement, it would be impossible for me to finish this challenging journey.

Results from Chapter 1 have been published as Hartman, Sheri; Chen, Ruohui; Tam, Rowena; Narayan, Hari; Natarajan, Loki; and Liu, Lin. *Fitbit Use and Activity Levels From Intervention to 2 Years After: Secondary Analysis of a Randomized Controlled Trial.* JMIR Mhealth Uhealth. 2022 Jun 30;10(6):e37086. The dissertation author was one of the coauthors of this paper.

Chapter 2, in full, has been submitted for publication as Chen, Ruohui; Rosenberg, Dori; Di, Chongzhi; Zablocki, Rong; Hartman, Sheri; LaCroix, Andrea; Tu, Xin; Natarajan, Loki and Liu, Lin. *A Linear Mixed Model Approach for Measurement Error Adjustment: Applications to Sedentary Behavior Assessment from Wearable Devices*, submitted to Annals of Applied Statistics. The dissertation author was the primary author of this paper.

Chapter 3, in full, has been submitted for publication as Chen, Ruohui; Lin, Tuo; Liu, Lin; Liu, Jinyuan; Chen, Ruifeng; Zou, Jingjing; Liu, Chenyu; Natarajan, Loki; and Tu, Xin. *A Double Robust Estimator for Mann Whiney Wilcoxon Rank Sum Test When Applied for Causal Inference in Observational Studies*, submitted to Journal of Applied Statistics. The dissertation author was the primary author of this paper.

Data collection and analysis for Chapter 1 were supported by the National Cancer Institute (NCI) and the National Institutes of Health (NIH/NCI K07CA181323). HKN was supported by a Padres Pedal the Cause/Rady grant (PTC2020) and the University of California San Diego Moore Cancer Center, and by a Specialized Cancer Control Support grant (NIH/NCI P30CA023100). Data collection and analysis for Chapter 2 and Chapter 3 were supported by prior funding from the National Institute on Aging (U01AG006781). All statements in this dissertation, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the National Institute on Aging or the National Institutes of Health. We thank the participants of the Adult Changes in Thought (ACT) study for the data they have provided and the many ACT investigators and staff who steward that data.

VITA

2014	Bachelor of Science in Finance. Iowa State University, Ames, IA
2014	Bachelor of Science in Software Engineering. SiChuan University, ChengDu, SiChuan
2018	Master of Science in Biostatistics. University of Pennsylvania, Philadelphia, PA
2018 - 2023	Graduate Researcher. University of California San Diego, CA
2023	Doctor of Philosophy in Biostatistics. University of California San Diego, CA

PUBLICATIONS

Hartman,S., **Chen,R.**, Tam,R., Narayan,H., Natarajan,L., and Liu,L. **Fitbit Use and Activity Levels Two Years Post Interventions: Secondary Analysis of a Randomized Controlled Trial**, *JMIR Mhealth Uhealth*, 2022, 10(6):e37086, doi: 10.2196/37086.

Chen, R., Lin,T., Liu, L., Liu, J., Chen, R., Liu, C., Zou, J., Natarajan, L., and Tu, X.M. **A Double Robust Estimator for Mann Whitney Wilcoxon Rank Sum Test When Applied for Causal Inference in Observational Studies**, submitted.

Sanderson-Cimino, M.*, **Ruohui, C.***, Tu,X.M., Elman,J.A., Jak, A.J, and Kremen, W.S. **Misinterpreting cognitive change over multiple timepoints: When practice effects meet age-related decline**, *Neuropsychology*, accepted, Feb 2023. (* denotes Joint first authors)

Quach E. N., Yang, K., **Chen, R.**, Tu, J., Xu M., Tu, X., and Zhang, X. **Post hoc Power Analysis: Conceptually Valid Approach for Power based on Observed Study Data**,*General Psychiatry*, 2022.

Cohen,J., Potluri,V., Porrett,P., **Chen,R.**, Roselli,M., Shults,J., Sawinski,D., and Reese,P. **Leveraging marginal structural modeling with Cox regression to assess the survival benefit of accepting vs declining kidney allograft offers**, *American Journal of Transplantation*, (2019) 45:14, 2548-2562, doi:10.1111/ajt.15290.

Megan Chiu, Peiting Kuo, Khrissa Lecrone, Andrew Garcia, **Ruohui Chen**, Natalie Quach, Xin M Tu, and David Pride. **Comparison of the APAS Independence Automated Plate Reader System with manual standard-of-care for processing urine culture specimens**, *Microbiology Spectrum*, 2022, 11(5):e01442-22, doi:10.1128/spectrum.01442-22.

ABSTRACT OF THE DISSERTATION

Measurement Error and Causal Inference: Implications in the analysis of mobile-health data

by

Ruohui Chen
Doctor of Philosophy in Biostatistics

University of California San Diego, 2023

Professor Loki Natarajan, Co Chair
Professor Lin Liu, Co Chair

Wearable devices have been gaining popularity in biomedical studies and clinical trials. In recent years, wearable devices have become more common in the study design stage and for data collection purposes. Wearable devices, such as accelerometers and Fitbit, have not only made collecting data for participants much easier than before but also can capture subjects' activities along with other important biometrics more objectively than surveys and other traditional data-collecting methods. However, despite the potential benefit of using those technology-based trackers to collect data and potentially boost wearers' activity levels, very little is known about how individuals use these trackers on a daily basis or how tracker use relates to increasing physical activity or changing sedentary behaviors. Additional research is needed to understand how best to utilize trackers in interventions to support self-monitoring and effectively change behaviors. Furthermore, statistical methods for correcting estimates from activity measures that contained measure-

ment error, and investigating causal inference between lifestyle interventions and activity level have not been fully exploited. There is a need for novel statistical approaches to answer the above questions in both randomized control trials and observational studies.

The goal of this dissertation is to develop appropriate and innovative statistical methods to answer the questions fore-mentioned, while trying to close the gap between available dense continuous mobile health data and appropriate statistical methods.

The dissertation consists of three main chapters. In chapter one, we used minute-level activity data collected from Fitbit trackers in a randomized controlled trial of breast cancer survivors to examine physical activity levels and adherence to Fitbit use. We examined patterns of activity level and Fitbit use for both the 12-week intervention period and the 2-year follow-up period and compared patterns between the intervention group and the control group. We found that within the first 3-month intervention period, the Exercise group has a higher average of MVPA and adherence to Fitbit use than the Wellness group, but the trend of MVPA and adherence to Fitbit use are no differences between the two groups. Besides that, both the Exercise and Wellness group showed a dropping trend of MVPA and adherence to Fitbit use in the follow-up period, but the Exercise group has a much slower dropping trend than the Wellness group.

Realizing the amount of measurement errors and extreme values contained in the activity data captured by those wearable devices in chapter one, and motivated by the existence of measurement errors in sedentary behavior assessment arising from different sources poses serious challenges for conducting statistical analysis and obtaining unbiased estimates, especially without validation data [1], in chapter two, we proposed to use structure models consisting of Linear Mixed Effect Models and Generalized Linear Models to obtain unbiased estimates of the relationship between exposures subject to measurement errors and outcome of interest, after appropriately accounting for the errors in devices' measurement. In the motivating example of chapter two, we found that without accounting for errors in the measurements, we may end up inappropriately exaggerating the effect of sedentary time on subjects' BMI and disseminating invalid health guidance to the population.

To investigate causal inference between lifestyle interventions and activity level while addressing the extreme values of the measurements from the wearable devices, in chapter three, we proposed a double robust estimator to extend the traditional Mann Whitney Wilcoxon Rank Sum Test (MWWRST) for causal inference in observational studies. The proposed estimator not only addresses the limitations of existing alternatives for more robust and reliable inference when applying the MWWRST to observational study data, but also performs well for small sample sizes. Meanwhile, The results from the real weight-loss trial showed that in addition to the doubly robust properties, the proposed estimator also effectively addressed outliers and extreme values.

Chapter 1

Fitbit Use and Activity Levels Two Years Post Interventions: Secondary Analysis of a Randomized Controlled Trial

1.1 Introduction

Technology-based activity trackers are increasingly used in research to promote behavior change. These trackers permit self-monitoring of many exercise related variables for wearers, such as sitting time, sit-stand transitions, steps, walking distance, moderate to vigorous physical activity (MVPA), sleep duration, and heart rate. Compared to traditional self-report surveys, self-monitoring activity trackers have several advantages, such as minimizing recall bias, automatic recordings of activity, providing much richer and denser data for inference[2]. Previous studies have shown that physical activity can decrease cancer recurrence, mortality, and improve quality of life, but many breast cancer survivors decrease their activity levels as much as 50% from pre- to post diagnosis and for several months to years following diagnosis [3]. A large proportion of those breast cancer survivors do not meet the activity guidelines of 150 minutes moderate-to-vigorous activity per week or the equivalent of 30 minutes of daily activity proposed by the World Health Organization [4].Activity trackers can be used to investigate how physical activities are accumulated for subjects in a daily basis, so that subjects can self-monitoring their daily activity level and change current behavior if necessary. Even when subjects meet physical activity guidelines, sitting for prolonged periods can still compromise metabolic health. Previous studies have shown that prolonged sedentary time, mostly consist of screen-based leisure activities (e.g., television watching), screen-based work activities (e.g., computer use for work purposes) and time spend on transportation, can cause negative effects on metabolic health, and that breaking up sedentary time, such as increase sit-stand transitions, can be beneficial [5]. Multiple studies carried by Sheri J Hartman, Andrea Z LaCroix, Loki Natarajan, etc mentioned that given many older adults spend the majority of their waking hours sitting and many other challenges for them to do MVPA, decreasing sitting time and increasing the number of sit-to-stand transitions can bring much health benefits [2, 6]. In studies examining the sedentary-mortality relationship among subjects with diabetes found that for every 60min/day increase in sedentary behavior, independent of moderate-to-vigorous physical activity (MVPA) and other covariates, subjects with diabetes had a 13% increased risk of all-cause mortality [7]. Recent studies investigating relationship between sedentary time and disease incidence, mortality, and hospitalization in adults also showed that prolonged sedentary time was independently associated with deleterious health outcomes regardless of physical activity [8].

Previous research on physical activity and health has concentrated largely on quantifying the amount of time spent in activities involving high levels of energy expenditure, such as moderate-vigorous-physical activity, while neglecting substantial contribution of sedentary behaviours and light physical activities to the overall energy expenditure[9, 10]. It is our contention that sedentary behavior is not simply the absence of moderate-to-vigorous physical activity, but rather is a unique set of behaviors, with unique environmental determinants and a range of potentially-unique health consequences. Sedentary behavior serves a distinct role for the population health, as it may influence obesity and other metabolic precursors of major chronic diseases (type 2 diabetes, cardiovascular disease, and breast and colon cancer). It is important to understand how emerging intervention modalities such as technology-based trackers can be used to help people increase their physical activity, and reduce sedentary behaviors.

For the first chapter of this dissertation, we used minute-level activity data collected from Fitbit trackers in a randomized controlled trial of breast cancer survivors to examine physical activity level and adherence of Fitbit use. We examined patterns of activity level and Fitbit use for both the 12-week intervention period and the 2 year follow-up period and compared patterns between intervention group and control group, by using Generalized Additive Mixed Model (GAMM) and Linear Mixed Effect Model.

1.2 Study cohort

Subjects in this study were enrolled in a randomized controlled trial (RCT) of a 12-week physical activity intervention. All subjects were female breast cancer survivors, in the age range of 21 to 85 years old, who were diagnosed less than 5 years before study enrollment and had completed chemotherapy or radiation treatment. The RCT has a total of 87 subjects, and we were focusing on the 75 subjects who consented to a 2-year follow-up. During the first 3-month intervention period, 37 subjects in the intervention group received intervention phone calls, at week 2, and week 6 and automatic emails every 3 days for data synchronizing throughout the 12-week intervention period. Starting right after the 12-week intervention, intervention subjects received a 2-year follow-up period. The control (or wellness) group has 38 subjects that were followed for 2 years. Subjects in the wellness group still received the intervention emails every 3 days and were offered the phone calls at week 2 and week 6, however, few people choose to complete the calls at the designated times.

Table 1.1: Baseline Characteristics for Participants

Chracteristics (N)	Exercise (N=37)	Waitlist (N=38)	All (N=75)
Age in years, mean(SD)	58.2(11.5)	56.2(9.1)	57.2(10.4)
Married status, n(%)	27(72.9)	27(71.1)	54(72.0)
BMI, kg/m^2, mean (SD)	26.7(6.4)	27.7(6.4)	27.2(6.4)
Education, n(%)			
Some college or less	11(29.7)	9(23.7)	20(26.7)
College graduate	15(40.6)	20(52.6)	35(46.7)
Master or higher	11(29.7)	9(23.7)	20(26.6)
Ethnicity, n(%)			
Not Hispanic/Latino	30(81.1)	33(86.8)	63(84.0)
Hispanic/Latino	7(18.9)	5(13.2)	12(16.0)
Race, n(%)			
White	30(80.1)	31(81.6)	61(81.3)
NonWhite	7(18.9)	7(18.4)	14(18.7)
Cancer stage, n(%)			
Stage 1	22(59.5)	22(57.9)	44(58.7)
Stage 2	11(29.7)	13(34.2)	24(32.0)
Stage 3	4(10.8)	3(7.9)	7(9.3)
Received chemotherapy, n(%)	21(56.7)	20(52.6)	41(54.7)
Time since surgery, months, mean(SD)	31.4(17.0)	30.6(16.0)	30.9(16.4)

1.3 Methods

1.3.1 Outcome measures

We used Fitbit-measured MVPA (moderate to vigorous physical activity) to measure subject's activity level [11, 12]. Daily MVPA is the total MVPA minutes in a day.

Daily adherence of wearing the Fitbit was defined as wearing the Fitbit for more than 10 hours or logging at least some activity (more than 1 min of MVPA). In order to study the trend of adherence for each subject over time, we transformed this binary daily adherence to a continuous percentage weekly rolling average adherence. The weekly rolling average adherence of wearing the fitbit was calculated by the percent of days in a rolling weekly period that the participant logged a valid day of wear (more than 10 hours of wear or more than 1 min MVPA).

1.3.2 Statistical analysis

Twelve patients who did not consent for 2-year follow-up study were excluded from the analysis and we compared patient characteristics between them and those who were included in the analysis. Patient baseline characteristics were also compared between the exercise and wellness groups. For exercise group, we specified knots at week 2, and week 6 to emphasize the time points when subjects receive intervention phone calls, and specified knot at week 12 to emphasize the end of the intervention period when using GAMM for trend comparison between the exercise and wellness group. Since wellness group didn't have the intervention period, we did not specify any knots for the first 3-month of the follow-up period in wellness group.

Mean weekly rolling average of adherence, mean daily MVPA and mean change from 3-month to follow-up period

For comparing the change in mean outcomes between two intervention groups at different periods, the model set up is

$$E[Y] = \beta_0 + \beta_1 * Group + \beta_2 * Period + \beta_3 * Group * Period + b_0 + b_1 * Period$$

where b_0 and b_1 are random effects, β_1 indicates the mean outcomes difference between the exercise group and the wellness group at first 3-month period, with wellness group serve as reference for group variable and first 3-month period serve as reference for period variable. $\beta_1+\beta_3$ indicates mean outcomes difference between the two groups at 2-year follow-up period. β_2 indicates mean outcomes difference between first 3-month and 2-year followup for the wellness group. $\beta_2 + \beta_3$ indicates mean outcomes difference between first 3-month and 2-year followup for the exercise group. We used t-test statistics from the regression to get the p-value for each combination of covariates. For example, for p-value of $\beta_1 + \beta_3$, the t-statistics would be $\frac{\hat{\beta}_1+\hat{\beta}_3}{\sqrt{var(\beta_1)+var(\beta_3)}+2\times cov(\beta1,\beta3)}$, the p-value would be $2 \times (1 - P(T \leq tscore))$, with df = number of observations - number parameters.

The p-values of group comparisons along with estimated mean and standard error of average adherence and MVPA using the LME was shown in table 3 in the results section.

Trend comparison of weekly rolling average of adherence and daily MVPA between two intervention groups for the overall study period, the initial 3-month period, and the 2-year follow-up period

A Generalized additive mixed effects model (GAMM) was used to compare trend of adherence and MVPA between the Exercise and Wellness group. The basic model set up is

$$g(y) = \beta_0 + \beta_1 * Group + s(Time) + s(Time) * Group,$$

in which β_0 is the fixed intercept, β_1 is the coefficient for 'Group', $s(Time)$ is the smooth term for 'Time', and $s(Time) * Group$ is the interaction term between 'Time' and 'Group'. Time is treated as a continuous variable.

For the goodness of fit of the chosen models, we used the model's deviance explained for both the MVPA and weekly rolling average of adherence, and the adjusted R-square to assess percentage of variances that are explained. We also used the minimized Generalized Cross Validation score (GCV) of the GAMM fitted for smoothness selection.

In general, the GAMMs without a specified number of knots will have relatively small smoothing parameters allowing for more curvatures. On the other hand, the GAMMs with specified knots will have relatively large smoothing parameters corresponding straight line estimates.

To select the best fitting model, in terms of the interaction term between time and group and knots specification in GAMM, we conducted model comparisons using analysis of variance (ANOVA) and model's Akaike Information Criteria (AIC). Graph of the best fit was used to display the trends of adherence and MVPA over the study period.

When the non-linear term (smooth term) in the GAMM was not significant, and graphical patterns of weekly rolling average adherence and MVPA showed linear trend, we used linear random effects model (LME) as a sensitivity analysis, to examine if the simpler LME and GAMM gave consistent results. The LME model set up is

$$E[Y] = \beta_0 + \beta_1 * Time + \beta_2 * Group + \beta_3 * Time * Group + b_0 + b_1 * Time,$$

in which b_0 is the random intercept for each subject, b_1 is the random slope for each subject.

Trend Comparison of weekly rolling average of adherence and daily MVPA between 3-month and 2-year follow up in the exercise group

To compare the weekly rolling average of adherence and daily MVPA within the exercise group between the 3-month active intervention period and the follow-up period, we used the linear random effects model with a random slope to estimate the slope of weekly rolling average of adherence and the slope of daily MVPA during the 3-month intervention and during 2-year follow up period for each individual. Then we used paired Wilcoxon rank sum test to compare the slope between 3-month and 2-year follow up period.

Since this is a randomized trial study, and the covariates between the two arms are balanced, thus the final models we use are unadjusted for potential covariates. We also examined the models with adjustment for potential covariates and there are no significant covariates, indicating no covariates were significantly different between the two arms at baseline.

1.4 Results

Patient characteristics

All the subjects in the original study were female breast cancer survivors, in the age between 21 and 85 years old, were diagnosed with less than 5 years before study enrollment, and had completed chemotherapy or radiation treatment. For 75 subjects who consented and were included in this analysis, their demographics were shown as table 1.1. As we can see from the table, there was no significant difference between the exercise and wellness groups.

We also did not observe any significant difference in baseline characteristics between subjects who were included in the study and those who were excluded from the study.

Mean weekly rolling average and mean MVPA

The mean weekly rolling average adherence and MVPA for both exercise and wellness group at different periods, and p-values for mean level group comparisons were shown as Table 1.2. As we can see from Table 1.2, for the overall and first 3-month period, the exercise group had higher average adherence and MVPA than the wellness group. However, during the 2-year follow up period, there was no significant difference in either average adherence or MVPA between the exercise and wellness groups. In addition, within the exercise and wellness group, the first 3-month period had higher mean adherence and MVPA than the 2-year follow up period. We also found that the change of mean adherence and mean MVPA from 3-month to 2-year follow-up period is significantly different between the two intervention groups, with more decrease in the exercise group.

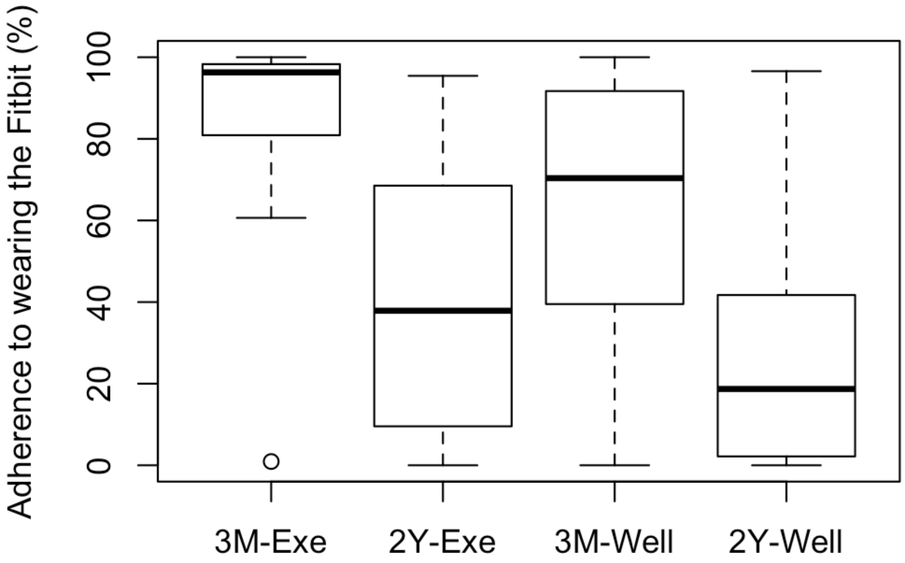
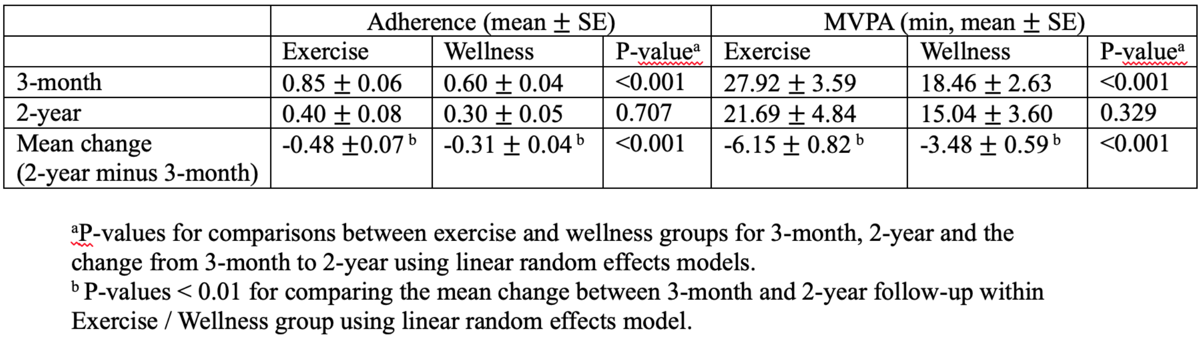


Figure 1.1: Average adherence

Table 1.2: Estimated Mean and SE for the average adherence and MVPA using LME



	Adherence (mean ± SE)			MVPA (min, mean ± SE)		
	Exercise	Wellness	P-value[a]	Exercise	Wellness	P-value[a]
3-month	0.85 ± 0.06	0.60 ± 0.04	<0.001	27.92 ± 3.59	18.46 ± 2.63	<0.001
2-year	0.40 ± 0.08	0.30 ± 0.05	0.707	21.69 ± 4.84	15.04 ± 3.60	0.329
Mean change (2-year minus 3-month)	-0.48 ±0.07 [b]	-0.31 ± 0.04 [b]	<0.001	-6.15 ± 0.82 [b]	-3.48 ± 0.59 [b]	<0.001

[a] P-values for comparisons between exercise and wellness groups for 3-month, 2-year and the change from 3-month to 2-year using linear random effects models.

[b] P-values < 0.01 for comparing the mean change between 3-month and 2-year follow-up within Exercise / Wellness group using linear random effects model.

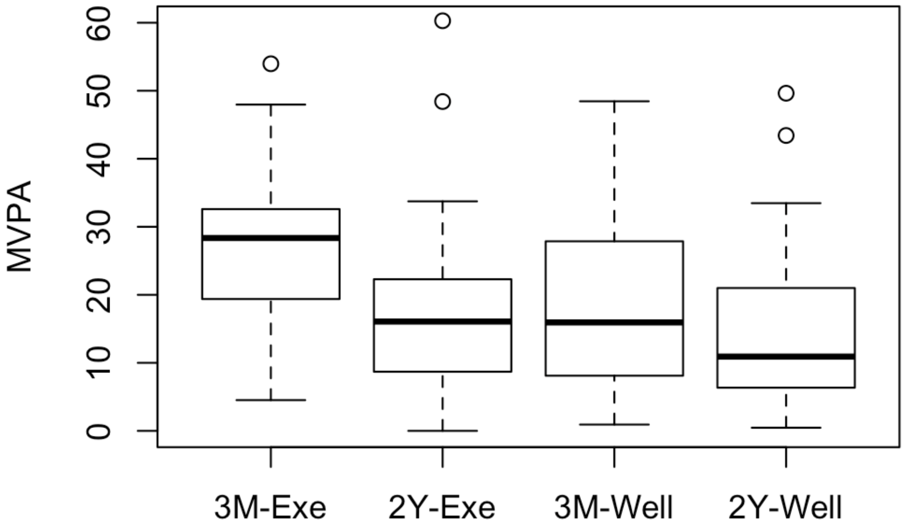


Figure 1.2: Average MVPA

Overall trend comparison of weekly rolling average of adherence and daily MVPA between the exercise and wellness groups

We examined GAMM models with different smoothing bases, such as penalized cubic regression spline (CR), cyclic penalized cubic regression (CC), and a shrinkage version of penalized cubic regression spline (CS), with and without prespecified change points. The change points were specified at week 2, week 6, and week 12 to emphasize intervention phone calls, at week 2, week 6, and the end of the study measurement visit at week 12. The GAMM with cubic regression basis and without specified time change points was selected as the best model for comparing the overall trend of adherence and daily MVPA between exercise and wellness groups, and has the smallest AIC compare to the other models we fitted. For the weekly rolling average of adherence, the p-value for the interaction between smooth 'time' term and 'group' was significant (p-value < 0.001), indicating the overall trend of adherence is significantly different between the exercise and wellness groups (Figure 1.4). For the daily MVPA, the overall trend was also significantly different between the exercise and wellness groups (p-value < 0.001, Figure 1.4). The curvature trend of the daily MVPA in the wellness group during the final months is mostly caused by a few subjects that maintained a relatively high activity level at the end of the study.

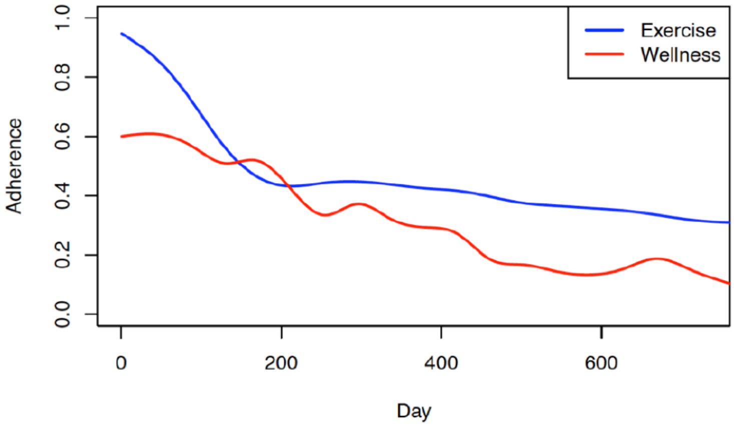


Figure 1.3: Overall adherence trend comparison between Exercise and Wellness group

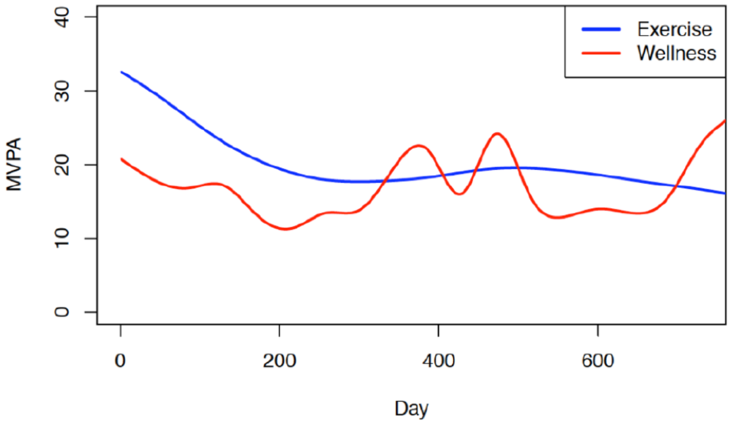


Figure 1.4: Overall MVPA trend comparison between Exercise and Wellness group

Comparison of weekly rolling average of adherence and daily MVPA during 3-month intervention between the exercise and wellness groups

Similar to the overall trend analysis, we examined GAMM models with different smoothing bases, with and without pre-specified change points. The GAMM with cubic regression basis was selected as the best model for comparing the 3-month trend of adherence and daily MVPA between the exercise and wellness groups. For exercise group, the GAMM was specified with knots at week 2, week 6, and week 12 to emphasize the time point subjects received intervention phone calls and the end of the measurement visit. For wellness group, no such knots were specified since few people choose to complete the calls at the designated times. For the weekly rolling average of adherence, the p-value of the smoothing term for time was not significant (p-value = 0.12), indicating a linear trend of adherence, and the p-value for the interaction between smooth 'time' term and 'group' was also not significant (p-value = 0.24, Figure 1.4), indicating that the trend of adherence was not significantly different between the exercise and wellness groups during this 3-month period. For the daily MVPA, we also found a linear trend for both group (p-value of the smoothing term = 0.15), and the trend of MVPA was also not significantly different between the exercise and wellness groups over this 3-month period (p-value = 0.99 for the interaction, Figure 1.4). Under the condition that smoothing terms in GAMM are not significant in trend of adherence and trend of MVPA analysis, the LME sensitivity analysis provided consistent results.

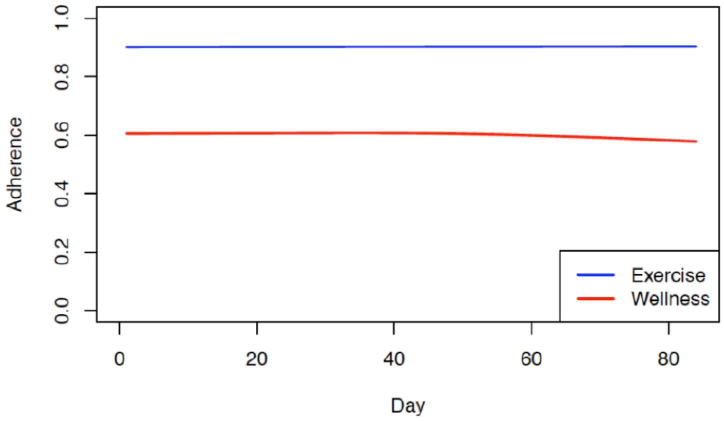


Figure 1.5: Comparison of adherence trend during the first 3-month between the Exercise and Wellness group

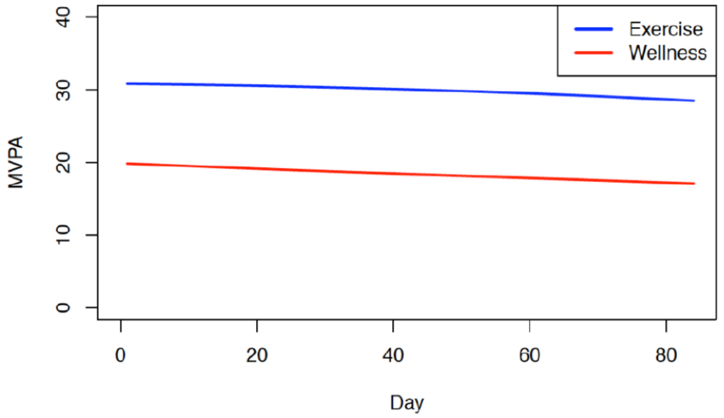


Figure 1.6: Comparison of MVPA trend during the first 3-month between the Exercise and Wellness group

Comparison of weekly rolling average of adherence and daily MVPA during the 2-year follow up between the exercise and wellness groups

We did not specify time change points in this analysis since there was no clinical intervention and no prior knowledge about the time change points during the 2-year follow-up period. The GAMM with cubic regression basis and without specified time change points was selected as the best model for assessing the trend of adherence and daily MVPA between the exercise and wellness groups. For the weekly rolling average of adherence, we found that the p-value of the smoothing term for time was significant (p-value < 0.001), indicating the trend of adherence is non-linear, and the trend of adherence was significantly different between the exercise and wellness groups during the 2-year follow up (p-value < 0.001, Figure 1.4). For the daily MVPA, we also found that the trend of MVPA was significantly non-linear(p-value < 0.001), and the trend of MVPA was significantly different between the exercise and wellness groups during the 2-year follow up (p-value < 0.001, Figure 1.4). We can see that the variation in the activity level is much higher than the variation in the adherence. Even at the end of the follow-up period, some subjects in the wellness group still maintained a relatively high activity level, causing the curvature in the wellness group.

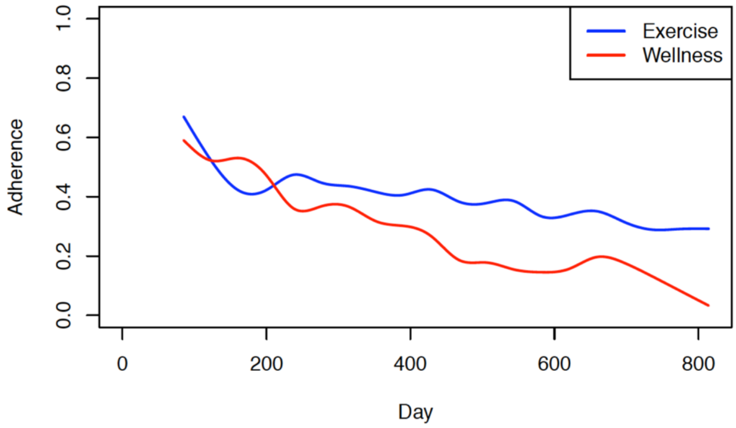


Figure 1.7: Comparison of adherence trend during the 2-year follow up between the Exercise and Wellness group

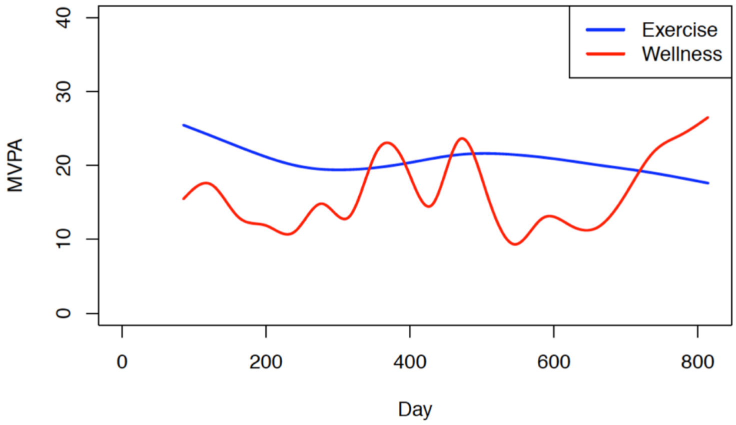


Figure 1.8: Comparison of MVPA trend during the 2-year follow up between the Exercise and Wellness group

1.5 Dicussion

We found that within the first 3-month intervention period, Exercise group has higher average of MVPA and adherence of fitbit use than the Wellness group, but the trend of MVPA and adherence of fitbit use are no difference between the two groups. Besides that, both the Exercise and Wellness group showed dropping trend of MVPA and adherence of fitbit use in the follow-up period, but Exercise group has a much slower dropping trend than the Wellness group. These insights may enhance our ability to effectively utilize activity trackers to promote behavior change.

1.6 Acknowledgement

The analysis results from Chapter 1 were partially published by Hartman, Sheri; Chen, Ruohui; Tam, Rowena; Narayan, Hari; Natarajan, Loki; and Liu, Lin. *Fitbit Use and Activity Levels From Intervention to 2 Years After: Secondary Analysis of a Randomized Controlled Trial.* JMIR Mhealth Uhealth. 2022 Jun 30;10(6):e37086. The dissertation author was one of the coauthors of the publication and the primary author of Chapter 1 of this dissertation.

Chapter 2

Measurement Error LME

2.1 Introduction

Current research suggests that a sedentary lifestyle can increase potential health risks, such as all-cause mortality rate, cancer risks, and risks for metabolic diseases [13]. Many health organizations have proposed scientific advisories on sedentary behavior to encourage people to exercise and minimize their sedentary time. The World Health Organization 2020 guidelines on physical activity and sedentary behavior recommend that adults engage in 75-300 minutes of moderate to vigorous physical activities (MVPA) weekly, and reduce sedentary time [14]. The American Heart Association also issued recommendations, indicating minimizing sedentary behavior can lower cardiovascular morbidity and mortality [15]. However, compared to physical activity recommendations, guidelines for sedentary behavior have been non-specific and do not indicate how much sedentary behavior is ”acceptable” nor do they quantify the amount by which sedentary behavior needs to be reduced in order to confer health benefits. In order to create specific guidelines, we have to address the measurement error inherent in current estimates of sedentary behavior. There are many sources for such measurement errors, such as inaccurate calibration of measurement instruments, inaccurate observations, and recording errors. Those measurement errors can be correlated with the true value of observations, the explanatory variables, and the response variables, making an intractable obstacle to obtaining accurate assessment of sedentary behavior, and to obtaining valid estimates of associations between these health behaviors and health outcomes [16, 17, 18].

The existence of measurement errors in sedentary behavior assessment arising from different sources poses serious challenges for conducting statistical analysis and obtaining unbiased estimates, especially without validation data [1]. Many large community-based studies use surrogate tools, such as self-report for activities during the day, which are subject to recall biases and contain both random and systemic errors. While a variety of technology-based trackers for measuring physical activity and sedentary behavior have emerged over the last few decades, very few 'gold standards' are established for evaluating the recorded measures. Failure to model the measurement errors appropriately can not only undermine the study design, but also lead to invalid conclusions, reduced statistical power, biased exposure-disease risk estimates, and misclassified risk groups [19, 20, 21, 22, 23].

Statistical methods for modeling measurement errors in dietary assessment have been intensively studied in the past decades [24, 25]. However, research to account for measurement errors in sensor-based recording activities, specifically sedentary behavior, is less studied, especially when there are multiple replicates of recordings for subjects in studies, as is typically the case for data from wearable sensors. The few published studies have focused on physical activity using regression calibration, and functional or Bayesian techniques to evaluate and correct for errors. For example, Ferrari et al (2007) [26], Nusser et al (2012) [27], and Beyler et al (2013) [28] proposed a multi-level equation-based modeling process to evaluate different types of physical activity measures and estimate the validity coefficients and attenuation factors, while accounting for measurement errors. Lim et al (2015) [29] proposed using regression calibration method to account for measurement errors in a self-reported physical activity survey in New York city. Agogo et al (2015) [30] and Jadhav et al (2022) [31] proposed a Bayesian-based method and a function-based method to model measurement error for physical activity data. Morrell et al. (2003) [32] proposed a method using the LME model on repeated measurements to obtain a predicted value to account for errors in the measurements. Our approach also exploits repeated measures but our focus is on analytic derivations to prove the unbiasedness of our proposed estimates. To our knowledge, no studies to date have evaluated measurement error for sedentary behavior derived from sensors, with a focus on providing statistical calculations, as well as, real data applications, of the process for correcting bias caused by measurement errors.

While theoretical approaches for measurement error correction could in principle be transported from one setting (e.g., diet) to another (e.g., sensor-based sedentary behavior), it is necessary to conduct a careful evaluation of the unique measurement properties of a given device and health behavior, in order to develop rigorous and domain-specific measurement correction tools. In this article, we conduct such an analysis of sedentary behavior derived from two commonly used wearable sensors (ActiGraph and activPAL). Leveraging the availability of multiple replicates measured from both devices for each subject, we propose to use structure models consisting of Linear Mixed Effect Models and Generalized Linear Models to obtain unbiased estimates of the relationship between exposures subject to measurement errors and outcome of interest, after appropriately accounting for the errors in devices' measurement. Section 2.2 introduces the sedentary behavior study that provides the motivation for the current work. Section 2.3 introduces the structure models to appropriately account for measurement errors and the proof of measurement error correction process. Section 2.4 implements Monte Carlo simulation to evaluate the proposed method. In Section 2.5 we apply our proposed method to the sedentary behavior study. Section 2.6 discusses the contributions and limitations of the proposed method.

2.2 Motivating Data Example

2.2.1 Study Cohort

Our proposed method and data application were motivated by the Adult Changes in Thought(ACT) study, which is an ongoing longitudinal cohort study of community-dwelling older adults that were greater than 65 years old and without evidence of Alzheimer's Disease or dementia in Washington State. The ACT study was initiated in 1994 to investigate risk factors for the development of dementia and has since provided a unique set of opportunities to additionally study a wide range of non-cognitive factors of healthy aging. Pertinent to the current study, the ACT activity monitor sub-study (ACT-AM) was initiated in 2016, adding a device-based activity component to capture the spectrum of sedentary and physical activity patterns [33]. Participants were instructed to wear a hip-worn triaxial ActiGraph ActiGraph (ActiGraph LLC, Pensacola, FL, USA) which captured 30Hz movement accelerations in three spatial axes, and a thigh-worn activPAL micro3 (PAL Technologies, Glasgow, Scotland, UK), which captured postures (e.g., sitting, standing, moving) [34]. Participants were instructed to wear both devices at the same time for 7 days. Participants also recorded self-reported sleep logs, in which participants recorded the time of waking up and going to bed, throughout their device wear. Ethics approval was obtained from the Kaiser Permanente Washington institutional review board. All participants provided written informed consent. There were total 980 subjects included in the data analysis; Table 2.1 provides demographic summary statistics of the study sample.

2.2.2 Sedentary Behavior Measures

Our data comprised concurrent wear of two devices, namely thigh-worn activPAL and hip-worn ActiGraph. The thigh-worn activPAL has been widely used in previous studies and is considered a gold standard for measuring postures, specifically sitting. Minutes spent in a sitting or lying posture during waking hours were summed over the day to provide a daily total sedentary time estimate for the activPAL [35]. For the ActiGraph monitor, sedentary behavior is commonly estimated using cut-points; we used the standard cut-point of <100 counts per minute for capturing participant's daily total sedentary time [36]. More than 95 % of subjects have at least 5 days of wearing both devices in the study, indicating the percentage of missing days of wearing devices for participants is very low.

We compared estimates of total daily sedentary time over the wearing period for each subject from these two devices via boxplots and summary statisics. From Fig 2.1 and Table 2.2, we can see that the daily average sitting time from activPAL is

Table 2.1: Descriptive Characteristics Of The Study Cohort

Characteristics	Subjects Included (n=980)
Age in years, mean(SD)	77.0(6.6)
Male,n(%)	428(44.7)
Race/Ethnicity, n(%)	
Hispanic or non-White	100(10.2)
non-Hispanic White	876(89.8)
Education, n(%)	
Less than high school	15(1.5)
Completed high school	79(8.1)
Some College	157(16.0)
Completed College	729(74.4)
BMI, n(%)	
Underweight ($\leq$ 18.5)	8(0.8)
Normal (18.5 - 24)	357(37.2)
Overweight (25 - 29)	378(39.4)
Obese ($\geq$ 30)	216(22.5)

slightly lower than the same recording from ActiGraph with higher variance. The longer sitting time measured by the ActiGraph could be caused by measurement error inherent when using cut-points to delineate (in)activity. Meanwhile, the Pearson correlation between measures from activPAL and measures from ActiGraph is around 0.64 with a p-value less than 0.001, indicating that there is a significant positive relationship between measures from the two devices.

Fig 2.2.2 is the scatter plot between measures from activPAL and measures from ActiGraph, the blue line is added as the reference for 'y=x', representing the perfect agreement between the two device estimates. Each dot represents a subject. We can see that the majority of the dots are above the blue reference line, indicating that most of the measures from activPAL are smaller than the measures from ActiGraph. Besides the scatter plot, we also used the Bland-Altman plot to

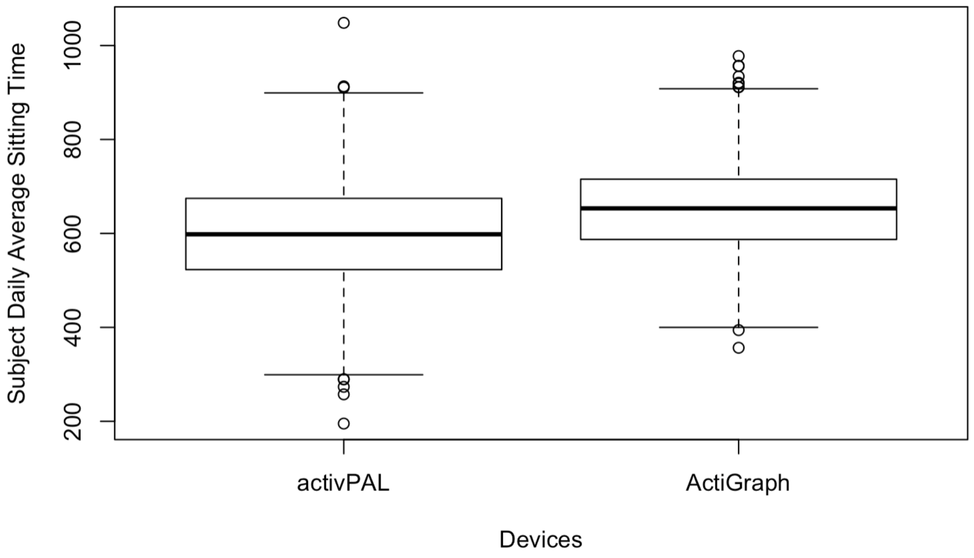


Figure 2.1: Boxplot of Daily Average Total Sedentary Time (mins) from activPAL and ActiGraph

Table 2.2: Daily Average Sedentary Time (mins) from activPAL and ActiGraph

Device (mins)	Min	Median	Mean(SD)	Max
activPAL	195.2	598.6	598.6(120.5)	1048.3
ActiGraph	356.6	655.9	654.9(97.6)	977.6

investigate the concordance between measures from the two devices [37]. In Fig 2.2.2, the blue dotted line indicates the mean difference measures between the two devices, and the light blue dashed line indicates the 95% confidence interval for the difference. From the Bland-Altman plot, we can see that the mean measures from activPAL are slightly smaller than the mean measures from ActiGraph, while many of the points are spread over the 95% confidence interval bands, indicating substantial variability in agreement between the two devices.

In summary, the plots and summary statistics clearly indicate that sedentary behavior estimates from the two devices are not identical. Importantly, even though the activPAL is considered to be more accurate for capturing sedentary behavior compared to the ActiGraph, measures from both devices likely contain measurement errors, which are due to inaccurate device calibration, inappropriate device wearing styles, and recording errors. Naïvely using the measures without appropriately accounting for the measurement errors can give us biased estimates and ultimately lead to invalid conclusions. We investigate these issues in the next sections, and propose using structure models to account for the measurement errors, with the ultimate goal of obtaining unbiased and consistent estimates.

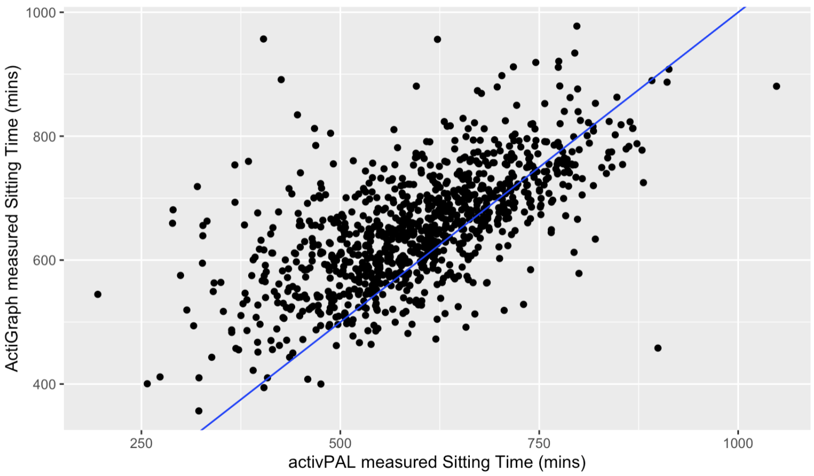


Figure 2.2: Total Sedentary Time estimates between activPAL and ActiGraph

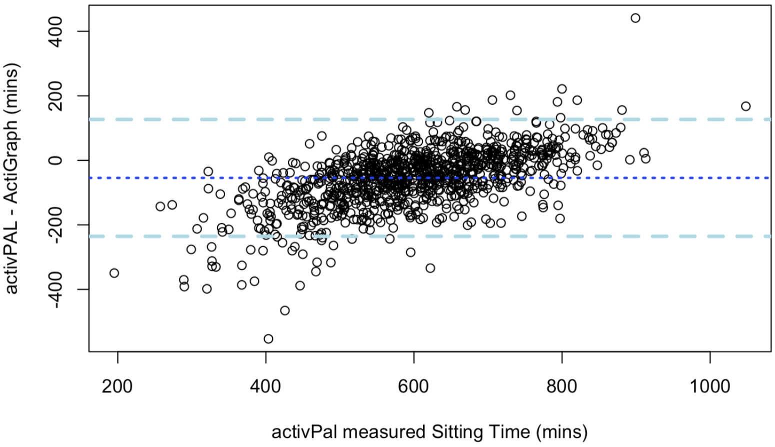


Figure 2.3: Total Sedentary Time Difference between activPAL and ActiGraph

2.3 Structure Models

2.3.1 Model Setup

We first establish terminology and the models. Let n denote the number of subjects, and assume every subject has the same number of replicates J. Consider our measurement error and outcome models as below:

$$W_{ij} = \sum_{s=1}^{m} \gamma_{0s} A_{is} + \gamma_1 X_i + U_{ij} \tag{2.1}$$

$$Y_i = \beta_0 + \beta_x X_i + \sum_{c=1}^{p} \beta_c C_{ic} + \epsilon_i \tag{2.2}$$

where W_{ij} denotes observed sedentary behavior measures for subject i from a device, with j represents repeated measures. A_{is} denotes a covariate that is without measurement error and informative for the measures of W_{ij} for subject i and there are a total of $m-1$ such covariates, with A_{i1} equals 1 for all subjects to include an intercept in the model. C_{ic} denotes a covariate without measurement error that is correlated with an outcome measure Y_i, p denotes the number of such covariates for subject i. Meanwhile X_i represents the centered true measures without measurement error with $X_i \sim N(0, \sigma_x^2)$, $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. We introduce some matrix notations next to simplify the analytic derivations to follow.

Let $\boldsymbol{X} = \Big(X_1, X_2, ..., X_n\Big)^T_{1\times n}$, $\boldsymbol{Y} = \Big(Y_1, Y_2, ..., Y_n\Big)^T_{1\times n}$, and $\boldsymbol{\epsilon} = \Big(\epsilon_1, \epsilon_2, ..., \epsilon_n\Big)^T_{1\times n}$.

Let $\boldsymbol{U_i}$ denote a vector of the errors U_{ij} for subject i and $\boldsymbol{U}$ denote a vector of $\boldsymbol{U_i}$ and, then $\boldsymbol{U_i} = \Big(U_{i1}, U_{i2}, ..., U_{iJ}\Big)^T_{1\times J} \sim N(0, \boldsymbol{\Sigma_u})$ and $\boldsymbol{U} = \Big(\boldsymbol{U_1}, \boldsymbol{U_2}, ..., \boldsymbol{U_n}\Big)^T_{1\times nJ} \sim N(0, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma_u} & & & & \\ & \boldsymbol{\Sigma_u} & & & \\ & & . & & \\ & & & . & \\ & & & & \boldsymbol{\Sigma_u} \end{bmatrix}_{nJ\times nJ}$$

with

$$\boldsymbol{\Sigma_u} = \begin{bmatrix} \sigma_u^2 & \rho\sigma_u^2 \cdots & \\ & \sigma_u^2 \cdots & \\ & & \sigma_u^2 \cdots \end{bmatrix}_{J \times J},$$

where ρ denotes the error correlation between replicates within each subject for the device. Note we are assuming error correlations are exchangeable which is reasonable in our case, since the measures are collected in close proximity, e.g., daily over 7 days.

Let $\boldsymbol{C}$ denote a matrix of C_{ic} so that $\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C_1} \\ \boldsymbol{C_2} \\ . \\ . \\ \boldsymbol{C_n} \end{pmatrix}_{n \times p}$, where $\boldsymbol{C_i} = (C_{i1}, ..., C_{ip})_{1 \times p}$ indicates a vector of covariates without measurement errors for subject i; and let $\boldsymbol{\beta_c}$ denote a vector of parameters β_c that is the coefficient for each covariate without measurement error $\boldsymbol{\beta_c} = \left(\beta_1, \beta_2, ..., \beta_p\right)^T_{1 \times p}$.

Let $\boldsymbol{A}$ denote a matrix that

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{\tilde{A}_{11}} \\ . \\ \boldsymbol{\tilde{A}_{1J}} \\ \cdots \\ \boldsymbol{\tilde{A}_{21}} \\ . \\ \boldsymbol{\tilde{A}_{2J}} \\ \cdots \\ \vdots \\ \cdots \\ \boldsymbol{\tilde{A}_{n1}} \\ . \\ \boldsymbol{\tilde{A}_{nJ}} \end{pmatrix}_{nJ \times m}$$

where $\boldsymbol{\tilde{A}_{ij}} = (A_{i1}, A_{i2}, ..., A_{im})$ with the first column being constant 1 to incorporate intercept. Let $\boldsymbol{\gamma_0}$ denote a vector

$$\boldsymbol{\gamma_0} = \Big(\gamma_{01}, \gamma_{02}, ..., \gamma_{0m}\Big)^T_{1\times m},$$

and $\boldsymbol{W}$ denote a vector of the replicates W_{ij} for subject i

$$\boldsymbol{W} = \Big(W_{11}, .., W_{1J}, W_{21}, .., W_{2J}, .., W_{n1}, .., W_{nJ}\Big)^T_{1\times nJ}$$

We also define $\boldsymbol{Z_i} = \boldsymbol{1}_{J\times 1}$, a vector of length J of 1s, and $\boldsymbol{Z}$ is a matrix of $\boldsymbol{Z_i}$

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z_1} & & & & \\ & \boldsymbol{Z_2} & & & \\ & & . & & \\ & & & . & \\ & & & & \boldsymbol{Z_n} \end{bmatrix}_{nJ\times n}$$

Then we can re-write the measurement error model (2.1) and outcome model (2.2) as

$$\boldsymbol{W} = \boldsymbol{A\gamma_0} + \boldsymbol{Z}\gamma_1\boldsymbol{X} + \boldsymbol{U} \tag{2.3}$$

$$\boldsymbol{Y} = \beta_0\boldsymbol{1}_{n\times 1} + \beta_x\boldsymbol{X} + \boldsymbol{C\beta_c} + \boldsymbol{\epsilon} \tag{2.4}$$

Therefore, $\gamma_1\boldsymbol{X} \sim N(0, \gamma_1^2\sigma_x^2\boldsymbol{I_n})$, where $\boldsymbol{I_n}$ is a $n \times n$ identity matrix. For notational brevity, let $\boldsymbol{G} = Var(\gamma_1\boldsymbol{X}) = \gamma_1^2\sigma_x^2\boldsymbol{I_n}$. Then, using standard linear mixed effects (LME) model theory [38], we can estimate $\boldsymbol{\gamma_0}$ in the measurement error model by $\boldsymbol{\hat{\gamma}_0} = \boldsymbol{(A^TV^{-1}A)^{-1}A^TV^{-1}W}$, where $\boldsymbol{V}$ represents the variance for $\boldsymbol{W}$ given $\boldsymbol{A}$ with $\boldsymbol{V} = \boldsymbol{ZGZ^T} + \boldsymbol{\Sigma}$. We can then estimate the Best Linear Unbiased Predictors (BLUP) from the measurement error model (2.3), $\widehat{\gamma_1\boldsymbol{X}} = \boldsymbol{GZ^TV^{-1}(W - A\hat{\gamma}_0)}$.

2.3.2 Measurement Error Correction and Unbiasedness

It is well-known that naïvely plugging in a replicate value, e.g., W_{ij} or the average across replicates as a substitute for the true measures will usually give biased estimates of β_x [20]. We will demonstrate this point for our application in a later section, here we prove that using the BLUP (from the LME model) in the outcome model gives us unbiased estimators for β_x when the outcome is continuous. In a subsequent section, we also evaluate the performance of the proposed method for both continuous and binary outcomes for different sample sizes through simulations.

Let $\boldsymbol{H} = \boldsymbol{I_{nJ}} - \boldsymbol{A(A^TV^{-1}A)^{-1}A^TV^{-1}}$, where $\boldsymbol{I_{nJ}}$ is a $nJ \times nJ$ identity matrix. Note that $\boldsymbol{H}$ is idempotent since $\boldsymbol{H^2} = \boldsymbol{H}$, and that $\boldsymbol{V^{-1}H} = \boldsymbol{H^TV^{-1}}$. We also have $\boldsymbol{HA\gamma_0} = \boldsymbol{0}$, where $\boldsymbol{0}$ is a $nJ \times 1$ matrix (vector), and therefore $\boldsymbol{HW} = \boldsymbol{H}(\boldsymbol{Z}\gamma_1\boldsymbol{X} + \boldsymbol{U})$. Therefore, the estimated $\widehat{\gamma_1\boldsymbol{X}} = \boldsymbol{GZ^TV^{-1}HW} = \boldsymbol{GZ^TV^{-1}H}(\boldsymbol{Z}\gamma_1\boldsymbol{X} + \boldsymbol{U})$.

The joint distribution of $\boldsymbol{Y}$ and $\gamma_1\boldsymbol{X}$ is multivariate normal, assuming independence between $\boldsymbol{X}$ and $\boldsymbol{C}$, we can write:

$$\begin{bmatrix} \boldsymbol{Y} \\ \widehat{\gamma_1\boldsymbol{X}} \end{bmatrix} = \begin{bmatrix} \beta_0\boldsymbol{1}_{n\times 1} & \beta_x\boldsymbol{I_n} & \boldsymbol{C} & \boldsymbol{I_n} & 0 \\ 0 & \boldsymbol{GZ^TV^{-1}HZ}\gamma_1 & 0 & 0 & \boldsymbol{GZ^TV^{-1}H} \end{bmatrix} \begin{bmatrix} 1 \\ \boldsymbol{X} \\ \boldsymbol{\beta_c} \\ \boldsymbol{\epsilon} \\ \boldsymbol{U} \end{bmatrix}$$

.

$$\begin{aligned} Var(\widehat{\gamma_1\boldsymbol{X}}) &= Var(\boldsymbol{GZ^TV^{-1}H)}(\boldsymbol{Z}\gamma_1\boldsymbol{X} + \boldsymbol{U})) \\ &= \boldsymbol{GZ^TV^{-1}H}[Var(\boldsymbol{Z}\gamma_1\boldsymbol{X} + \boldsymbol{U})]\boldsymbol{H^T(V^{-1})^TZG^T} \\ &= \boldsymbol{GZ^TV^{-1}HVH^TV^{-1}ZG^T} \\ &= \boldsymbol{GZ^TH^TV^{-1}ZG^T} \end{aligned}$$

$$
\begin{aligned}
Cov(\boldsymbol{Y},\widehat{\gamma_1\boldsymbol{X}}) &= Cov(\beta_0\mathbf{1}_{n\times1}+\beta_x\boldsymbol{X}+\boldsymbol{C}\boldsymbol{\beta_c}+\boldsymbol{\epsilon},\boldsymbol{GZ^TV^{-1}H}(\boldsymbol{Z}\gamma_1\boldsymbol{X}+\boldsymbol{U}))\\
&= Cov(\beta_x\boldsymbol{X},\boldsymbol{GZ^TV^{-1}HZ}\gamma_1\boldsymbol{X})\\
&= \beta_x Cov(\boldsymbol{X},\gamma_1\boldsymbol{X})\boldsymbol{(GZ^TV^{-1}HZ)^T}\\
&= \beta_x\gamma_1 Var(\boldsymbol{X})\boldsymbol{Z^TH^T(V^{-1})^TZG^T}\\
&= \frac{\beta_x}{\gamma_1}Var(\gamma_1\boldsymbol{X})\boldsymbol{Z^TH^TV^{-1}ZG^T}\\
&= \frac{\beta_x}{\gamma_1}\boldsymbol{GZ^TH^TV^{-1}ZG^T}
\end{aligned}
$$

Then by the properties of multivariate normal distribution, we have:

$$
\begin{aligned}
E(\boldsymbol{Y}|\widehat{\gamma_1\boldsymbol{X}},\boldsymbol{C}) &= E(\boldsymbol{Y}|\boldsymbol{C})+Cov(\boldsymbol{Y},\widehat{\gamma_1\boldsymbol{X}})[Var(\widehat{\gamma_1\boldsymbol{X}})]^{-1}(\widehat{\gamma_1\boldsymbol{X}}-E(\widehat{\gamma_1\boldsymbol{X}}))\\
&= \beta_0\mathbf{1}_{n\times1}+\boldsymbol{C}\boldsymbol{\beta_c}+\frac{\beta_x}{\gamma_1}\widehat{\gamma_1\boldsymbol{X}}\\
&= (\mathbf{1}_{n\times1},\widehat{\boldsymbol{X}},\boldsymbol{C})\begin{pmatrix}\beta_0\\ \beta_x\\ \boldsymbol{\beta_c}\end{pmatrix}
\end{aligned}
$$

Let $\boldsymbol{D}$ denote the design matrix, which is $\boldsymbol{D}=(\mathbf{1}_{n\times1},\widehat{\boldsymbol{X}},\boldsymbol{C})$. By the ordinary least square estimation, we have:

$$
\widehat{\boldsymbol{\beta}}=\begin{pmatrix}\beta_0\\ \beta_x\\ \boldsymbol{\beta_c}\end{pmatrix}=\boldsymbol{(D^TD)^{-1}D^TY}
$$

Then we show that $\boldsymbol{(D^TD)^{-1}D^TY}$ is an unbiased estimator for β_0, β_x and $\boldsymbol{\beta_c}$:

$$
\begin{aligned}
E(\widehat{\boldsymbol{\beta}}) &= \boldsymbol{(D^TD)^{-1}D^T}E(\boldsymbol{Y}|\widehat{\gamma_1\boldsymbol{X}},\boldsymbol{C})\\
&= \boldsymbol{(D^TD)^{-1}D^TD}\widehat{\boldsymbol{\beta}}\\
&= \widehat{\boldsymbol{\beta}}
\end{aligned}
$$

Therefore using the BLUP of $\gamma_1\boldsymbol{X}$, instead of naïvely observed measures which contain errors, gives us unbiased estimates for parameters in the outcome model. However, the above proof holds if and only if $\boldsymbol{C}$ is independent of $\boldsymbol{X}$, namely that the error-free covariates are independent of the true measures for each subject. We explored situations when $\boldsymbol{C}$ and $\boldsymbol{X}$ are correlated for both continuous and binary outcomes $\boldsymbol{Y}$ through simulations in the next section.

2.4 Simulation

2.4.1 Simulation Model Setup

Recapitulating the earlier notation, let Y_i represent a clinical outcome of interest for subject i, C_i denote a covariate measured without error, and W_{ij} represent observed sedentary behavior measures from a device, where j indexes replicate measures (e.g., $j = 1, 2, \cdots 7$ if the device is worn daily for 1 week). X_i represents the true centered average sedentary time for subject i over the measurement period (e.g., 1 week). For demonstration purposes, we consider a regression model of the clinical outcome Y_i on covariate X_i through a link function $f(y)$. We specify the following measurement error model for the covariate measured with error,

$$W_{ij} = \gamma_0 + \gamma_1 X_i + U_{ij}; \tag{2.5}$$

and we specify an outcome model for the centered true average measure given the covariate C_i:

$$f(Y_i) = \beta_0 + \beta_x X_i + \beta_c C_i + \epsilon_i. \tag{2.6}$$

The equations (2.5) and (2.6) are simplified cases of (2.1) and (2.2) with no covariates A_i and a single covariate C_i when we consider a linear regression of a continuous outcome Y_i. The goal of the analysis is to estimate β_x, the effect of covariate X with measurement error.

Let $\boldsymbol{W_i} = (W_{i1}, W_{i2}, \cdots, W_{iJ})^T$ and $\boldsymbol{U_i} = (U_{i1}, U_{i2}, \cdots, U_{iJ})^T$ be vectors representing the J replicates for subject i, we assume

$$\begin{aligned} \boldsymbol{U_i} &\sim N(0, \boldsymbol{\Sigma_u}), \quad \epsilon_i \sim N(0, \sigma_\epsilon^2), \\ \boldsymbol{W_i} \mid X_i &\sim N(\gamma_0 + \gamma_1 X_i, \boldsymbol{\Omega}), \\ \boldsymbol{\Omega} &= \gamma_1^2 \sigma_x^2 \boldsymbol{I}_{J\times J} + \boldsymbol{\Sigma_u}, \quad \boldsymbol{\Sigma_u} = \begin{pmatrix} \sigma_u^2 & \rho\sigma_u^2 & \cdots & \\ & \sigma_u^2 & \cdots & \\ & & \sigma_u^2 & \cdots \end{pmatrix}_{J\times J}. \end{aligned}$$

X and C will be generated as follows:

$$\begin{pmatrix} X \\ C \end{pmatrix} = N(\begin{pmatrix} \mu_x \\ \mu_c \end{pmatrix}, \Sigma_{xc}), \quad \Sigma_{xc} = \begin{pmatrix} \sigma_x^2, & \rho_{xc}\sigma_x\sigma_c \\ \rho_{xc}\sigma_x\sigma_c, & \sigma_c^2 \end{pmatrix}.$$

In the above model set-up, ρ denotes the error correlations between replicates within each subject for the device, ρ_{xc} represents the correlation between the latent variable X and variable without measurement error C, $N()$ indicates the normal distribution and $\boldsymbol{I}_{J\times J}$ is a $J \times J$ identity matrix. We also note that for simplicity we assume that the number of replicates J is the same across subjects; our methods will be easily generalized to the unbalanced case.

We conducted a series of simulations to examine the performance of the proposed method for accounting for measurement error in continuous and binary outcome models, and compared the proposed method to the naïve method that uses the error-prone measures without adjusting measurement errors. All simulations were performed with a Monte Carlo sample of 1000. We examined the performance of these methods in estimating β_0, β_x and β_c. The performance metrics include mean estimates of the coefficient, estimated asymptotic and empirical standard errors, relative bias and coverage probability of 95% confidence interval . To mimic the dataset in the ACT-AM study, all parameters in the simulation are set based on prior analysis. Using data structures as in (2.5) and (2.6) , we showed the simulation results for sample sizes $n = 50, 100, 500$ for continuous outcome, and sample sizes $n = 100, 200, 500$ for binary outcomes; each subject had J=7 replicates of measures.

We also conducted simulations with missing completely at random for replicates of each subject, and the results are similar to using complete data.

2.4.2 Continuous Clinical Outcome

We considered a linear regression for a continuous outcome Y_i with $f(y) = y$, then

$$Y_i|(X_i, C_i) \sim N(\beta_0 + \beta_x X_i + \beta_c C_i, \sigma_\epsilon^2).$$

Parameters and generated measures were as follows:

$$\begin{aligned}
&\beta_0 = 10, \quad \beta_x = 2.95, \quad \beta_c = 3, \\
&\gamma_0 = 1, \quad \gamma_1 = 1(W_1) \quad or \quad 2(W_2), \quad \mu_c = 1, \quad \sigma_c = 1, \\
&\mu_x = 0, \quad \sigma_x = 2, \quad \sigma_\epsilon = 1, \quad \sigma_u = 1, \\
&\ \rho = 0.1 \quad \text{or} \quad 0.3, \quad \rho_{xc} = 0 \quad \text{or} \quad 0.5
\end{aligned}$$

For the correlation between the latent variable X and variable without measurement error C, we explored the cases when $\rho_{xc} = 0$ and $\rho_{xc} = 0.5$, meanwhile varied the value of attenuation bias γ_1, where W_1 represents an unbiased case with measurements with $\gamma_1 = 1$, and W_2 represents measurements with $\gamma_1 = 2$.

2.4.3 Binary Clinical Outcome

We considered a logistic regression for a binary outcome

$$Y_i|(X_i, C_i) \sim Bernoulli(Pr), \quad Pr = \frac{1}{1 + exp(-(\beta_0 + \beta_x X_i + \beta_c C_i))}.$$

Parameters in the outcome model (β) were set up as follows, other parameters were set up the same as in continuous case:

$$\beta_0 = 0.1, \quad \beta_x = 0.1, \quad \beta_c = 0.1$$

2.4.4 Simulation Results

We compared the performance of the method naïvely using the measures containing errors (denoted as W_1 and W_2 assuming the attenuated bias γ_1 =1 and 2) and the proposed approach using BLUP under 1000 Monte Carlo samples. The $BLUP_1$ and $BLUP_2$ represent using the Best Linear Unbiased Prediction (BLUP) from W_1 and W_2, respectively. Shown in the appendix Table A.1 and Table A.2 are the simulation results for the estimated β_0, β_x, and β_c and standard errors of these estimates for the continuous and binary outcome models, respectively. For both continuous and binary outcomes, the standard errors of the estimates (i.e., the asymptotic standard errors) from the proposed approach using BLUP were very similar to their empirical standard errors, and as expected, these standard errors decreased as the sample size increased. In Fig 2.4. (Appendix Table A.3 and Table A.4), we summarized the relative bias of estimated β_x under different sample sizes while comparing using $BLUP_1$ and $BLUP_2$ to naively using W1 and W2 for both continuous and binary outcomes with different magnitudes of ρ and ρ_{xc}. As we can see from Fig 2.4., the proposed BLUP method has a much smaller relative bias than naively using measures containing errors for both continuous and binary outcomes under different correlations.

When the latent truth X is independent of the covariate C ($\rho_{xc} = 0$), the estimates of β_0, β_x and β_c from the proposed BLUP approach were close to the truth for both the continuous and binary outcome cases, even for relatively small sample size. In contrast, naïvely using measures containing measurement errors gave us biased estimates for β_0 and β_x, and the bias increases as the correlation for replicates within subject ρ and the magnitude of attenuation bias γ_1 increases. The relative bias was given in Fig 2.4. For example, in the continuous outcome case, when $\rho = 0.1$ and $\rho_{xc} = 0$, the relative bias of estimated β_x is around 0.8% for rather small sample size $n = 50$ using $BLUP_1$ and $BLUP_2$, but is around 3.8% if naïvely using W_1 and even worse around 50% if naïvely using W_2, which has greater attenuation bias than W_1 (Fig 2.4.a). In the binary outcome case, a similar trend follows but with a slightly increased relative bias for all methods in comparison; taking the $\rho = 0.1$ and $\rho_{xc} = 0$ for binary outcome case for example, the relative

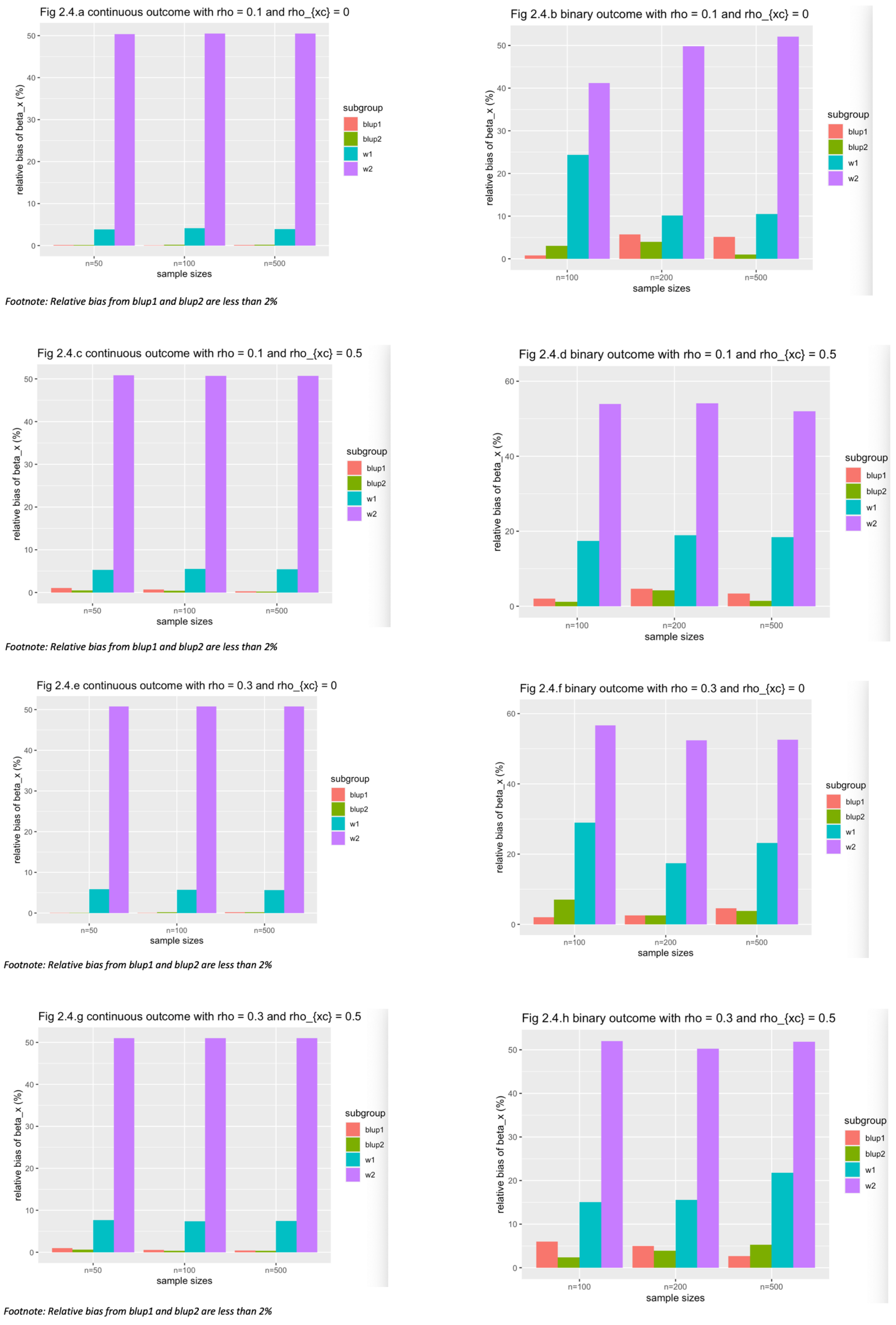


Fig 2.4.a continuous outcome with rho = 0.1 and rho_{xc} = 0

Footnote: Relative bias from blup1 and blup2 are less than 2%

Fig 2.4.b binary outcome with rho = 0.1 and rho_{xc} = 0

Fig 2.4.c continuous outcome with rho = 0.1 and rho_{xc} = 0.5

Footnote: Relative bias from blup1 and blup2 are less than 2%

Fig 2.4.d binary outcome with rho = 0.1 and rho_{xc} = 0.5

Fig 2.4.e continuous outcome with rho = 0.3 and rho_{xc} = 0

Footnote: Relative bias from blup1 and blup2 are less than 2%

Fig 2.4.f binary outcome with rho = 0.3 and rho_{xc} = 0

Fig 2.4.g continuous outcome with rho = 0.3 and rho_{xc} = 0.5

Footnote: Relative bias from blup1 and blup2 are less than 2%

Fig 2.4.h binary outcome with rho = 0.3 and rho_{xc} = 0.5

Figure 2.4: Relative bias of estimated β_x using different methods for continuous outcome and binary outcome

bias of estimated β_x is around 0.8% and 3.0% for n=100 when using $BLUP_1$ and $BLUP_2$, in comparison to a much higher 24.4% and 41.2% when naïvely using W_1 and W_2 (Fig 2.4.b). Of note, in our simulations, we set $\gamma_1 = 2$ for W_2, which resulted in attenuation of β_x when naïve plugging in W_2. A different choice of γ_1 (e.g., $\gamma < 1$) would result in inflation of β_x, namely erroneous exaggerated effects of the exposure-disease associations.

When X is correlated with C ($\rho_{xc} = 0.5$), the proposed structure models using BLUP were still able to have much less biased estimates of β_0 and β_x than the naïve plug-in approach. For example, in continuous outcome case, when $\rho = 0.1$ and $\rho_{xc} = 0.5$, the relative bias of estimated β_x is around 1.0% and 0.5% when using $BLUP_1$ and $BLUP_2$, which are much lower than 5.3% and 50.9% the relative bias of estimated β_x for n=50 when naïvely using W_1 and W_2 (Fig 2.4.c). A similar trend was also observed in binary outcome cases with slightly increased relative bias.

Besides that, we have also examined the 95% CI coverage probability for using $BLUP_1$, $BLUP_2$, W_1, and W_2, and found that for continuous outcome the coverage of the proposed method using BLUP is much better than using the naïve plug-in approach (Appendix Fig A.1). On the other hand, for binary outcome, due to the wide confidence interval of the estimates, coverage probability is similar among all approaches, but as attenuation bias γ_1 contained in the measurements increases, the coverage drops for the naïve plug-in approach, while the BLUP still maintained good coverage adjusting the measurement errors.

2.5 Data Analysis

In sedentary behavior studies, research generally focuses on whether an individual's activity level or sitting time has an impact on their health. We implemented our proposed method on data from the ACT Study, to investigate the impacts of increasing subjects' daily total sitting time on BMI and obesity status (i.e., BMI is equal to or greater than 30 kg/m^2). We compared our proposed method to the naïve estimate using the subject level average of the repeated measures from activPAL and ActiGraph without measurement error correction. Participants were instructed to wear both devices at the same time for 7 days, and analysis results using BLUP from fitting LME using the activPAL measures and the ActiGraph measures are quite close.

Table 2.3: Estimated β_x in Outcome Model (BMI/Obesity) for Total Sedentary Time.

Outcome	BLUP$_{activPAL}$		BLUP$_{ActiGraph}$		activPAL		ActiGraph	
	$\hat{\beta}_x$ (sd)	p-value	$\hat{\beta}_x$ (sd)	p-value	$\hat{\beta}_x$ (sd)	p-value	$\hat{\beta}_x$ (sd)	p-value
BMI	0.80(0.09)	<0.001	0.77(0.10)	<0.001	0.55 (0.10)	<0.001	0.36 (0.11)	<0.001
Obesity (y/n)	0.41 (0.06)	<0.001	0.40 (0.06)	<0.001	0.28 (0.05)	<0.001	0.23 (0.06)	<0.001

Shown in Table 2.3 are the estimates of β_x in equation (2.2) with associated standard error and p-values while controlling other variables that don't contain errors, such as age, gender, education level, and ethnicity. $BLUP_{activPAL}$ and $BLUP_{ActiGraph}$ indicate the proposed method using BLUP from devices activPAL and ActiGraph, while activPAL and ActiGraph indicate naïvely using measures contained errors from each device.

In Table 2.3, even though for both methods, the standard errors are small, and p-values < 0.001, the estimated $\hat{\beta}_x$ in the outcome (BMI) model for total sedentary time differ substantially between the two approaches. For the proposed approach, for a 1-hour increase in the total sedentary time,average BMI is expected to increase by 0.80, while holding other covariates constant. However, naïvely using the activPAL and ActiGraph measures that contained errors gave us a falsely underestimated effect. From the naïve use of the activPAL measures, with 1-hour increase in the total sedentary time, average BMI is expected to increase by 0.55, which is about seventy percent of the estimated effect using BLUP. Meanwhile, from the naïve plug-in approach by the ActiGraph measures, with 1-hour increase in the total sedentary time, average BMI is expected to increase by 0.36, which is less than half of the estimated effect after accounting for measurement errors. Therefore without accounting for errors in the measurements, we may inappropriately underestimate the effect of sedentary time on subjects' BMI and disseminate invalid health guidance to the population.

For the binary obesity status case, we see a similar trend as in continuous BMI. For a 1-hour increase in the total sedentary time, the odds of being obese will increase by around 50% (exp(0.41) - 1, exp(0.40) -1) using the proposed approach to account for the measurement errors, however, the odds of being obese will increase by around 32% (exp(0.28) - 1) from naïvely using activPAL measures and 26% (exp(0.23) - 1) from naïvely using ActiGraph measures. This falsely underestimated effect is also likely due to measurement errors, which can cause biased estimates and ultimately lead to false conclusions. Of note, both devices are also subject to random errors, and as indicated by our simulations the BLUP method may correct for both random and systematic biases in either device. Although we do not know the true β_x for the data application, the results are consistent with our simulations where we observed shrunk estimates from the error-prone W_1 and W_2, while the BLUP method was able to adjust the measurement errors and give unbiased estimates.

2.6 Discussion

With the increasing use of wearable devices based on different technologies, there is growing interest in ascertaining how to use these rich data sources to obtain unbiased risk estimates and draw valid conclusions in health behavior studies. Although wearable sensors are likely less error-prone than self-report, however, they are subject to errors, e.g., device malfunction, or calibration issues. Notably, due to different correlation structures between measurements and errors, and other correlated variables in the model, the direction of impact on risk estimates of naïvely using measures containing errors is difficult to model or predict. Therefore, instead of estimating the impact of the measurement error, previous studies had tried to take advantage of the replicated measures to reduce the bias for estimates of the relationship between exposures subject to measurement errors and outcome of interest. For example, Rosner and Polk (1983) proposed that the average of repeated measures of subjects' blood pressure within a relatively short period of time tends to be close to their true blood pressure level. Averaging the replicates of measures may be a good idea for measurement errors that are random. However, when there are systemic correlated errors in the measurements, which can be caused by inaccurate instruments., simply using the average of the replicates may still contain a significant amount of measurement errors. Rosner et al. (1989) proposed two methods, the linear approximation method and the likelihood approximation method, to reduce bias caused by either random or systematic measurement errors for estimates of relative risk. Nonetheless, these methods require a separate validation study to be applicable. Morrel et al (2003) proposed a method using the mixed effects estimates to get a baseline measure closer to the truth than using a single measurement. However,they did not elucidate the theoretical advantages of using the BLUP, which when used correctly can be more robust to bias caused by measurement errors as demonstrated in this paper.

We developed a version of structure models by combining the Linear Mixed Effect Models and Generalized Linear Models to account for measurement errors, so that we can obtain unbiased estimations of parameters of interest. We achieved this by taking advantage of multiple replicates available from daily device wear, and proposed using the BLUP instead of naïvely using measures directly provided by the devices. We showed that using BLUP will give unbiased estimates for the conditional associations between the true exposure, i.e. sedentary time in our application, and clinical outcomes, when the true exposure is independent of other covariates that are measured without errors. Through intensive simulations for both continuous and binary outcomes, we demonstrated that the proposed method performed very well and achieved accurate parameter estimates, in scenarios with more general correlation structures and even for relatively small sample sizes. We also applied the proposed approach to an existing study and compared the results with naïve plug-in approach to demonstrate discrepancies between uncorrected estimates based on device outputs versus the error-corrected BLUP approach proposed herein.

Focusing on sedentary behavior, we compared two commonly used devices for sedentary time assessment, the activPAL and the ActiGraph. Our data analysis

(Table 2.3) indicates less attenuation of regression coefficients for the activPAL versus ActiGraph compared to the proposed BLUP method. While we do not know the "truth" in this data analysis application, our results are in line with health behavior research: the activPAL is a validated tool for posture classification (e.g., sitting, standing, stepping), and is believed to be less biased for sedentary behavior estimation than the ActiGraph, which uses thresholds based on energy expenditure to classify sedentary behavior and thus is prone to systematic biases [39]. Of note, both devices are also subject to random errors, and as indicated by our simulations the BLUP method may correct for both random and systematic biases.

While the proposed LME-based structure models can correct the measurement errors in the exposure, it is important to note that multiple replicates are needed for the proposed method to be applicable. Since per best practices, health behavior researchers already require and collect multiple repeated days of device wear, this potential drawback, can be accommodated for sedentary behavior research, which is the focus of our application. Importantly, the LME structures can be straightforwardly implemented in different settings through standard statistical software, such as R and SAS, and generalized to other behaviors such as physical activity or sleep research.

While our approach sets a rigorous foundation, there is undoubtedly scope to expand and improve our methods. Although our data analysis study cohort was relatively complete, for future work, we are interested in expanding this work to accommodate different missingness mechanisms, such as MAR and MNAR. Meanwhile, currently we treat measures within a short period of time (7 days) as replicates of each other, we also aim to extend the current setting to longitudinal data with different cluster sizes.

2.7 Acknowledgement

Chapter 2, in full, has been submitted for publication as Chen, Ruohui; Rosenberg, Dori; Di, Chongzhi; Zablocki, Rong; Hartman, Sheri; LaCroix, Andrea; Tu,

Xin; Natarajan, Loki and Liu, Lin. *A Linear Mixed Model Approach for Measurement Error Adjustment: Applications to Sedentary Behavior Assessment from Wearable Devices*, submitted to Annals of Applied Statistics. The dissertation author was the primary author on this paper.

Chapter 3

A Double Robust Estimator for Mann Whitney Wilcoxon Rank Sum Test When Applied for Causal Inference in Observational Studies

3.1 Introduction

In randomized control trials, the non-parametric Mann-Whitney-Wilcoxon rank sum test (MWWRST) is widely used as an alternative to the two-sample t-test when data distributions are highly skewed, especially with outliers. For non-randomized studies, this rank-based test generally yields invalid results for causal inference. Although one may remove or winsorize outliers and apply mean-based methods such as regression, propensity score matching, marginal structure models, results from such analyses are difficult to interpret as they are subjective and depend on how the outliers are handled.

Shown in Table 3.1 are sample means and standard deviations of subjects' activity levels at the end of intervention, along with maximum (Max), interquartile

Table 3.1: Two sample t-test vs. MWWRST for group difference of Weighted Sum Activity Count of MVPA.

Two sample t-test			
	Intervention mean(SD)	Control mean(SD)	p-value(t)
AC MVPA	3595378 (1652517)	3231656 (1634587)	0.091
Max/IQR	5.53	5.80	
MWWRST			
	Estimate(SE)		p-value
δ	0.430(0.035)		0.039

range (IQR) and p-value from the two-sample t-test, to assess intervention effects for breast cancer survivors in a randomized control weight-loss study (see Section 3.5.2 for study details). There are clearly outliers in both groups based on the common winsorizing approach, e.g., 3 times of IQR criterion [40]. While the t-test fails to capture any significant difference, the MWWRST shows a significant difference in mean rank between the two groups. The outliers in this study have a significant impact on study findings with important implications for clinical practice.

For non-randomized observational studies, Wu et al.(2014) [41] introduced an approach for causal inference by incorporating inverse probability weighting (IPW) into the MWWRST. Their approach addressed limitations of an earlier attempt by Rosenbaum (2002) [42], in which a constant individual treatment effect $\tau = y_1 - y_0$ between two potential outcomes (y_1, y_0) is imposed in order to use a randomization technique for inference. This assumption is not only implausible, but also unverifiable in practice. Mao et al. (2018) [43] and Zhang et al. (2019) [44] extended the IPW approach of Wu et al. (2014) to develop doubly robust estimators for more robust inference. However, their doubly robust estimators have major limitations.

In Mao (2018), parametric logistic and linear regression were used for the propensity score of the IPW and outcome model of the augmented component of the doubly robust estimator. Since the logistic regression cannot address overdispersion [45, 46], it may not be correct for some real studies. The parametric linear regression is even more problematic, since it not only fails to address outliers, which calls for the MWWRST in the first place, but also limits its applications to normal data. Thus, both parametric models, especially the second one, can be wrong for most real study data and as such this doubly robust estimator may even be less robust than the IPW estimator in Wu et al. (2014), which uses a semiparametric generalized linear model with the logit link, or restricted moment [47], for the propensity score.

Zhang et al. (2019) presented two doubly robust estimators. One estimator also posits a parametric outcome model, which like Mao's approach does not address outliers, since no reliable estimator of this outcome model can be obtained in the presence of outliers. The second estimator addresses the limitation of Mao's parametric outcome model by using a semiparametric regression for between-subject attributes, or generalized probability index model (GPI) as so termed in the paper, since it generalizes the probability index model (PI) developed by Thas et al. (2012) [48]. A major limitation of their approach is the reliance on bootstrap for inference. First, applications of bootstrap within the current context are highly inefficient, since the MWWRST statistic based on a sample size n requires computational times in the order of $O\left(n^2\right)$, where $n = n_1 + n_2$ and n_k denotes the sample size of group k $(= 0, 1)$. For example, in a similar application involving modeling between-subject attributes for microbiome beta diversity outcomes [49],the run time for asymptotic inference is about 35 times less compared to a permutation-based approach with 1,000 permutations. For the simulation set up in their paper, we find that run time for the their bootstrap inference is generally around 40 times the run time for the proposed doubly robust estimator in this paper with asymptotic inference. Second, the bootstrap procedure for their doubly robust estimators has unknown performance even for large samples, since it is being applied to quite a complex U-statistics setting with both estimated model parameters for the propensity score and outcome regression models. As no investigation of large sample properties was conducted, applications of their doubly robust estimators raise questions about inference validity when applied to real study data.

In this paper, we develop an approach to address all the aforementioned limitations by leveraging functional response models (FRM), a class of semiparamatric regression models for between-subject attributes that include both PI and GPI. In Section 3.2, we review potential outcomes and causal effects for the MWWRST. In Section 3.3, we present IPW estimators, outcome regression (mean score imputation) estimators and doubly robust estimators by combing the IPW and outcome regression estimators. In Section 3.4, we discuss joint inference for all the three estimators by leveraging the FRM and U-statistics based generalized estimating equations. In Section 3.5, we examine performances of the proposed doubly robust estimator for small and large samples using both simulated and real study data. We discuss future directions in Section 3.6.

3.2 Causal Effect for Mann-Whitney-Wilcoxon Rank-sum Test

There are two equivalent forms of the MWWRST [50]. We use the U-statistics expression, also known as the Mann–Whitney form, for the development below unless stated otherwise.

Consider two independent groups, indexed by k $(= 0, 1)$, with sample size n_k, and let $y_{i_k k}$ denote an observed continuous outcome from the i_k subject in group k. The MWWRST is given by:

$$\text{Mann–Whitney form of MWWRST}: \quad \widehat{\delta}_n = \frac{1}{n_1}\frac{1}{n_0}\sum_{i_1=1}^{n_1}\sum_{j_0=1}^{n_0} I\left(y_{i_1 1} \leq y_{j_0 0}\right). \quad (3.1)$$

The U-statistic $\widehat{\delta}_n$ is an unbiased estimator of $\tilde{\delta} = E\left[I\left(y_{i_1 1} \leq y_{j_0 0}\right)\right]$ [50]. The MWWRST tests whether the mean rank of $y_{i_k k}$ is the same between the two groups, since the null $H_0 : \tilde{\delta} = \frac{1}{2}$ holds true if and only if $y_{i_k k}$ has the same mean rank [51]. Only under some assumptions does equal mean rank imply equal median [52]. This correct interpretation is important, since there is a long-standing misconception that the MWWRST always compare medians of two distributions.

To apply the concept of potential outcomes to non-randomized studies, consider a sample of size n, and let z_i $(= 0, 1)$ denote an indicator of treatment assignment

, or exposure status, and let (y_{i1}, y_{i0}) denote the potential outcomes corresponding to the two treatment conditions k $(= 0, 1)$. Then, for a randomized controlled trial (RCT), $y_{ik} \perp z_i$ and we observe one of the potential outcome y_{i1}, denoted as $y_{i_1 1}$, or y_{i0}, denoted as $y_{i_0 0}$, depending on whether the subject is randomized to group $k = 1$ or $k = 0$ $(1 \leq i_k \leq n_k,\, n = n_0 + n_1)$. As in causal inference about treatment effects based on comparing the means of y_{ik}, we would like to use the following to indicate treatment effect:

$$\Delta = E\left[I\left(y_{i1} \leq y_{i0}\right)\right], \tag{3.2}$$

and then apply the counterfactual consistency assumption to estimate Δ using the observed $y_{i_k k}$ [41]. Unfortunately, such an approach does not work.

First, unlike mean-based models, it is impossible to estimate the rank-based Δ using observed $y_{i_k k}$, because only one of the potential outcomes for each subject is observed. Second, $\tilde{\delta} \neq \Delta$, which is easily seen by noting the fact that the within-subject pair (y_{i1}, y_{i0}) generally has smaller variability than the between-subject pair (y_{i1}, y_{j0}).

Therefore we define causal effects for the MWWRST by:

$$\delta = E\left[I\left(y_{i1} \leq y_{j0}\right)\right], \quad \text{for any } (i, j) \in C_2^n, \tag{3.3}$$

where C_q^n denotes the set of $\binom{n}{q}$ combinations of q distinct elements $(i_1, i_2, \ldots i_q)$ from the integer set $\{1, \ldots, n\}$. For RCTs, δ is equal to $\tilde{\delta}$ and can be consistently estimated by $\widehat{\delta}_n$ in (3.1) based on the observed $(y_{i_1 1}, y_{j_0 0})$ $(i_1 \in C_1^{n_1},\, j_0 \in C_1^{n_0})$, which is a subset of the potential outcomes $(\, y_{i1}, y_{j0})$ $((i, j) \in C_2^n)$. For non-RCTs, $\widehat{\delta}_n$ is no longer a consistent estimator of δ.

Note that as MWWRST is usually called for dealing with outliers, we will consider unbounded continuous or count outcomes for y_{ik}. When ties are present, the null in this case may be expressed as $H_0 : E\left(I\left(y_{i1} < y_{j0}\right)\right) + \frac{1}{2}E\left(I\left(y_{i1} = y_{j0}\right)\right) = \frac{1}{2}$ [41]. Thus, for count outcomes, we redefine $f_{\mathbf{i}}$ in (3.4) with $I\left(y_{i1} \leq y_{j0}\right)$ replaced by $I\left(y_{i1} < y_{j0}\right) + \frac{1}{2}I\left(y_{i1} = y_{j0}\right)$. Without loss of generality, we focus on (unbounded) continuous y_{ik} unless stated otherwise.

3.3 Doubly Robust Estimator for Mann-Whitney-Wilcoxon Rank-sum Test

We start with a brief review of IPW estimator for the MWWRST.

3.3.1 Inverse Probability Weighting Estimator

Let $p = E(z_i)$. For RCTs, p is a known constant indicating the probability of treatment assignment (exposure status). Let

$$f_{\mathbf{i}} = \frac{1}{2}\left(\frac{z_i(1-z_j)}{p(1-p)} I(y_{i1} \le y_{j0}) + \frac{z_j(1-z_i)}{p(1-p)} I(y_{j1} \le y_{i0})\right), \quad \mathbf{i} = (i,j) \in C_2^n. \tag{3.4}$$

Although only one of the potential outcomes (y_{i1}, y_{i0}) is observed, $f_{\mathbf{i}}$ in the above is well-defined, since $z_i(1-z_j) = 0$ for pairs of subjects who are assigned the same treatment. For an RCT, $\widehat{\delta}_n$ in (3.1) is a consistent estimator of δ and can be expressed as:

$$\begin{aligned} \widehat{\delta}_n &= (n_1 n_0)^{-1} \sum_{i_1=1}^{n_1} \sum_{j_0=1}^{n_0} I(y_{i_1 1} \le y_{j_0 0}) \qquad (3.5) \\ &= \left[\frac{1}{2}\sum_{\mathbf{i}\in C_2^n}\left(\frac{z_i(1-z_j)}{p(1-p)} + \frac{z_j(1-z_i)}{p(1-p)}\right)\right]^{-1}\left(\sum_{\mathbf{i}\in C_2^n} f_{\mathbf{i}}\right) \\ &= A_n^{-1}\left(\sum_{\mathbf{i}\in C_2^n} f_{\mathbf{i}}\right). \end{aligned}$$

, where $A_n^{-1} = \left[\frac{1}{2}\sum_{\mathbf{i}\in C_2^n}\left(\frac{z_i(1-z_j)}{p(1-p)} + \frac{z_j(1-z_i)}{p(1-p)}\right)\right]^{-1}$. By the theory of U-statistics [50],

$$\delta = E(f_{\mathbf{i}}), \quad \sqrt{n}\binom{n}{2}^{-1} \sum_{\mathbf{i}\in C_2^n} (f_{\mathbf{i}} - \delta) \to_d N\left(0, \sigma_\delta^2\right), \quad \binom{n}{2}^{-1} A_n \to_p 1, \tag{3.6}$$

where $\to_d$ ($\to_p$) denotes convergence in distribution (probability) and σ_δ^2 denotes the asymptotic variance of
$\sqrt{n}\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}(f_{\mathbf{i}} - \delta)$. It follows from the Slutsky's theorem that $\widehat{\delta}_n$ has the same asymptotic distribution as $\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n} f_{\mathbf{i}}$.

For observational studies, z_i generally depends on y_{ik}, and as a result, $\widehat{\delta}_n$ in (3.5) is generally a biased estimator of δ. Suppose there exists a vector of confounders, or covariates, $\mathbf{w}_i$, such that $y_{ik} \perp z_i \mid \mathbf{w}_i$ $(k = 0, 1)$. Let $\pi_i = E(z_i \mid \mathbf{w}_i)$, $\mathbf{w_i} = \{\mathbf{w}_i, \mathbf{w}_j\}$ and

$$f_{\mathbf{i}}^{IPW} = \frac{1}{2}\left(\frac{z_i(1-z_j)}{\pi_i(1-\pi_j)} I(y_{i1} \leq y_{j0}) + \frac{z_j(1-z_i)}{\pi_j(1-\pi_i)} I(y_{j1} \leq y_{i0})\right), \quad \mathbf{i} = (i,j) \in C_2^n. \tag{3.7}$$

Then we have:

$$\begin{aligned} E\left(f_{\mathbf{i}}^{IPW}\right) &= \frac{1}{2} E\left[\frac{z_i(1-z_j)}{\pi_i(1-\pi_j)} I(y_{i1} \leq y_{j0}) + \frac{z_j(1-z_i)}{\pi_j(1-\pi_i)} I(y_{j1} \leq y_{i0})\right] \\ &= \frac{1}{2} E\left[\frac{I(y_{i1} \leq y_{j0})}{\pi_i(1-\pi_j)} E(z_i(1-z_j) \mid \mathbf{w_i}) + \frac{I(y_{j1} \leq y_{i0})}{\pi_j(1-\pi_i)} E(z_j(1-z_i) \mid \mathbf{w_i})\right] \\ &= \delta. \end{aligned}$$

If π_i is known, then as in the case of RCTs $\widehat{\delta}_n^{IPW}$ below is a consistent and asymptotically normal estimator of δ:

$$\widehat{\delta}_n^{IPW} = \binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} f_{\mathbf{i}}^{IPW} \to_d \delta, \quad \sqrt{n}\left(\widehat{\delta}_n^{IPW} - \delta\right) \to_d N\left(0, \sigma_\delta^2\right), \tag{3.8}$$

where σ_δ^2 denotes the asymptotic variance.

In practice, π_i is unknown and can be modeled using any parametric [43, 53, 54] or semiparametric GLM with a link for a binary response such as the logit link [41, 44]. Let $\pi_i = \pi(\mathbf{w}_i; \boldsymbol{\eta})$ with $\boldsymbol{\eta}$ indicating an unknown vector of parameters. We can estimate $\boldsymbol{\eta}$ using maximum likelihood or estimating equations. Let $\widehat{\boldsymbol{\eta}}_n$ denote such an estimator of $\boldsymbol{\eta}$. In this case, $\widehat{\delta}_n^{IPW}$ in (3.8) will be a function of $\widehat{\boldsymbol{\eta}}_n$, i.e., $\widehat{\delta}_n^{IPW}(\widehat{\boldsymbol{\eta}}_n)$. By using a Taylor series expansion, we can obtain the asymptotic variance of $\widehat{\delta}_n^{IPW}(\widehat{\boldsymbol{\eta}}_n)$ that accounts for sampling variability of $\widehat{\boldsymbol{\eta}}_n$ for inference about δ [41, 43, 53, 54].

3.3.2 Outcome Regression Estimator

The IPW in Section 3.3.1 only uses the observed pairs $(y_{i_1 1}, y_{j_0 0})$ $(i_1 \in C_1^{n_1}$, $j_0 \in C_1^{n_0})$. Alternatively, we can posit a model to relate outcome y_{ik} with $\mathbf{w}_i$ and impute missing y_{ik}.

This approach was considered by Mao (2018) and Zhang et al. (2019). However, as noted in Section 3.1, parametric models are at odds with the reason for employing the MWWRST in the first place. Moreover, in all real study applications, no reliable estimator can be obtained for such a parametric model because of outliers.

Zhang et al. (2019) introduced another estimator by positing a semiparametric generalized probability index (GPI) model:

$$E\left(I\left(y_{i1} \leq y_{j0}\right) \mid \mathbf{w}_{\mathbf{i}}\right) = g\left(\mathbf{w}_{\mathbf{i}}; \boldsymbol{\gamma}\right), \quad \mathbf{i} = (i, j) \in C_2^n, \tag{3.9}$$

where $\boldsymbol{\gamma}$ denotes a vector of parameters. The GPI above is a member of functional response models (FRM) for between-subject attributes (see Section 3.4 for more details about FRM). This particular form of FRM has been used in a similar context to address outliers for linear regression in Chen et al. (2014)[55] and Chen et al. (2016) [56]. We will refer to the GPI in (3.9) as an FRM in the following discussion unless stated otherwise. The FRM in (3.9) is much more robust as it only models the conditional mean of $I\left(y_{i1} \leq y_{j0}\right)$ given $\mathbf{w}_{\mathbf{i}}$. In addition, it allows us to directly impute missing $I\left(y_{i1} \leq y_{j0}\right)$ due to unobserved y_{i1} or y_{j0}.

If $\boldsymbol{\gamma}$ is known, then under $y_{ik} \perp z_i \mid \mathbf{w}_i$ and $y_{jk} \perp z_j \mid \mathbf{w}_j$, with $\mathbf{w}_i$ and $\mathbf{w}_j$ indicate vectors of covariates for different subject i and j, we can impute missing $I\left(\left\{y_{i1} \leq y_{j0}\right\}\right)$ or $I\left(\left\{y_{j1} \leq y_{i0}\right\}\right)$ with the mean score (MS), $g\left(\mathbf{w}_{\mathbf{i}}; \boldsymbol{\gamma}\right)$ or $g\left(\mathbf{w}_{\mathbf{i}^c}; \boldsymbol{\gamma}\right)$ with $\mathbf{w}_{\mathbf{i}} = \left\{\mathbf{w}_i, \mathbf{w}_j\right\}$ and $\mathbf{w}_{\mathbf{i}^c} = \left\{\mathbf{w}_j, \mathbf{w}_i\right\}$. For example, under the logit link and additive linear predictor, we can posit:

$$g\left(\mathbf{w}_{\mathbf{i}}; \boldsymbol{\gamma}\right) = \operatorname{expit}\left(\gamma_0 + \boldsymbol{\gamma}_{11}^{\top}\mathbf{w}_i + \boldsymbol{\gamma}_{10}^{\top}\mathbf{w}_j\right), \quad g\left(\mathbf{w}_{\mathbf{i}^c}; \boldsymbol{\gamma}\right) = \operatorname{expit}\left(\gamma_0 + \boldsymbol{\gamma}_{11}^{\top}\mathbf{w}_j + \boldsymbol{\gamma}_{10}^{\top}\mathbf{w}_i\right).$$

where expit$(\cdot)$ denotes the inverse of the logit link.

Note that $g\left(\mathbf{w}_{\mathbf{i}}; \boldsymbol{\gamma}\right)$ is the mean of the semiparametric FRM for between-subject attributes $\left\{I\left(y_{i1} \leq y_{j0}\right), \mathbf{w}_{\mathbf{i}}\right\}$ in (3.9), which generally does not have a direct relationship with semiparametric GLM for within-subject attributes $\left\{y_{ik}, \mathbf{w}_i\right\}$. For example, if conditioning on $\mathbf{w}_i$, y_{ik} follows a semiparametric linear regression:

$$y_{ik} = \beta_0 + \boldsymbol{\beta}_1^{\top}\mathbf{w}_{ik} + \epsilon_{ik}, \quad 1 \leq i \leq n_k, \quad k = 0, 1.$$

then the link function g^{-1} for the semiparametric FRM in (3.9) is determined by the distribution of $\epsilon_{ij} = \epsilon_{i1} - \epsilon_{j0}$. For normal-distributed ϵ_{ik}, ϵ_{ij} is also normal and g^{-1} is the probit link:

$$g\left(\mathbf{w}_{\mathbf{i}};\boldsymbol{\gamma}\right) = \Phi(\gamma_0 + \boldsymbol{\gamma}_{11}^{\top}\mathbf{w}_i + \boldsymbol{\gamma}_{10}^{\top}\mathbf{w}_j), \tag{3.10}$$

where $\Phi\left(\cdot\right)$ denotes the cumulative distribution function (CDF) of the standard normal $N\left(0,1\right)$. If ϵ_{ik} follows other distributions such as t or logistic, ϵ_{ij} will not have a t or logistic distribution. In practice, we can use differen link functions for within-subject binary responses for the between-subject FRM in (3.10).

The following $f_{\mathbf{i}}^{MSI}$ based on such mean score imputed (MSI) $g\left(\mathbf{w}_{\mathbf{i}};\boldsymbol{\gamma}\right)$ and $g\left(\mathbf{w}_{\mathbf{i}^c};\boldsymbol{\gamma}\right)$ for the unobserved $I\left(y_{i1} \leq y_{j0}\right)$ and $I\left(y_{j1} \leq y_{i0}\right)$ is well-defined for all $\binom{n}{2}$ subject pairs of the combined sample:

$$\begin{aligned} f_{\mathbf{i}}^{MSI} \quad &= \quad \frac{1}{2}\left[z_i\left(1-z_j\right)I\left(y_{i1} \leq y_{j0}\right) + \left(1 - z_i\left(1-z_j\right)\right)g\left(\mathbf{w}_{\mathbf{i}};\boldsymbol{\gamma}\right)\right] \\ &\quad\; + \frac{1}{2}\left[z_j\left(1-z_i\right)I\left(y_{j1} \leq y_{i0}\right) + \left(1 - z_j\left(1-z_i\right)\right)g\left(\mathbf{w}_{\mathbf{i}^c};\boldsymbol{\gamma}\right)\right]. \end{aligned} \tag{3.11}$$

Also, we have:

$$\begin{aligned} E\left(f_{\mathbf{i}}^{MSI}\right) &= \frac{1}{2}E\left[z_i\left(1-z_j\right)I\left(y_{i1} \leq y_{j0}\right) + \left(1 - z_i\left(1-z_j\right)\right)g\left(\mathbf{w}_{\mathbf{i}};\boldsymbol{\gamma}\right)\right] \\ &+ \frac{1}{2}E\left[z_j\left(1-z_i\right)I\left(y_{j1} \leq y_{i0}\right) + \left(1 - z_j\left(1-z_i\right)\right)g\left(\mathbf{w}_{\mathbf{i^c}};\boldsymbol{\gamma}\right)\right] \\ &= \frac{1}{2}[E\left\{z_i\left(1-z_j\right)\left[E\left(I\left(y_{i1} \leq y_{j0}\right) - g\left(\mathbf{w}_{\mathbf{i}};\boldsymbol{\gamma}\right)\right) \mid \mathbf{w}_{\mathbf{i}}\right]\right\} + E\left[g\left(\mathbf{w}_{\mathbf{i}};\boldsymbol{\gamma}\right)\right]] \\ &+ \frac{1}{2}[E\left\{z_j\left(1-z_i\right)\left[E\left(I\left(y_{j1} \leq y_{i0}\right) - g\left(\mathbf{w}_{\mathbf{i^c}};\boldsymbol{\gamma}\right)\right) \mid \mathbf{w}_{\mathbf{i^c}}\right]\right\} + E\left[g\left(\mathbf{w}_{\mathbf{i^c}};\boldsymbol{\gamma}\right)\right]] \\ &= E\left[\frac{1}{2}E\left(I\left(y_{i1} \leq y_{j0}\right) \mid \mathbf{w}_{\mathbf{i}}\right) + \frac{1}{2}E\left(I\left(y_{j1} \leq y_{i0}\right) \mid \mathbf{w}_{\mathbf{i^c}}\right)\right] \\ &= \delta. \end{aligned}$$

Thus by the theory of U-statistics, the estimator $\widehat{\delta}_n^{MSI}$ based on the mean score imputed $f_{\mathbf{i}}^{MSI}$ is consistent and asymptotically normal with asymptotic variance τ_δ^2:

$$\widehat{\delta}_n^{MSI} = \binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} f_{\mathbf{i}}^{MSI} \to_d \delta, \quad \sqrt{n}\left(\widehat{\delta}_n^{MSI} - \delta\right) \to_d N\left(0, \tau_\delta^2\right).$$

If $\boldsymbol{\gamma}$ is unknown as in most applications, we can estimate $\boldsymbol{\gamma}$ using U-statistics based generalized estimating equations for sesmiparametric FRM (see Section 3.4 for details) [44, 53]. As in the case of IPW estimators above, we can also account for sampling variability in estimated $\widehat{\boldsymbol{\gamma}}_n$ in the asymptotic variance of $\widehat{\delta}_n^{MSI}\left(\widehat{\boldsymbol{\gamma}}_n\right)$ [55, 56]. This asymptotic inference addresses the limitation of bootstrap inference used in estimators proposed by Zhang et al. (2019).

3.3.3 Doubly Robust Estimator

First, we assume $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ are known and discuss how to construct a doubly robust estimator of δ by combining the two.

To this end, let

$$\begin{aligned} f_{\mathbf{i}}^{DR} &= \frac{1}{2}\left[\frac{z_i\left(1-z_j\right)}{\pi_i\left(1-\pi_j\right)} I\left(y_{i1} \leq y_{j0}\right)+\left(1-\frac{z_i\left(1-z_j\right)}{\pi_i\left(1-\pi_j\right)}\right) g_{\mathbf{i}}\right] \\ &\quad +\frac{1}{2}\left[\frac{z_j\left(1-z_i\right)}{\pi_j\left(1-\pi_i\right)} I\left(y_{j1} \leq y_{i0}\right)+\left(1-\frac{z_j\left(1-z_i\right)}{\pi_j\left(1-\pi_i\right)}\right) g_{\mathbf{i}^c}\right], \\ \pi_i &= \pi\left(\mathbf{w}_i ; \boldsymbol{\eta}\right), \quad g_{\mathbf{i}}=g\left(\mathbf{w}_{\mathbf{i}} ; \boldsymbol{\gamma}\right), \quad g_{\mathbf{i}^c}=g\left(\mathbf{w}_{\mathbf{i}^c} ; \boldsymbol{\gamma}\right), \quad \mathbf{i}=(i, j) \in C_2^n . \end{aligned} \tag{3.12}$$

The above $f_{\mathbf{i}}^{DR}$ is well-defined for all $\binom{n}{2}$ subject pairs of the combined sasmple. We now show that $E\left(f_{\mathbf{i}}^{DR}\right)=\delta$ if one of the regression models, $\pi\left(\mathbf{w}_i ; \boldsymbol{\eta}\right)$ and $g\left(\mathbf{w}_{\mathbf{i}} ; \boldsymbol{\gamma}\right)$, is correctly specified. Since the two terms in $f_{\mathbf{i}}^{DR}$ have the same mean, it suffices to show this result only for the first term, i.e.,

$$\begin{aligned} E\left(I_{\mathbf{i}}\right) &= E\left[\frac{z_i\left(1-z_j\right)}{\pi_i\left(1-\pi_j\right)} I\left(y_{i1} \leq y_{j0}\right)+\left(1-\frac{z_i\left(1-z_j\right)}{\pi_i\left(1-\pi_j\right)}\right) g_{\mathbf{i}}\right] \\ &= \delta . \end{aligned}$$

1. If $E\left(I\left(y_{i1} \leq y_{j0}\right) \mid \mathbf{w}_{\mathbf{i}}\right)=g\left(\mathbf{w}_{\mathbf{i}} ; \boldsymbol{\gamma}\right)$ is correctly specified, then we have:

$$\begin{aligned} E\left(I_{\mathbf{i}}\right) &= E\left(I\left(y_{i1} \leq y_{j0}\right)\right)+E\left[\frac{z_i(1-z_j)-\pi_i\left(1-\pi_j\right)}{\pi_i\left(1-\pi_j\right)}\left(I\left(y_{i1} \leq y_{j0}\right)-g\left(\mathbf{w}_{\mathbf{i}} ; \boldsymbol{\gamma}\right)\right)\right] \\ &= E\left(I\left(y_{i1} \leq y_{j0}\right)\right) \\ &\quad + E\left\{E\left[\frac{z_i(1-z_j)-\pi_i\left(1-\pi_j\right)}{\pi_i\left(1-\pi_j\right)}\left(I\left(y_{i1} \leq y_{j0}\right)-g\left(\mathbf{w}_{\mathbf{i}} ; \boldsymbol{\gamma}\right)\right) \mid \mathbf{w}_{\mathbf{i}}, z_i, z_j\right]\right\} \\ &= E\left(I\left(y_{i1} \leq y_{j0}\right)\right)+E\left\{\frac{z_i(1-z_j)-\pi_i\left(1-\pi_j\right)}{\pi_i\left(1-\pi_j\right)}\left[g\left(\mathbf{w}_{\mathbf{i}} ; \boldsymbol{\gamma}\right)-g\left(\mathbf{w}_{\mathbf{i}} ; \boldsymbol{\gamma}\right)\right]\right\} \\ &= \delta . \end{aligned}$$

2. If $E\left(z_i \mid \mathbf{w}_{\mathbf{i}}\right) = \pi\left(\mathbf{w}_i; \boldsymbol{\eta}\right)$ is correctly specified, then we have:

$$\begin{aligned}
E\left(I_{\mathbf{i}}\right) &= E\left[\frac{z_i\left(1-z_j\right)}{\pi_i\left(1-\pi_j\right)} I\left(y_{i1} \leq y_{j0}\right) + \left(1 - \frac{z_i\left(1-z_j\right)}{\pi_i\left(1-\pi_j\right)}\right) g\left(\mathbf{w}_{\mathbf{i}}; \boldsymbol{\gamma}\right)\right] \\
&= E\left\{E\left[\frac{z_i\left(1-z_j\right)}{\pi_i\left(1-\pi_j\right)} I\left(y_{i1} \leq y_{j0}\right)\right] \mid \mathbf{w}_{\mathbf{i}}, y_{i1}, y_{j0}\right\} \\
&+ E\left\{E\left[\left(1 - \frac{z_i\left(1-z_j\right)}{\pi_i\left(1-\pi_j\right)}\right) g\left(\mathbf{w}_{\mathbf{i}}; \boldsymbol{\gamma}\right)\right] \mid \mathbf{w}_{\mathbf{i}}\right\} \\
&= E\left[\frac{I\left(y_{i1} \leq y_{j0}\right)}{\pi_i\left(1-\pi_j\right)} E\left(z_i\left(1-z_j\right) \mid \mathbf{w}_{\mathbf{i}}\right)\right] \\
&+ E\left[g\left(\mathbf{w}_{\mathbf{i}}; \boldsymbol{\gamma}\right)\left(1 - \frac{1}{\pi_i\left(1-\pi_j\right)} E\left(z_i\left(1-z_j\right) \mid \mathbf{w}_{\mathbf{i}}\right)\right)\right] \\
&= \delta.
\end{aligned}$$

Thus, if $\pi\left(\mathbf{w}_i; \boldsymbol{\eta}\right)$ or $g\left(\mathbf{w}_{\mathbf{i}}; \boldsymbol{\gamma}\right)$ is correctly specified, it follows from the theory of U-statistics that the estimator $\widehat{\delta}_n^{DR}$ below based on $f_{\mathbf{i}}^{DR}$ in (3.12) is consistent and asymptotically normal with asymptotic variance ς_δ^2:

$$\widehat{\delta}_n^{DR} = \binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} f_{\mathbf{i}}^{DR} \rightarrow_d \delta, \quad \sqrt{n}\left(\widehat{\delta}_n^{DR} - \delta\right) \rightarrow_d N\left(0, \varsigma_\delta^2\right).$$

Since $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ are both unknown as in most applications, we may again first estimate $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ using maximum likelihood (as in Mao, 2018 and Zhang et al, 2019) or U-statistics based generalized estimating equations through semiparametric GLM and FRM (as in Wu et al., 2014 and Chen et al., 2016). Given such estimators, we then compute $\widehat{\delta}_n^{DR}\left(\widehat{\boldsymbol{\eta}}_n, \widehat{\boldsymbol{\gamma}}_n\right)$ and its asymptotic variance estimates. When modeling the outcome regression using FRM, the asymptotic variance inference again addresses the limitation of bootstrap inference in Zhang et al. (2019). Alternatively, we may utilize the flexibility of FRM to jointly estimate δ, $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$, as we discuss next.

3.4 Inference for Functional Response Models

We first provide a brief overview of functional response models (FRM). More details can be found in Kowalski and Tu (2007) [50] and other citations below.

Let y_i and $\mathbf{x}_i$ denote some response and a vector of predictors (or covariates) from the ith subject $(1 \leq i \leq n)$. The class of functional response model (FRM) is defined by:

$$E\left[f\left(y_{i_1}, \ldots, y_{i_q}; \boldsymbol{\theta}\right) \mid \mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_q}\right] = h\left(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_q}; \boldsymbol{\theta}\right), \quad (i_1, \ldots, i_q) \in C_q^n, \tag{3.13}$$

where $f(\cdot)$ is some function, $h(\cdot)$ some smooth function (e.g., continuous second-order derivatives), C_q^n denotes the set of $\binom{n}{q}$ combinations of q (≥ 1) distinct elements $(i_1, \ldots, i_q)$ from the integer set $\{1, \ldots, n\}$ and $\boldsymbol{\theta}$ a vector of parameters. For $q = 1$, $f(y_i; \boldsymbol{\theta}) = y_i$ and $h(\mathbf{x}_i; \boldsymbol{\theta}) = h\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)$, (3.13) reduces to the semiparametric, or restricted moment, generalized linear models: $E\left(y_i \mid \mathbf{x}_{i_1}\right) = h\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)$, $1 \leq i \leq n$. For $q = 2$, $f(y_i, y_j; \theta) = I(y_i \leq y_j)$ and $h(\mathbf{x}_i, \mathbf{x}_j; \theta) = E[I(y_i \leq y_j)] = \theta$, (3.13) reduces to (3.3) for causal effects of the MWWRST. By extending traditional semiparametric regression models for within-subject attributes to between-subject attributes, the FRM has been utilized to model between-subject relationships in a wide range of applications such as social network connectivity [57], Beta-diversity in microbiome research [49],and a host of popular reliability indices such as Pearson and concordance correlation coefficients [53, 54, 58, 59]. Within the current context, we utilize the FRM to facilitate infernce when applying the doubly robust estimator to observational study data. We focus on joint inference about $\boldsymbol{\theta} = \left(\delta, \boldsymbol{\eta}^\top, \boldsymbol{\gamma}^\top\right)^\top$ for the doubly robsut estimator $\widehat{\delta}_n^{DR}$. Similar FRMs are readily developed for joint inference about the parameters for the IPW and MSI estimators.

Consider an FRM of the form:

$$
\begin{aligned}
& E\left(\mathbf{f}_{\mathbf{i}} \mid \mathbf{x}_{\mathbf{i}}\right)=\mathbf{h}_{\mathbf{i}}\left(\mathbf{x}_{\mathbf{i}};\boldsymbol{\theta}\right), \quad \mathbf{i}=(i,j)\in C_2^n, \qquad (3.14)\\
& \qquad \mathbf{f}_{\mathbf{i}}=\left(f_{\mathbf{i}1},f_{\mathbf{i}2},f_{\mathbf{i}3}\right)^{\top}, \quad \mathbf{h}_{\mathbf{i}}=\left(h_{\mathbf{i}1},h_{\mathbf{i}2},h_{\mathbf{i}3}\right)^{\top}, \quad \mathbf{x}_{\mathbf{i}}=\left\{\mathbf{x}_i,\mathbf{x}_j\right\},\\
& \qquad f_{\mathbf{i}1}=\frac{1}{2}(z_i+z_j), \quad f_{\mathbf{i}2}=\frac{1}{2}\left[I(y_{i1}\le y_{j0})+I(y_{j1}\le y_{i0})\right],\\
& \qquad f_{\mathbf{i}3}=\frac{1}{2}\left[\frac{z_i(1-z_j)}{\pi_i(1-\pi_j)}I(y_{i1}\le y_{j0})+\left(1-\frac{z_i(1-z_j)}{\pi_i(1-\pi_j)}\right)g_{\mathbf{i}}\right]\\
& \qquad\quad +\frac{1}{2}\left[\frac{z_j(1-z_i)}{\pi_j(1-\pi_i)}I(y_{j1}\le y_{i0})+\left(1-\frac{z_j(1-z_i)}{\pi_j(1-\pi_i)}\right)g_{\mathbf{i}^c}\right],\\
& h_{\mathbf{i}1}\left(\mathbf{w}_{\mathbf{i}};\boldsymbol{\theta}\right)=\frac{1}{2}(\pi\left(\mathbf{w}_i;\boldsymbol{\eta}\right)+\pi\left(\mathbf{w}_j;\boldsymbol{\eta}\right)), \quad h_{\mathbf{i}2}\left(\mathbf{w}_{\mathbf{i}};\boldsymbol{\theta}\right)=\frac{1}{2}\left[g\left(\mathbf{w}_{\mathbf{i}},\boldsymbol{\gamma}\right)+g\left(\mathbf{w}_{\mathbf{i}^c},\boldsymbol{\gamma}\right)\right],\\
& h_{\mathbf{i}3}\left(\mathbf{w}_{\mathbf{i}};\boldsymbol{\theta}\right)=\delta, \quad \pi_i=\pi\left(\mathbf{w}_i;\boldsymbol{\eta}\right), \quad \pi_j=\pi\left(\mathbf{w}_j;\boldsymbol{\eta}\right),\\
& \qquad g_{\mathbf{i}}=g\left(\mathbf{w}_{\mathbf{i}},\boldsymbol{\gamma}\right), \quad g_{\mathbf{i}^c}=g\left(\mathbf{w}_{\mathbf{i}^c},\boldsymbol{\gamma}\right),\\
& \qquad \boldsymbol{\theta}=\left(\boldsymbol{\eta}^{\top},\boldsymbol{\gamma}^{\top},\delta\right)^{\top}, \quad \boldsymbol{\eta}=\left(\eta_0,\boldsymbol{\eta}_1^{\top}\right)^{\top}, \quad \boldsymbol{\gamma}=\left(\gamma_0,\boldsymbol{\gamma}_{11}^{\top},\boldsymbol{\gamma}_{10}^{\top}\right)^{\top}.
\end{aligned}
$$

Let

$$
\begin{aligned}
& S_{\mathbf{i}}=\mathbf{f}_{\mathbf{i}}-\mathbf{h}_{\mathbf{i}}, \quad D_{\mathbf{i}}=\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{h}_{\mathbf{i}}\left(\boldsymbol{\theta}\right), \quad \mathbf{i}=(i,j)\in C_2^n.\\
& \mathbf{V}_{\mathbf{i}}=Var\left(\mathbf{f}_{\mathbf{i}}\mid\mathbf{w}_{\mathbf{i}}\right)=\begin{pmatrix} V_{\mathbf{i}1} & 0 & 0\\ 0 & V_{\mathbf{i}2} & 0\\ 0 & 0 & V_{\mathbf{i}3}\end{pmatrix}^{\frac{1}{2}} R\left(\boldsymbol{\alpha}\right)\begin{pmatrix} V_{\mathbf{i}1} & 0 & 0\\ 0 & V_{\mathbf{i}2} & 0\\ 0 & 0 & V_{\mathbf{i}3}\end{pmatrix}^{\frac{1}{2}},\\
& V_{\mathbf{i}1}=Var\left(f_{\mathbf{i}1}\mid\mathbf{w}_{\mathbf{i}}\right), \quad V_{\mathbf{i}2}=Var\left(f_{\mathbf{i}2}\mid\mathbf{w}_{\mathbf{i}}\right), \quad V_{\mathbf{i}3}=Var\left(f_{\mathbf{i}3}\mid\mathbf{w}_{\mathbf{i}}\right),
\end{aligned}
$$

where $V_{\mathbf{i}1}$, $V_{\mathbf{i}2}$ and $V_{\mathbf{i}3}$ are readily evaluated (see Appendix) and $R\left(\boldsymbol{\alpha}\right)$ denotes a working correlation matrix. Inference about $\boldsymbol{\theta}$ is based on class of U-statistics based generalized estimating equations (UGEE):

$$\mathbf{U}_n\left(\boldsymbol{\theta}\right)=\sum_{\mathbf{i}\in C_2^n}\mathbf{U}_{n,\mathbf{i}}=\sum_{\mathbf{i}\in C_2^n}D_{\mathbf{i}}V_{\mathbf{i}}^{-1}S_{\mathbf{i}}=\mathbf{0}. \qquad (3.15)$$

If either $\pi\left(\mathbf{w}_i;\boldsymbol{\eta}\right)$ or $g\left(\mathbf{w}_{\mathbf{i}},\boldsymbol{\gamma}\right)$ is correctly specified, then we have:

$$E\left(\mathbf{U}_{n,\mathbf{i}}\right)=E\left(D_{\mathbf{i}}V_{\mathbf{i}}^{-1}S_{\mathbf{i}}\right)=E\left[E\left(D_{\mathbf{i}}V_{\mathbf{i}}^{-1}S_{\mathbf{i}}\mid\mathbf{w}_{\mathbf{i}}\right)\right]=\mathbf{0}.$$

Thus the UGEE in (3.15) is unbiased, yielding consistent estimators of $\boldsymbol{\theta}$ if either $\pi\left(\mathbf{w}_i;\boldsymbol{\eta}\right)$ or $g\left(\mathbf{w}_{\mathbf{i}},\boldsymbol{\gamma}\right)$ is correctly specified. Further, UGEE estimators $\widehat{\boldsymbol{\theta}}$ are also asymptotically normal under mild regularity conditions. We summarize the asymptotic properties in a theorem below for ease of reference.

Theorem 1. Let

$$\mathbf{v}_i = E\left(\mathbf{U}_{n,\mathbf{i}} \mid y_{i1}, y_{i0}, z_i, \mathbf{w}_i\right), \quad \Sigma = Var\left(\mathbf{v}_i\right), \quad B = E\left(D_{\mathbf{i}} V_{\mathbf{i}}^{-1} D_{\mathbf{i}}^{\top}\right), \quad D_{\mathbf{i}} = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{h}_{\mathbf{i}}. \tag{3.16}$$

Then, under mild regularity conditions, we have:

1. $\widehat{\boldsymbol{\theta}}$ is consistent.

2. If $\sqrt{n}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) = \mathbf{O}_p(1)$, i.e., $\widehat{\boldsymbol{\alpha}}$ is $\sqrt{n}$-consistent [50], $\widehat{\boldsymbol{\theta}}$ is asymptotically normal:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \to_d N\left(\mathbf{0}, \Sigma_\theta = 4B^{-1}\Sigma B^{-\top}\right). \tag{3.17}$$

To estimate Σ_θ, we note that $\Sigma = Var\left(\mathbf{v}_i\right) = E\left(\mathbf{v}_i \mathbf{v}_i^{\top}\right)$. Thus,

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathbf{v}}_i \widehat{\mathbf{v}}_i^{\top}, \quad \widehat{\mathbf{v}}_i = \frac{1}{n-1} \sum_{j \neq i}^{n} \mathbf{U}_{n,ij}. \tag{3.18}$$

Also, we estimate B by $\widehat{B} = \binom{n}{2}^{-1} \sum_{(i,j) \in C_2^n} \widehat{D}_{\mathbf{i}} \widehat{V}_{\mathbf{i}}^{-1} \widehat{D}_{\mathbf{i}}^{\top}$, where $\widehat{D}_{\mathbf{i}}$ ($\widehat{V}_{\mathbf{i}}$) denotes $D_{\mathbf{i}}$ ($V_{\mathbf{i}}$) with $\widehat{\boldsymbol{\theta}}$ substituting for $\boldsymbol{\theta}$. Thus, a consistent estimate of Σ_θ is given by: $\widehat{\Sigma}_\theta = 4\widehat{B}^{-1}\widehat{\Sigma}\widehat{B}^{-\top}$.

3.5 Application

We illustrate the proposed approach with both simulated and real data. We start with investigating performance of the doubly robust estimator for small and large samples by simulation and then present an application to a real weight-loss trial to improve physical activities for breast cancer survivors. In all examples, we set a two-sided type I $\alpha = 0.05$. All analyses are carried out using codes developed using the R software platform [60].

3.5.1 Simulation Study

In order to investigate causal effect between two treatment groups under confounding bias, we generate data from the following setup for the potential outcome, confounder and treatment assignment mechanism:

$$\begin{aligned}
& y_{ik} = \beta_0 + \beta_1 I\left(z_i = k\right) + \beta_2 w_i + b_i + \epsilon_{ik}, \qquad (3.19)\\
& \operatorname{logit}(E(z_i \mid w_i)) = \eta_0 + \eta_1 w_i, \quad \pi\left(w_i; \eta\right) = \frac{\exp\left(\eta_0 + \eta_1 w_i\right)}{1 + \exp\left(\eta_0 + \eta_1 w_i\right)},\\
& \epsilon_{ik} \sim (\chi_1^2 - 1)\sqrt{\frac{\sigma^2}{2}}, \quad b_i \sim (\chi_1^2 - 1)\sqrt{\frac{\sigma_b^2}{2}},\\
& w_i \sim N(\mu_w, \sigma_w^2), \quad z_i \sim Bern(\pi\left(w_i; \boldsymbol{\eta}\right)), \quad k = 0, 1, \quad 1 \le i \le n,
\end{aligned}$$

It follows that:

$$\begin{aligned}
y_{i1} &= \beta_0 + \beta_1 + \beta_2 w_i + b_i + \epsilon_{i1}\\
y_{j0} &= \beta_0 + 0 + \beta_2 w_j + b_j + \epsilon_{j0}\\
E[I((y_{i1} - y_{j0}) \le 0 \mid \mathbf{w_i})] &= P((\epsilon_{i1} + b_i) - (\epsilon_{j0} + b_j) \le -\beta_2(w_i - w_j) - \beta_1 \mid \mathbf{w_i})\\
&= \Phi(\gamma_0 + \gamma_{11}^T w_i + \gamma_{10}^T w_j)\\
\gamma_0 = -\frac{1}{\sqrt{2(\sigma^2 + \sigma_b^2)}}\beta_1, &\quad \gamma_{11}^T = -\frac{1}{\sqrt{2(\sigma^2 + \sigma_b^2)}}\beta_2, \quad \gamma_{10}^T = \frac{1}{\sqrt{2(\sigma^2 + \sigma_b^2)}}\beta_2
\end{aligned}$$

where $\Phi\left(\cdot\right)$ is the cumulative distribution function (CDF) of the standard normal with mean 0 and standard deviation 1, $z_i = 1$ for the treated condition, $Bern(\pi_i)$ denotes a Bernoulli distribution with mean π_i, w_i is a baseline covariate connecting z_i and y_{ik}, and y_{ik} is the potential outcome under the influence of w_i, i.e., $\mathbf{y}_i \perp z_i \mid \mathbf{w}_i$. The amount of confounding bias for $\mathbf{y}_i$ is controlled through η_1.

Parameters for the simulation study are set to:

$$\begin{aligned}
& \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top = (0, 0, 1), \quad \boldsymbol{\eta} = (\eta_0, \eta_1) = (1, -1),\\
& \sigma^2 = \sigma_b^2 = 1, \quad u_w = 1, \quad \sigma_w^2 = 0.25.
\end{aligned}$$

With a negative η_1 and positive β_2, the potential outcomes y_{ik} have a smaller mean in the treated group ($z_i = 1$), i.e., $\delta = E\left(\frac{1}{2}(I\left\{y_{i1} \le y_{j0} \mid \mathbf{w_i}\right\}) + I\left\{y_{j1} \le y_{i0} \mid \mathbf{w_i}\right\}\right) > \frac{1}{2}$. The type I error is set to $\alpha = 0.05$.

Shown in Table 3.2 are estimated $\boldsymbol{\theta}$ for the FRM models for the three causal estimators, $\widehat{\delta}_{IPW}$, $\widehat{\delta}_{MSI}$ and $\widehat{\delta}_{DR}$, along with their asymptotic and empirical standard errors for the different sample sizes under 1,000 Monte Carlo samples. Table 3.2 also includes estimates $\widehat{\delta}_{MMW}$ from the standard MWWRST in (3.1) based on observed outcome $y_{i_k k}$, which were set to the observed potential outcome y_{ik} corresponding to the value of z_i, i.e., $y_{i_1 1}$ $(y_{i_0 0})$ if $z_i = 1$ (0) with $n_1 = \sum_{i=1}^{n} z_i$ and $n_0 = n - n_1$. The estimates of $\boldsymbol{\eta}$ and standard errors for $\widehat{\delta}_{IPW}$ and estimates of $\boldsymbol{\gamma}$ and standard errors for $\widehat{\delta}_{MSI}$ were all close to their counterparts for $\widehat{\delta}_{DR}$ across all sample sizes. All three causal estimates were quite close to the true $\delta = 0.5$ and the empirical type I errors were close to the nominal $\alpha = 0.05$. In contrast, estimates of $\widehat{\delta}_{MMW}$ exhibited high bias in estimating δ. The asymptotic standard errors of the causal estimators were nearly identical to their empirical counterparts. The DR estimator $\widehat{\delta}_{DR}$ had the smaller standard errors across all sample sizes, except for $n = 50$ in which case the asymptotic standard error of $\widehat{\delta}_{DR}$ was slightly larger, demonstrating the higher efficiency of $\widehat{\delta}_{DR}$ over $\widehat{\delta}_{IPW}$ and $\widehat{\delta}_{MSI}$.

To demonstrate the doubly robust properties of $\widehat{\delta}_{DR}$, we mis-specified $\pi\left(w_i; \boldsymbol{\eta}\right)$ as a constant $\pi\left(w_i; \boldsymbol{\eta}\right) = \text{expit}(\eta_0)$ for the IPW component or mis-specified $g\left(\mathbf{w_i}, \boldsymbol{\gamma}\right)$ as a constant $g\left(\mathbf{w_i}, \boldsymbol{\gamma}\right) = \Phi(\gamma_0)$ for the MSI component. Let $\widehat{\delta}_{DR}^{IPW}$ $(\widehat{\delta}_{DR}^{MSI})$ denote the resulting DR estimator with the IPW (MSI) component specified correctly. Shown in Table 3.3 are the estimated $\boldsymbol{\theta}$ for the FRM models for the DR estimator for the two scenarios, along with asymptotic and empirical standard errors for different sample sizes under 1,000 Monte Carlo samples. Both DR estimates were close to the true $\delta = 0.5$ across all sample sizes. The empirical type I errors were close to the nominal level $\alpha = 0.05$, albeit exhibiting small upward bias for $n = 50$. Table 3.3 also includes estimates from $\widehat{\delta}_{IPW}$ under mis-specified propensity score and $\widehat{\delta}_{MSI}$ under mis-specified outcome model for $n = 400$ (large sample size used to reduce sampling variability). As expected, both $\widehat{\delta}_{IPW}$ and $\widehat{\delta}_{MSI}$ showed significant amount of bias.

3.5.2 Data Application

We use data from Reach for Health (RFH) Study, a randomized control weight-loss trial to improve physical activities for non-diabetic breast cancer survivors conducted at the University of California San Diego, to illustrate how the proposed approach may be used to address outliers in real studies. By using a randomized control trial to create confounders, we can see how the proposed approach ad-

Table 3.2: Comparison of estimates of δ and standard errors (asymptotic vs. empirical) between FRM and traditional MWW approaches for the Simulation Study. For Double Robust estimator, both IPW and MSI are correctly specified.

	η_0	η_1	γ_0	γ_{11}	γ_{10}	δ	α
				$n = 50$			
$\widehat{\delta}_{DR}$	1.087 (0.678/0.746)	-1.071 (0.613/0.678)	0.093 (0.964/1.915)	-0.583 (0.825/1.285)	0.590 (1.127/1.943)	0.493 (0.085/0.080)	0.059
$\widehat{\delta}_{IPW}$	1.079 (0.679/0.737)	-1.068 (0.612/0.659)				0.513 (0.086/0.083)	0.055
$\widehat{\delta}_{MSI}$			0.101 (0.933/1.884)	-0.582 (0.792/1.117)	0.601 (1.084/1.896)	0.511 (0.084/0.083)	0.063
$\widehat{\delta}_{MWW}$						0.562 (0.080/0.110)	0.310
				$n = 200$			
$\widehat{\delta}_{DR}$	1.010 (0.327/0.334)	-1.009 (0.304/0.314)	0.045 (0.793/0.857)	-0.495 (0.596/0.624)	0.511 (0.741/0.830)	0.497 (0.040/0.040)	0.049
$\widehat{\delta}_{IPW}$	1.009 (0.328/0.335)	-1.006 (0.306/0.320)				0.502 (0.041/0.042)	0.048
$\widehat{\delta}_{MSI}$			0.046 (0.804/0.862)	-0.495 (0.607/0.648)	0.508 (0.744/0.842)	0.501 (0.041/0.041)	0.051
$\widehat{\delta}_{MWW}$						0.558 (0.041/0.040)	0.190
				$n = 400$			
$\widehat{\delta}_{DR}$	1.004 (0.238/0.248)	-1.007 (0.215/0.223)	0.041 (0.572/0.610)	-0.496 (0.485/0.505)	0.505 (0.559/0.564)	0.499 (0.037/0.037)	0.050
$\widehat{\delta}_{IPW}$	1.003 (0.238/0.250)	-1.007 (0.214/0.221)				0.501 (0.039/0.039)	0.050
$\widehat{\delta}_{MSI}$			0.039 (0.572/0.611)	-0.501 (0.488/0.512)	0.504 (0.560/0.565)	0.497 (0.038/0.039)	0.052
$\widehat{\delta}_{MWW}$						0.559 (0.037/0.038)	0.160

dresses confounding effects in the presence of outliers in real studies without being confounded by hidden bias [45, 46].

The RFH study has four arms that included a total of 333 participants that were overweight/obese (BMI $\geq 25kg/m^2$) and diagnosed with stage 1, 2, or 3 breast cancer within the past 10 years. The four-arms trial used a 2×2 factorial design with all participants randomly assigned to weight loss counseling vs. educational materials and to Metformin vs. placebo, with all interventions conducted over 6 months [61].

For illustration purposes, we combine the two medication groups (Metformin vs. placebo subjects) within each of the lifestyle intervention groups (weight loss counseling vs. educational materials) to consider only the effects of the lifestyle intervention. For the subjects included in this analysis, their demographics are shown in Table 3.4. The weight loss counseling serves as the group that would re-

Table 3.3: Comparison estimates and standard errors (asymptotic vs. empirical) of doubly robust estimator with only one component correctly specified and IPW (MSI) estimator with incorrectly specified propensity (imputation) function.

	η_0	η_1	γ_0	γ_{11}	γ_{10}	δ	α
				$n = 50$			
$\widehat{\delta}_{DR}^{IPW}$	1.088 (0.698/0.787)	-1.112 (0.675/0.783)	3.148 (0.975/1.963)	-	-	0.508 (0.091/0.089)	0.059
$\widehat{\delta}_{DR}^{MSI}$	-0.002 (0.286/0.295)	-	0.106 (0.981/1.924)	-0.580 (0.862/1.227)	0.573 (1.264/2.003)	0.509 (0.088/0.087)	0.064
				$n = 200$			
$\widehat{\delta}_{DR}^{IPW}$	1.009 (0.336/0.349)	-1.005 (0.322/0.339)	2.817 (0.815/1.052)	-	-	0.502 (0.043/0.044)	0.049
$\widehat{\delta}_{DR}^{MSI}$	-0.005 (0.138/0.142)	-	0.047 (0.822/0.893)	-0.489 (0.646/0.703)	0.507 (0.881/0.972)	0.501 (0.042/0.043)	0.051
				$n = 400$			
$\widehat{\delta}_{DR}^{IPW}$	1.003 (0.241/0.253)	-1.006 (0.215/0.227)	2.309 (0.751/0.803)	-	-	0.501 (0.040/0.041)	0.050
$\widehat{\delta}_{DR}^{MSI}$	-0.004 (0.100/0.100)	-	0.038 (0.575/0.614)	-0.501 (0.487/0.513)	0.505 (0.563/0.567)	0.497 (0.041/0.041)	0.051
$\widehat{\delta}_{MSI}$	-	-	2.311 (0.763/0.842)	-	-	0.563 (0.041/0.043)	0.164
$\widehat{\delta}_{IPW}$	-0.003 (0.107/0.114)	-	-	-	-	0.559 (0.041/0.042)	0.131

ceive lifestyle intervention consisting of 12 motivational interview calls from trained lifestyle coaches, while the educational materials group was given the 2010 US Dietary Guidelines. The general health score is derived from SF-36, which is a self-report questionnaire, with total scores ranging from 0 to 100 and with higher scores corresponding to better health [61, 62].

Assessments of behavioral outcomes were recorded on a daily basis. The primary interest was to compare subjects' weighted summation of day-level Moderate and Vigorous Physical Activity (MVPA) count between the weight loss counseling and educational materials group during the 6-month intervention period. For each subject i, the weighted summation day-level MVPA count is calculated by $\sum_{ijk} I(X_{ijk} \geq 1952) \times X_{ijk}$, where X_{ijk} denotes the activity count for the kth minute for subject i at day j and 1952 the threshold for counting as one minute of MVPA.

A common and challenging issue with the activity data is extreme values recorded by the device for subjects' activity levels, as shown in Table 3.1. As discussed in Section 3.1, the common approach of winsorizing outliers induces subjective opinions and rank-based methods such as the MWWRST objectively address this issue.

To use data from this RCT to illustrate the proposed approach for its ability to

Table 3.4: Demographics characteristics by groups.

	Intervention (N=166)	Control (N=167)
	mean (SD)	mean (SD)
Age at Enrollment	62.7(7.07)	62.5(7.01)
Age at Diagnosis	60.1(7.08)	60.0(6.80)
Year from Diagnosis		
to study enrollment	2.7(2.04)	2.5(1.79)
General Health (SF-36^a)	73.6(18.54)	70.0(18.96)
BMI	29.1(5.12)	30.4(5.09)
	N(%)	N(%)
White	105 (81.4%)	118 (84.9%)
College Education or greater	65 (50.4%)	70 (50.4%)
High Blood Pressure	61 (47.3%)	67 (48.2%)
Received Chemotherapy	67 (51.9%)	73 (52.5%)
Received Radiation Therapy	95 (73.6%)	100 (71.9%)
Smoking Status		
Never	66 (51.2%)	82 (59.0%)
Former	62 (48.1%)	55 (39.6%)
Current	1 (0.7%)	2 (1.4%)
Breast Cancer Stage		
I	68 (52.7%)	64 (46.0%)
II	41 (31.8%)	49 (35.3%)
III	20 (15.5%)	26 (18.7%)

address confounding bias, we assumed a hypothetical scenario in which the study intervention was not efficacious and the observed group activity difference was the result of confounding bias in selecting subjects who were more (less) likely to exercise for the intervention (control) group. We then selected four covariates, age at diagnosis, BMI, high cholesterol (a binary with 0 indicating low and 1 indicating high cholesterol) and general health score (higher indicating better health), as such confounders. We assigned the lowest 166 values of age at diagnosis, BMI and high cholesterol, and the highest 166 values of general health score to the intervention group. The remaining 167 values of the four covariates were assigned to the control group.

Shown in Table 3.5 are the means (percent) and standard deviations of the three (one) continuous (binary) covariates. As expected, the intervention group was healthier compared to the control group with respect to the four covariates, and thus was more likely to exercise than the control group. Since we only changed values of the four covariates for the study subjects, the traditional MWWRST yielded the same test statistic and p-value as shown in Table 3.1. But, the significant difference of activity level between the groups now was the result of confounding bias due to the four covariates, rather than the intervention effect.

To address this bias, we applied the proposed doubly robust approach by mod-

Table 3.5: Demographics characteristics by groups of imbalanced data.

	Intervention (N=166) mean (SD)	Control (N=167) mean (SD)
Age at Diagnosis	54.3(3.18)	65.4(4.94)
General Health Score	87.3(8.48)	57.3(13.49)
BMI	25.8(1.70)	33.5(4.49)
	N(%)	N(%)
High Cholesterol	6 (4.7%)	131 (94.2%)

eling the effects of the confounders through the IPW and MSI components by modeling the propensity score and outcome regression using the respective semi-parametric GLM and FRM with the logit link. The estimate of δ_{DR} was 0.501 with standard error = 0.043 and p-value = 0.973. Thus, the doubly robust estimator successfully addressed the biasing effects of the four confounders and indicated no treatment difference between the two groups of this hypothetical observational study.

3.6 Discussion

In this paper, we developed a doubly robust estimator to address the limitations of existing alternatives for more robust and reliable inference when applying the MWWRST to observational study data. We investigated the performance of the proposed method through both simulated and real data. The simulation study results demonstrated good performances even for samples as small as 50 when one of the propensity score and outcome regression model is correctly specified. The results from the real weight-loss trial showed that in addition to the doubly robust properties, the proposed estimator also effectively addressed outliers.

The proposed estimator is limited to cross-sectional study data. Work is currently underway to extend this approach to longitudinal cohort studies with missing data to facilitate causal inference for more complex study data arising in biomedical, clinical, epidemiological and psychosocial research.

3.7 Acknowledgement

Chapter 3, in full, has been submitted for publication as Chen, Ruohui; Lin, Tuo; Liu, Lin; Liu, Jinyuan; Chen, Ruifeng; Zou, Jingjing; Liu, Chenyu; Natarajan, Loki; and Tu, Xin. *A Double Robust Estimator for Mann Whiney Wilcoxon Rank Sum Test When Applied for Causal Inference in Observational Studies*, submitted to Journal of Applied Statistics. The dissertation author was the primary author on this paper.

Chapter 4

Conclusions and Future Work

As a bio-statistician with formal training in statistics and computation, I seek to develop practical and theoretically justified statistical methods, accompanied by robust software implementations, for the analysis of datasets generated from modern technologies, such as wearable devices.

Domain scientists are well aware of sources of identification and prior information that should 'count' as evidence in favor of their scientific hypothesis (or as evidence against their null hypothesis). Yet, a scientist's hands may be tied when no existing statistical methods can rigorously account for information they believe is relevant. As a bio-statistician, I seek to solve this conundrum by modeling the scientist's insight, formalizing inferential questions, and developing novel methods that can answer these questions in both theory and practice. In the dissertation, I am working towards this goal and trying to close the gap between the available dense excessive mobile health data and appropriate statistical methods.

In chapter one, we took advantage of a randomized controlled trial (RCT) of a 12-week physical activity intervention, trying to understand how emerging intervention modalities can be used to help people increase their physical activity, and reduce sedentary behaviors.

We used minute-level activity data collected from the Fitbit tracker, examined patterns of activity level and Fitbit use for both intervention and follow-up period, and compared patterns between the intervention and control groups, by using Generalized Additive Mixed Model (GAMM) and Linear Mixed Effect Model. We found that even though adherence to Fitbit use and physical activity level declined after the 12-week intervention period, the group that received a more active interventional strategy had a more stable trend and a higher level of adherence and physical activity in the follow-up period. These insights may enhance our ability to effectively utilize activity trackers to promote behavior change.

In chapter two, we tackled the problems that rise from measurement errors contained in wearable device recordings. There is a vast literature proposing statistical methods for adjusting for measurement errors in self-reported behaviors, such as in dietary intake. However, there is much less research on error correction for sensor-based device measures, especially sedentary behavior. We addressed this gap. Exploiting the excessive multiple-day assessments typically collected when sensor devices are deployed, we proposed a two-stage linear mixed effect model (LME) based approach to correct bias caused by measurement errors. We provided theoretical proof of the debiasing process using the Best Linear Unbiased Predictors (BLUP), and used both simulation and real data from a cohort study to demonstrate the performance of the proposed approach while comparing to the naïve plug-in approach that directly uses device measures without appropriately adjusting measurement errors. Our results indicate that employing our easy-to-implement BLUP correction method can greatly reduce biases in disease risk estimates and thus enhance the validity of study findings.

While the proposed LME-based structure models can correct the measurement errors in the exposure, it is important to note that multiple replicates are needed for the proposed method to be applicable. Since per best practices, health behavior researchers already require and collect multiple repeated days of device wear, this potential drawback, can be accommodated for sedentary behavior research, which is the focus of our application. Importantly, the LME structures can be straightforwardly implemented in different settings through standard statistical software, such as R and SAS, and generalized to other behaviors such as physical activity or sleep research.

Overall, Our approach sets a rigorous foundation, there is undoubtedly scope to expand and improve our methods. Although our data analysis study cohort was relatively complete, for future work, we are interested in expanding this work to accommodate missingness mechanisms. Meanwhile, currently, we treat measures within a short period of time (7 days) as replicates of each other, we also aim to extend the current setting to longitudinal data with different cluster sizes.

In chapter three, we proposed a double robust estimator that extended the traditional Mann-Whitney-Wilcoxon MWW rank sum test (MWWRST) for observational studies for causal inference. The Mann-Whitney-Wilcoxon MWW rank sum test (MWWRST) is widely used to compare two treatment groups in randomized control trials when data distributions are highly skewed, especially in the presence of outliers. However, the MWWRST generally yields invalid results for causal inference when applied to observational study data. We addressed this limitation by leveraging functional response models (FRM), a class of semiparametric regression models for between-subject attributes. As rank-based tests such as the MWWRST are defined by between-subject attributes, the FRM provides a native and effective paradigm to model such attributes to develop doubly robust estimators. We demonstrated the performances of the proposed approach through both simulated and real study data. The simulation study results demonstrated good performances even for samples as small as 50 when one of the propensity score and outcome regression model is correctly specified. The results from the real weight-loss trial showed that in addition to the doubly robust properties, the proposed estimator also effectively addressed outliers.

The proposed double robust estimator is limited to cross-sectional study data. Work is currently underway to extend this approach to longitudinal cohort studies with missing data to facilitate causal inference for more complex study data arising in biomedical, clinical, epidemiological, and psychosocial research.

Bibliography

[1] BLung-Fei Lee and Jungsywan H. Sepanski. “Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data.” In: *Journal of the American Statistical Association.* 90(429) (1995), pp. 130–140.

[2] Hartman SJ, Dillon LW, LaCroix AZ, Natarajan L, Sears DD, Owen N, Dunstan DW, Sallis JF, Schenk S, Allison M, Takemoto M, Herweck AM, Nguyen B, and Rosenberg DE. “Interrupting Sitting Time in Postmenopausal Women: Protocol for the Rise for Health Randomized Controlled Trial”. In: *JMIR Res Protoc* 10(5).65 (2021).

[3] Ballard-Barbash R, George SM, Alfano CM, and Schmitz K. “Physical activity across the cancer continuum. Oncology”. In: *Oncology* 6 (2013), p. 27.

[4] Melinda L irwin, Diane Crumley, Anne McTiernan, Leslie Bernstein, Richard Baumgartner, Frank D Gillilandm, Andrea Kriska, and Rachel Ballard-Barbash. “Physical activity levels before and after a diagnosis of breast carcinoma: the Health, Eating, Activity, and Lifestyle (HEAL) study”. In: *Cancer* 24 (2003), pp. 401–409.

[5] Owen N, Healy GN, Matthews CE, and Dunstan DW. “Too much sitting: the population health science of sedentary behavior.” In: *Exerc Sport Sci Rev.* 38(3):105-113 (2010).

[6] LaCroix AZ, Bellettiere John, Rillamas-Sun E, and et al. “Association of Light Physical Activity Measured by Accelerometry and Incidence of Coronary Heart Disease and Cardiovascular Disease in Older Women.” In: *JAMA Netw Open.* 2(3):105-113 (2019).

[7] Loprinzi PD and Sng E. “The effects of objectively measured sedentary behavior on all-cause mortality in a national sample of adults with diabetes.” In: *Prev Med.* 86:55-7.65 (2016).

[8] Biswas A, Oh PI, Faulkner GE, Bajaj RR, Silver MA, Mitchell MS, and et al. “Sedentary time and its association with risk for disease incidence, mortality”. In: *Ann Intern Med* 162(2) (2015).

[9] Hills AP, Mokhtar N, and Byrne NM. “Assessment of Physical Activity and Energy Expenditure: An Overview of Objective Measures”. In: *Front Nutr.* 1:5 (2014).

[10] Westerterp and Klaas R. “Physical activity and physical activity induced energy expenditure in humans: measurement, determinants, and effects.” In: *Frontiers in physiology* 4-90 (2013).

[11] Guinhouya CB, Hubert H, Soubrier S, Vilhelm C, Lemdani M, and Durocher A. “Moderate-to-vigorous physical activity among children: discrepancies in accelerometry-based cut-off points.” In: *Obesity (Silver Spring).* 14(5):774-7 (2006).

[12] Guinhouya CB, Lemdani M, Vilhelm C, Durocher A, and Hubert H. “Actigraph-defined moderate-to-vigorous physical activity cut-off points among children: statistical and biobehavioural relevance.” In: *Acta Paediatr.* 98(4):708-14 (2009).

[13] Park J. H., Moon J. H., Kim H. J., Kong M. H., and Oh Y. H. “Sedentary Lifestyle: Overview of Updated Evidence of Potential Health Risks”. In: *Korean journal of family medicine* 41(6) (2020).

[14] World Health Organization. *WHO guidelines on physical activity and sedentary behavior.* World Health Organization, 2020, viii, 93 p.

[15] Deborah Rohm Young, Marie-France Hivert, Sofiya Alhassan, Sarah M. Camhi, Jane F. Ferguson, Peter T. Katzmarzyk, Cora E. Lewis, Neville Owen, Cynthia K. Perry, Juned Siddique, and Celina M. Yong. “Sedentary Behavior and Cardiovascular Morbidity and Mortality: A Science Advisory From the American Heart Association.” In: *Circulation.* 134(13) (2016).

[16] John Bound and Alan B. Krueger. “The Extent of Measurement Error In Longitudinal Earnings Data: Do Two Wrongs Make A Right?” In: *Journal of Labor Economics.* 9(1) (1989), pp. 1–24.

[17] BE Ainsworth, CJ Caspersen, CE Matthews, LC Mâsse, T Baranowski, and W Zhu. “Recommendations to improve the accuracy of estimates of physical activity derived from self report.” In: *J Phys Act Health.* 103(14) (2012).

[18] Prentice, R.L., M. Pettinger, and G.L. Anderson. “Statistical issues arising in the Women’s Health Initiative.” In: *Biometrics.* 61(4) (2005).

[19] Kipnis V, Subar AF, Midthune D, Freedman LS, Ballard-Barbash R, Troiano RP, and Bingham S. “Structure of dietary measurement error: results of the OPEN biomarker study.” In: *Am J Epidemiol.* 158(1) (2003).

[20] Carroll R., D. Ruppert, and L. Stefanski. *Measurement Error in Nonlinear Models.* Monographs on Statistics and Applied Probability. London: Chapman Hall, 1995. ISBN: 9780471682271.

[21] Spiegelman D., B. Rosner, and R. Logan. “Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs.” In: *Journal of the American Statistical Association.* 95(449) (2000).

[22] Spiegelman D., S. Schneeweiss, and A. McDermott. “Measurement error correction for logistic regression models with an ’alloyed gold standard’.” In: *American Journal of Epidemiology.* 145(2) (1997).

[23] Thiébaut AC, Kipnis V, Schatzkin A, and Freedman LS. “The role of dietary measurement error in investigating the hypothesized link between dietary fat intake and breast cancer–a story with twists and turns.” In: *Cancer Invest.* 26(1) (2008).

[24] Loki Natarajan, Shirley W. Flatt, Xiaoying Sun, Anthony C. Gamst, and Jacqueline M. Major. “Validity and Systematic Error in Measuring Carotenoid Consumption with Dietary Self-report Instruments.” In: *Practice of Epidemiology.* 163(8) (2005).

[25] Loki Natarajan, Minya Pu, Juanjuan Fan, Richard A. Levine, Ruth E. Patterson, and Cynthia A. Thompson. “Measurement Error of Dietary Self-Report in Intervention Trials.” In: *American Journal of Epidemiology.* 172(7) (2010).

[26] Pietro Ferrari, Christine Friedenreich, and Charles E. Matthews. “The Role of Measurement Error in Estimating Levels of Physical Activity”. In: *American Journal of Epidemiology* 166(7), 832-840 (2007).

[27] Nusser, Nicholas K Beyler, Gregory J Welk, Alicia L Carriquiry, Wayne A Fuller, and M N King B. “Modeling errors in physical activity recall data”. In: *Journal of physical activity and health* 9.1 (2012), pp. 56–67.

[28] Nicholas Beyler, Susanne James-Burdumy, Martha Bleeker, Jane Fortson, and Max Benjamin. *Measurement Error Properties in an Accelerometer Sample of U.S. Elementary School Children.* Tech. rep. undated.

[29] Sungwoo Lim, Brett Wyker, Katherine Bartley, and Donna Eisenhower. “Measurement Error of Self-Reported Physical Activity Levels in New York City: Assessment and Correction”. In: *American Journal of Epidemiology* 181.9 (2015), pp. 648–655.

[30] George Agogo, Hilko van der Voet, Laura Trijsburg, Fred Eeuwijk, Pieter Veer, and Hendriek Boshuizen. “Measurement error modelling for accelerometer activity data using Bayesian integrated nested Laplace approximation”. In: July 2015.

[31] Sneha Jadhav, Carmen D. Tekwe, and Yuanyuan Luan. “A function-based approach to model the measurement error in wearable devices”. In: *Statistics in Medicine* 41.24 (2022), pp. 4886–4902.

[32] Morrell C.H., Brant L.J., Pearson J.D., Verbeke G., and Fleg J.L. “Applying linear mixed-effects models to the problem of measurement error in epidemiologic studies”. In: *Communications in Statistics: Simulation and Computation* 32, 437-459 (2003).

[33] Rosenberg DE, Walker R abd Greenwood-Hickman MA, and et al. “Device-assessed physical activity and sedentary behavior in a community-dwelling cohort of older adults.” In: *BMC Public Health.* 20:1256 (2020).

[34] Greenwood-Hickman, Mikael Anne, Nakandala Supun, Jankowska Marta M, and Natarajan Loki. “The CNN Hip Accelerometer Posture (CHAP) Method for Classifying Sitting Patterns from Hip Accelerometers.” In: *Medicine and Science in Sports and Exercise.* 18(4) (2021).

[35] HS An, Y Kim, and JM. Lee. “Accuracy of inclinometer functions of the activPAL and ActiGraph GT3X+: A focus on physical activity”. In: *Gait posture* 51 (2017).

[36] C Fischer, M Yıldırım, J Salmon, and M JM. Chinapaw. “Comparing different accelerometer cut-points for sedentary time in children.” In: *Pediatr Exerc Sci.* 24(2) (2012).

[37] J. M. Bland and D. G. Altman. “Calculating correlation coefficients with repeated observations: part 1–correlation within subjects”. In: *BMJ* 310 (1995).

[38] N M Laird and J H Ware. “Random-effects models for longitudinal data”. In: *Biometrics* 38(4) (1982).

[39] Greenwood-Hickman MA, Nakandala S, Jankowska MM, Rosenberg DE, Tuz-Zahra F, Bellettiere J, Carlson J, Hibbing PR, Zou J, Lacroix AZ, Kumar A, and Natarajan L. “The CNN Hip Accelerometer Posture (CHAP) Method for Classifying Sitting Patterns from Hip Accelerometers: A Validation Study”. In: *Med Sci Sports Exerc* 53(11) (2021).

[40] B. G. Tabachnick and L.S. Fidell. *Using multivariate statistics.* Pearson India Education Services, 2022.

[41] P. Wu, Y. Han, T. Chen, and X.M. Tu. “Causal inference for Mann-Whitney-Wilcoxon rank sum and other nonparametric statistics”. In: *Statistics in Medicine* 33(8) (2014).

[42] PR Rosenbaum. *Observational Studies.* Springer, 2002.

[43] L Mao. “On causal estimation using U-statistics”. In: *Biometrika* 105(1) (2018).

[44] Z. Zhang, S. Ma, C. Shen, and C. Liu. “Estimating Mann Whitney-type Causal Effects”. In: *International Statistical Review* 87(3) (2019).

[45] Tang W. and Tu XM. *Modern Clinical Trial Analysis*. Springer Science, New York, 2012.

[46] Tang W., He H., and Tu XM. *Applied Categorical Data Analysis*. Chapman & Hall/CRC, 2012.

[47] A.A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, New York, 2006.

[48] O. Thas, J. De Neve, L. Clement, and J. P. Ottoy. “Probabilistic index models”. In: *J. Roy. Statist. Soc. Ser. B* 74 (2012).

[49] J. Liu, X. Zhang, T. Chen, T. Wu, T. Lin, L. Jiang, S. Lang, L. Liu, L. Natarajan, J.X. Tu, T. Kosciolek, J. Morton, T.T Nguyen, B. Schnabl, R. Knight, C. Feng, Y. Zhong, and X.M. Tu. “A Semiparametric Model for Between-Subject Attributes: Applications to Beta-diversity of Microbiome Data”. In: *Biometrics* 78 (2021).

[50] Kowalski J. and Tu X.M. *Modern Applied U Statistics*. Wiley, New York, 2007.

[51] Spearman C. “The proof and measurement of association between two things”. In: *Amer. J. Psychol* 15 (1904).

[52] T. Lin, T. Chen, J. Liu, and X.M. Tu. “Extending the MWW rank sum test to survey data for comparing mean ranks”. In: *Statistics in Medicine* 40 (2021).

[53] Y. Ma, W. Tang, C. Feng, and X.M. Tu. “Inference for kappas for longitudinal study data: Applications to sexual health research”. In: *Biometrics* 64 (2008).

[54] Y. Ma, W. Tang, Q. Yu, and X.M. Tu. “Modeling Concordance Correlation Coefficient for longitudinal study data”. In: *Psychometrika* 75 (2010).

[55] R. Chen, T. Chen, N. Lu, H. Zhang, P. Wu, C. Feng, and X.M. Tu. "Extending the Mann-Whitney-Wilcoxon Rank Sum Test to longitudinal regression analysis". In: *Applied Statistics* 41(12) (2014).

[56] T. Chen, J. Kowalski, R. Chen, P. Wu, H. Zhang, C. Feng, and X.M. Tu. "Rank-preserving regression for longitudinal data with missing responses". In: *Statistics in Medicine* 35 (2016).

[57] N. Lu, A.M. White, P. Wu, H. He, J. Hu, C. Feng, and X.M. Tu. *Social Networking: Recent Trends, Emerging Issues and Future Outlook.* Nova Science, New York, 2013.

[58] T.S. King and V.M. Chinchilli. "A generalized concordance correlation coefficient for continuous and categorical data". In: *Statistics in Medicine* 20 (2001).

[59] Y. Ma. "On inference for Kendall's tau within a longitudinal data setting". In: *Journal of Applied Statistics* 39 (2012).

[60] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: `https://www.R-project.org/`.

[61] R. E. Patterson, C. R. Marinac, L. Natarajan, S. J. Hartman, L. Cadmus-Bertram, S. W. Flatt, H. Li, B. Parker, Oratowski-Coleman, A. J. Villaseñor, S. Godbole, and J. Kerr. "Recruitment strategies, design, and participant characteristics in a trial of weight-loss and metformin in breast cancer survivors". In: *Contemp Clin Trials* 47 (2016).

[62] Hartman SJ, Nelson SH, Marinac CR, Natarajan L, Parker BA, and Patterson RE. "The effects of weight loss and metformin on cognition among breast cancer survivors: Evidence from the Reach for Health study". In: *Psychooncology* 28(8):1640-1646 (2019).

Appendix A

Additional simulation results for Chapter 2

Table A.1: Comparisons of estimates of βs under different methods (asymptotic standard error/empirical standard errors) for continuous outcome simulation study under different sample size and correlation structures.

	β_0 (10)			β_x (2.95)			β_c (3)		
	n = 50	n = 100	n = 500	n = 50	n = 100	n = 500	n = 50	n = 100	n = 500
$\rho = 0.1$ and $\rho_{xc} = 0$									
BLUP$_1$	10.030 (0.650/ 0.665)	10.026 (0.329/ 0.332)	9.986 (0.204/ 0.226)	2.947 (0.114/ 0.115)	2.952 (0.058/ 0.058)	2.953 (0.036/ 0.039)	3.002 (0.227/ 0.223)	2.996 (0.114/ 0.105)	3.000 (0.070/ 0.064)
W$_1$	10.539 (0.648/ 0.658)	10.641 (0.318/ 0.305)	10.592 (0.198/ 0.212)	2.837 (0.113/ 0.117)	2.828 (0.055/ 0.053)	2.834 (0.034/ 0.035)	3.001 (0.228/ 0.223)	2.996 (0.114/ 0.105)	3.000 (0.070/ 0.063)
BLUP$_2$	9.974 (0.672/ 0.729)	9.981 (0.326/ 0.291)	10.031 (0.204/ 0.181)	2.953 (0.119/ 0.133)	2.956 (0.057/ 0.052)	2.944 (0.036/ 0.033)	3.002 (0.228/ 0.224)	2.998 (0.111/ 0.104)	2.998 (0.070/ 0.070)
W$_2$	8.633 (0.533/ 0.571)	8.719 (0.257/ 0.207)	8.696 (0.161/ 0.146)	1.464 (0.043/ 0.043)	1.460 (0.021/ 0.018)	1.460 (0.013/ 0.012)	3.001 (0.172/ 0.173)	2.998 (0.084/ 0.080)	2.999 (0.053/ 0.047)
$\rho = 0.3$ and $\rho_{xc} = 0$									
BLUP$_1$	10.055 (0.741/ 0.823)	9.967 (0.360/ 0.364)	9.985 (0.228/ 0.241)	2.949 (0.130/ 0.139)	2.952 (0.063/ 0.068)	2.955 (0.040/ 0.042)	2.997 (0.254/ 0.237)	3.002 (0.123/ 0.124)	2.998 (0.077/ 0.078)
W$_1$	10.905 (0.707/ 0.746)	10.816 (0.344/ 0.324)	10.809 (0.218/ 0.220)	2.778 (0.122/ 0.123)	2.782 (0.059/ 0.060)	2.784 (0.038/ 0.037)	2.996 (0.254/ 0.236)	3.003 (0.123/ 0.124)	2.998 (0.077/ 0.078)
BLUP$_2$	9.980 (0.752/ 0.871)	9.974 (0.366/ 0.418)	9.994 (0.228/ 0.260)	2.951 (0.132/ 0.148)	2.955 (0.064/ 0.072)	2.955 (0.040/ 0.045)	2.999 (0.250/ 0.263)	2.997 (0.124/ 0.131)	2.998 (0.078/ 0.083)
W$_2$	8.828 (0.554/ 0.553)	8.747 (0.269/ 0.327)	8.756 (0.169/ 0.182)	1.453 (0.045/ 0.044)	1.452 (0.022/ 0.027)	1.453 (0.014/ 0.015)	2.992 (0.182/ 0.166)	3.001 (0.088/ 0.093)	2.995 (0.055/ 0.058)
$\rho = 0.1$ and $\rho_{xc} = 0.5$									
BLUP$_1$	9.932 (0.619/ 0.686)	10.035 (0.305/ 0.302)	10.029 (0.193/ 0.210)	2.920 (0.132/ 0.150)	2.929 (0.065/ 0.066)	2.942 (0.041/ 0.043)	3.196 (0.259/ 0.296)	3.169 (0.127/ 0.139)	3.162 (0.080/ 0.079)
W$_1$	10.564 (0.594/ 0.611)	10.648 (0.293/ 0.279)	10.637 (0.186/ 0.190)	2.793 (0.126/ 0.132)	2.787 (0.062/ 0.061)	2.790 (0.039/ 0.039)	3.196 (0.259/ 0.296)	3.169 (0.127/ 0.139)	3.162 (0.080/ 0.079)
BLUP$_2$	9.951 (0.620/ 0.695)	10.013 (0.307/ 0.313)	10.027 (0.193/ 0.190)	2.935 (0.134/ 0.122)	2.937 (0.065/ 0.065)	2.944 (0.041/ 0.041)	3.138 (0.259/ 0.263)	3.155 (0.126/ 0.127)	3.158 (0.080/ 0.078)
W$_2$	8.679 (0.496/ 0.459)	8.709 (0.245/ 0.235)	8.713 (0.154/ 0.151)	1.450 (0.048/ 0.045)	1.454 (0.024/ 0.023)	1.454 (0.015/ 0.015)	3.193 (0.196/ 0.215)	3.244 (0.096/ 0.099)	3.242 (0.060/ 0.060)
$\rho = 0.3$ and $\rho_{xc} = 0.5$									
BLUP$_1$	9.976 (0.705/ 0.730)	10.039 (0.337/ 0.351)	10.041 (0.214/ 0.238)	2.922 (0.149/ 0.148)	2.933 (0.072/ 0.075)	2.947 (0.045/ 0.051)	3.238 (0.288/ 0.281)	3.221 (0.140/ 0.144)	3.217 (0.088/ 0.087)
W$_1$	10.870 (0.665/ 0.642)	10.864 (0.319/ 0.315)	10.875 (0.202/ 0.215)	2.725 (0.140/ 0.130)	2.732 (0.067/ 0.068)	2.731 (0.043/ 0.046)	3.238 (0.288/ 0.281)	3.221 (0.139/ 0.144)	3.217 (0.088/ 0.087)
BLUP$_2$	9.989 (0.695/ 0.723)	10.045 (0.338/ 0.350)	10.053 (0.214/ 0.221)	2.931 (0.148/ 0.151)	2.939 (0.072/ 0.075)	2.951 (0.045/ 0.047)	3.213 (0.287/ 0.295)	3.228 (0.139/ 0.134)	3.219 (0.088/ 0.086)
W$_2$	8.797 (0.529/ 0.521)	8.789 (0.257/ 0.256)	8.787 (0.162/ 0.164)	1.445 (0.051/ 0.052)	1.446 (0.025/ 0.025)	1.446 (0.016/ 0.016)	3.258 (0.207/ 0.208)	3.256 (0.101/ 0.104)	3.256 (0.063/ 0.066)

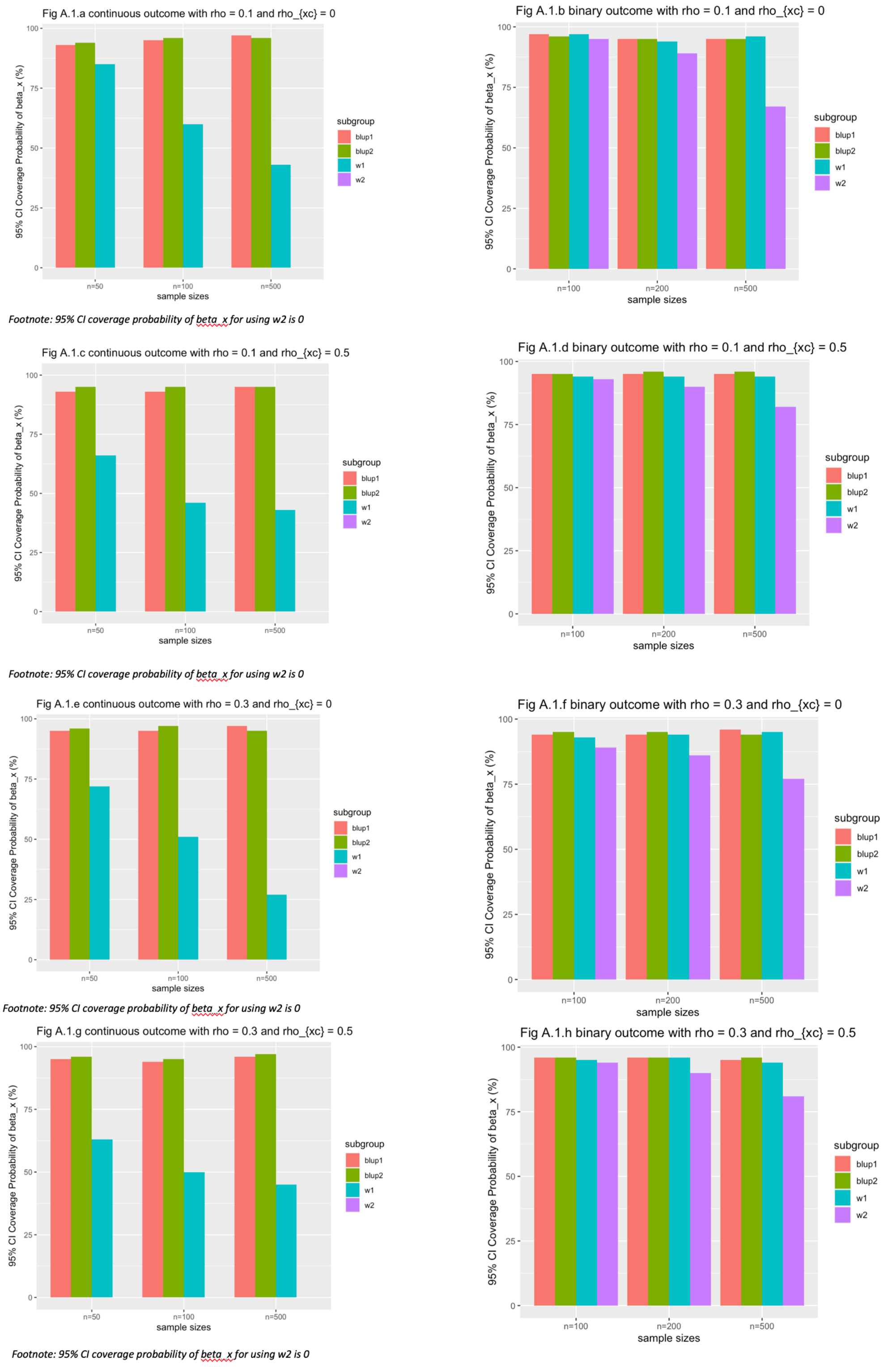


Fig A.1.a continuous outcome with rho = 0.1 and rho_{xc} = 0

Footnote: 95% CI coverage probability of beta_x for using w2 is 0

Fig A.1.b binary outcome with rho = 0.1 and rho_{xc} = 0

Fig A.1.c continuous outcome with rho = 0.1 and rho_{xc} = 0.5

Footnote: 95% CI coverage probability of beta_x for using w2 is 0

Fig A.1.d binary outcome with rho = 0.1 and rho_{xc} = 0.5

Fig A.1.e continuous outcome with rho = 0.3 and rho_{xc} = 0

Footnote: 95% CI coverage probability of beta_x for using w2 is 0

Fig A.1.f binary outcome with rho = 0.3 and rho_{xc} = 0

Fig A.1.g continuous outcome with rho = 0.3 and rho_{xc} = 0.5

Footnote: 95% CI coverage probability of beta_x for using w2 is 0

Fig A.1.h binary outcome with rho = 0.3 and rho_{xc} = 0.5

Figure A.1: 95 % Confidence Interval Coverage Probability of estimated β_x using different methods for continuous outcome and binary outcome

Table A.2: Comparisons of estimates of βs under different methods (asymptotic standard error/empirical standard errors) for binary outcome simulation study under different sample size and correlation structures.

$\rho = 0.1$ and $\rho_{xc} = 0$									
	β_0 (0.1)			β_x (0.1)			β_c (0.1)		
	n = 100	n = 200	n = 500	n = 100	n = 200	n = 500	n = 100	n = 200	n = 500
BLUP$_1$	0.108 (0.350/ 0.349)	0.097 (0.260/ 0.274)	0.097 (0.120/ 0.117)	0.099 (0.190/ 0.183)	0.106 (0.160/ 0.167)	0.105 (0.074/ 0.071)	0.105 (0.212/ 0.225)	0.109 (0.147/ 0.150)	0.102 (0.068/ 0.063)
W$_1$	0.083 (0.379/ 0.376)	0.113 (0.246/ 0.258)	0.112 (0.114/ 0.111)	0.124 (0.239/ 0.231)	0.090 (0.136/ 0.141)	0.089 (0.063/ 0.061)	0.105 (0.212/ 0.225)	0.109 (0.147/ 0.150)	0.102 (0.068/ 0.063)
BLUP$_2$	0.109 (0.359/ 0.378)	0.098 (0.261/ 0.278)	0.096 (0.121/ 0.120)	0.103 (0.195/ 0.204)	0.104 (0.159/ 0.165)	0.101 (0.074/ 0.075)	0.105 (0.213/ 0.226)	0.108 (0.149/ 0.151)	0.102 (0.068/ 0.063)
W$_2$	0.030 (0.420/ 0.426)	0.053 (0.297/ 0.315)	0.058 (0.138/ 0.139)	0.059 (0.101/ 0.103)	0.050 (0.072/ 0.071)	0.048 (0.033/ 0.034)	0.106 (0.212/ 0.225)	0.109 (0.147/ 0.149)	0.102 (0.068/ 0.063)
$\rho = 0.3$ and $\rho_{xc} = 0$									
BLUP$_1$	0.103 (0.378/ 0.402)	0.111 (0.262/ 0.278)	0.099 (0.164/ 0.162)	0.098 (0.238/ 0.263)	0.103 (0.164/ 0.180)	0.096 (0.102/ 0.102)	0.107 (0.213/ 0.222)	0.097 (0.147/ 0.149)	0.098 (0.092/ 0.092)
W$_1$	0.120 (0.349/ 0.368)	0.131 (0.243/ 0.256)	0.118 (0.152/ 0.151)	0.071 (0.190/ 0.208)	0.083 (0.132/ 0.145)	0.077 (0.082/ 0.081)	0.107 (0.213/ 0.222)	0.097 (0.147/ 0.149)	0.098 (0.092/ 0.092)
BLUP$_2$	0.104 (0.373/ 0.398)	0.094 (0.262/ 0.265)	0.102 (0.164/ 0.165)	0.093 (0.240/ 0.281)	0.103 (0.165/ 0.169)	0.103 (0.102/ 0.106)	0.107 (0.213/ 0.222)	0.097 (0.147/ 0.151)	0.098 (0.092/ 0.092)
W$_2$	0.060 (0.425/ 0.452)	0.071 (0.295/ 0.308)	0.052 (0.185/ 0.186)	0.043 (0.103/ 0.133)	0.048 (0.071/ 0.074)	0.047 (0.044/ 0.045)	0.107 (0.213/ 0.223)	0.097 (0.147/ 0.149)	0.098 (0.092/ 0.092)
$\rho = 0.1$ and $\rho_{xc} = 0.5$									
BLUP$_1$	0.103 (0.316/ 0.329)	0.097 (0.229/ 0.230)	0.097 (0.143/ 0.146)	0.098 (0.262/ 0.274)	0.095 (0.180/ 0.181)	0.097 (0.112/ 0.115)	0.115 (0.239/ 0.240)	0.115 (0.165/ 0.173)	0.116 (0.104/ 0.110)
W$_1$	0.088 (0.333/ 0.347)	0.112 (0.218/ 0.219)	0.088 (0.137/ 0.139)	0.083 (0.221/ 0.231)	0.081 (0.153/ 0.153)	0.082 (0.096/ 0.098)	0.115 (0.239/ 0.240)	0.115 (0.165/ 0.173)	0.116 (0.104/ 0.110)
BLUP$_2$	0.102 (0.323/ 0.331)	0.099 (0.230/ 0.231)	0.095 (0.144/ 0.149)	0.101 (0.261/ 0.271)	0.096 (0.180/ 0.181)	0.101 (0.113/ 0.117)	0.115 (0.239/ 0.240)	0.115 (0.165/ 0.173)	0.116 (0.104/ 0.110)
W$_2$	0.053 (0.380/ 0.390)	0.060 (0.263/ 0.266)	0.043 (0.164/ 0.164)	0.046 (0.119/ 0.118)	0.046 (0.083/ 0.083)	0.048 (0.052/ 0.052)	0.115 (0.243/ 0.244)	0.115 (0.168/ 0.175)	0.118 (0.105/ 0.113)
$\rho = 0.3$ and $\rho_{xc} = 0.5$									
BLUP$_1$	0.095 (0.318/ 0.351)	0.096 (0.232/ 0.236)	0.099 (0.145/ 0.153)	0.094 (0.266/ 0.284)	0.105 (0.184/ 0.185)	0.097 (0.115/ 0.118)	0.123 (0.237/ 0.254)	0.113 (0.164/ 0.164)	0.114 (0.103/ 0.106)
W$_1$	0.079 (0.313/ 0.312)	0.113 (0.216/ 0.221)	0.109 (0.136/ 0.137)	0.085 (0.225/ 0.279)	0.084 (0.157/ 0.157)	0.078 (0.092/ 0.094)	0.123 (0.237/ 0.254)	0.113 (0.164/ 0.164)	0.114 (0.103/ 0.106)
BLUP$_2$	0.097 (0.320/ 0.359)	0.098 (0.233/ 0.240)	0.096 (0.145/ 0.145)	0.102 (0.269/ 0.291)	0.103 (0.185/ 0.185)	0.095 (0.115/ 0.115)	0.123 (0.237/ 0.254)	0.113 (0.164/ 0.164)	0.115 (0.103/ 0.106)
W$_2$	0.021 (0.372/ 0.377)	0.040 (0.261/ 0.266)	0.051 (0.163/ 0.164)	0.048 (0.118/ 0.118)	0.050 (0.082/ 0.083)	0.048 (0.051/ 0.051)	0.122 (0.243/ 0.259)	0.116 (0.167/ 0.170)	0.115 (0.105/ 0.109)

Table A.3: Comparisons of relative bias and coverage probability for estimates of βs under different methods for continuous outcome simulation study under different sample size and correlation structures.

$\rho = 0.1$ and $\rho_{xc} = 0$

	β_0			β_x			β_c		
	n = 50	n = 100	n = 500	n = 50	n = 100	n = 500	n = 50	n = 100	n = 500
relative bias BLUP$_1$	0.30%	0.26%	0.14%	0.10%	0.07%	0.10%	0.06%	0.13%	0
coverage BLUP$_1$	89%	95%	97%	93%	95%	97%	96%	97%	99%
relative bias W$_1$	5.39%	6.41%	5.92%	3.83%	4.14%	3.93%	0.03%	0.13%	0
coverage W$_1$	85%	70%	11%	85%	60%	43%	95%	96%	99%
relative bias BLUP$_2$	0.26%	0.19%	0.31%	0.10%	0.20%	0.20%	0.06%	0.06%	0.06%
coverage BLUP$_2$	95%	94%	97%	94%	96%	96%	93%	97%	99%
relative bias W$_2$	13.67%	12.81%	13.04%	50.37%	50.51%	50.51%	0.03%	0.06%	0.03%
coverage W$_2$	25%	8%	0	0	0	0	97%	95%	99%
$\rho = 0.3$ and $\rho_{xc} = 0$									
relative bias BLUP$_1$	0.55%	0.33%	0.15%	0.03%	0.07%	0.17%	0.10%	0.06%	0.06%
coverage BLUP$_1$	90%	97%	94%	95%	95%	97%	96%	97%	95%
relative bias W$_1$	9.05%	8.16%	8.09%	5.83%	5.69%	5.63%	0.13%	0.10%	0.06%
coverage W$_1$	80%	63%	18%	72%	51%	27%	95%	96%	94%
relative bias BLUP$_2$	0.20%	0.26%	0.06%	0.03%	0.17%	0.17%	0.03%	0.10%	0.06%
coverage BLUP$_2$	92%	94%	95%	96%	97%	95%	96%	97%	96%
relative bias W$_2$	11.72%	12.53%	12.44%	50.75%	50.78%	50.75%	0.27%	0.03%	0.17%
coverage W$_2$	39%	9%	0	0	0	0	93%	96%	98%
$\rho = 0.1$ and $\rho_{xc} = 0.5$									
relative bias BLUP$_1$	0.68%	0.35%	0.29%	1.02%	0.71%	0.27%	6.53%	5.63%	5.40%
coverage BLUP$_1$	95%	93%	94%	93%	93%	95%	86%	85%	79%
relative bias W$_1$	5.64%	6.48%	6.37%	5.32%	5.53%	5.42%	6.53%	5.63%	5.40%
coverage W$_1$	74%	58%	29%	66%	46%	43%	83%	81%	77%
relative bias BLUP$_2$	0.49%	0.13%	0.27%	0.51%	0.44%	0.20%	4.60%	5.17%	5.27%
coverage BLUP$_2$	96%	94%	95%	95%	95%	95%	85%	84%	81%
relative bias W$_2$	13.21%	12.91%	12.87%	50.85%	50.71%	50.71%	6.43%	8.13%	8.07%
coverage W$_2$	20%	19%	0	0	0	0	88%	89%	80%
$\rho = 0.3$ and $\rho_{xc} = 0.5$									
relative bias BLUP$_1$	0.24%	0.39%	0.41%	0.95%	0.58%	0.44%	7.93%	7.37%	7.23%
coverage BLUP$_1$	95%	94%	96%	95%	94%	96%	87%	83%	76%
relative bias W$_1$	8.70%	8.64%	8.75%	7.63%	7.39%	7.42%	7.93%	7.37%	7.23%
coverage W$_1$	69%	61%	31%	63%	50%	45%	86%	80%	76%
relative bias BLUP$_2$	0.11%	0.45%	0.53%	0.64%	0.37%	0.31%	7.10%	7.60%	7.30%
coverage BLUP$_2$	94%	93%	95%	96%	95%	97%	88%	82%	77%
relative bias W$_2$	12.03%	12.11%	12.13%	51.02%	50.98%	50.98%	8.60%	8.53%	8.53%
coverage W$_2$	27%	12%	0	0	0	0	90%	86%	78%

Table A.4: Comparisons of relative bias and coverage probability for estimates of βs under different methods for binary outcome simulation study under different sample size and correlation structures.

$\rho = 0.1$ and $\rho_{xc} = 0$

	β_0			β_x			β_c		
	n = 100	n = 200	n = 500	n = 100	n = 200	n = 500	n = 100	n = 200	n = 500
relative bias BLUP$_1$	7.81%	3.00%	3.23%	0.80%	5.69%	5.12%	5.36%	9.33%	1.97%
coverage BLUP$_1$	96%	94%	96%	97%	95%	95%	95%	96%	96%
relative bias W$_1$	17.36%	13.21%	12.40%	24.38%	10.15%	10.50%	5.36%	9.32%	1.97%
coverage W$_1$	96%	94%	95%	97%	94%	96%	95%	95%	96%
relative bias BLUP$_2$	9.00%	2.00%	4.00%	3.00%	4.00%	1.00%	5.36%	8.00%	1.97%
coverage BLUP$_2$	95%	95%	96%	96%	95%	95%	95%	95%	96%
relative bias W$_2$	70.41%	47.41%	41.88%	41.18%	49.85%	52.08%	6.07%	9.04%	2.00%
coverage W$_2$	96%	94%	94%	95%	89%	67%	94%	95%	96%
$\rho = 0.3$ and $\rho_{xc} = 0$									
relative bias BLUP$_1$	3.11%	11.20%	0.87%	2.00%	2.53%	4.54%	7.23%	2.79%	2.30%
coverage BLUP$_1$	95%	94%	95%	94%	94%	96%	96%	97%	96%
relative bias W$_1$	20.10%	31.14%	17.76%	28.97%	17.38%	23.17%	7.26%	2.77%	2.30%
coverage W$_1$	95%	94%	95%	93%	94%	95%	96%	97%	95%
relative bias BLUP$_2$	4.00%	5.93%	1.80%	7.00%	2.54%	3.81%	7.23%	2.78%	2.30%
coverage BLUP$_2$	95%	94%	95%	95%	95%	94%	96%	97%	96%
relative bias W$_2$	39.59%	29.18%	47.61%	56.62%	52.39%	52.59%	6.80%	2.84%	2.40%
coverage W$_2$	95%	94%	94%	89%	86%	77%	95%	97%	96%
$\rho = 0.1$ and $\rho_{xc} = 0.5$									
relative bias BLUP$_1$	3.19%	2.76%	2.58%	2.06%	4.63%	3.38%	15.11%	14.71%	16.25%
coverage BLUP$_1$	94%	96%	95%	95%	95%	95%	96%	96%	94%
relative bias W$_1$	12.48%	11.57%	12.37%	17.39%	18.89%	18.42%	15.11%	14.70%	16.26%
coverage W$_1$	94%	95%	95%	94%	94%	94%	96%	95%	94%
relative bias BLUP$_2$	2.11%	0.25%	4.86%	1.16%	4.26%	1.42%	15.11%	14.71%	16.25%
coverage BLUP$_2$	95%	96%	95%	95%	96%	96%	96%	96%	94%
relative bias W$_2$	47.11%	39.66%	57.32%	53.97%	54.10%	52.02%	9.50%	14.69%	18.13%
coverage W$_2$	95%	95%	94%	93%	90%	82%	96%	96%	94%
$\rho = 0.3$ and $\rho_{xc} = 0.5$									
relative bias BLUP$_1$	5.06%	4.12%	1.07%	6.01%	4.95%	2.68%	22.76%	12.93%	14.07%
coverage BLUP$_1$	96%	95%	95%	96%	96%	95%	96%	96%	96%
relative bias W$_1$	20.62%	13.32%	8.57%	15.04%	15.55%	21.78%	22.81%	12.95%	14.06%
coverage W$_1$	96%	95%	96%	95%	96%	94%	96%	96%	95%
relative bias BLUP$_2$	3.19%	2.61%	3.87%	2.35%	3.93%	5.26%	22.76%	12.93%	14.92%
coverage BLUP$_2$	96%	96%	96%	96%	96%	96%	96%	96%	96%
relative bias W$_2$	78.74%	60.32%	48.57%	52.00%	50.26%	51.84%	22.43%	16.38%	15.08%
coverage W$_2$	96%	95%	94%	94%	90%	81%	95%	95%	94%

Appendix B

Proof of theorems in Chapter 3

1. Details of inferences in FRM equation (3.14)

It is readily checked $V_{\mathbf{i}1}$ and $V_{\mathbf{i}2}$ in (3.14) are as below:

$$\begin{aligned}
V_{\mathbf{i}1} &= Var(f_{\mathbf{i}1} \mid \mathbf{w_i}) = \frac{1}{4}[\pi_i\left(1-\pi_i\right)+\pi_j\left(1-\pi_j\right)] \\
V_{\mathbf{i}2} &= Var(f_{\mathbf{i}2} \mid \mathbf{w_i}) = \frac{1}{4}[g\left(\mathbf{w_i};\boldsymbol{\gamma}\right)\left(1-g\left(\mathbf{w_i};\boldsymbol{\gamma}\right)\right)+g\left(\mathbf{w_{i^c}};\boldsymbol{\gamma}\right)\left(1-g\left(\mathbf{w_{i^c}};\boldsymbol{\gamma}\right)\right)],
\end{aligned}$$

To simplify the evaluation of $V_{\mathbf{i}3}$ in (3.14), let:

$$\begin{aligned}
\pi_{\mathbf{i}} = \pi_i(1-\pi_j), \pi_{\mathbf{j}} = \pi_j(1-\pi_i) \\
r_{\mathbf{i}} = z_i(1-z_j), r_{\mathbf{j}} = z_j(1-z_i)
\end{aligned}$$

Given:

$$f_{\mathbf{i}3} = \frac{1}{2}[\frac{r_{\mathbf{i}}}{\pi_{\mathbf{i}}}I(y_{i1}\leq y_{j0})+\left(1-\frac{r_{\mathbf{i}}}{\pi_{\mathbf{i}}}\right)g\left(\mathbf{w_i},\boldsymbol{\gamma}\right)+\frac{r_{\mathbf{j}}}{\pi_{\mathbf{j}}}I(y_{j1}\leq y_{i0})+\left(1-\frac{r_{\mathbf{j}}}{\pi_{\mathbf{j}}}\right)g\left(\mathbf{w_{i^c}},\boldsymbol{\gamma}\right)],$$

It follows from the iterated conditional expectation that:

$$\begin{aligned}
V_{\mathbf{i}3} &= Var(f_{\mathbf{i}3} \mid \mathbf{w}_{\mathbf{i}}) = E[f_{\mathbf{i}3}^2 \mid \mathbf{w}_{\mathbf{i}}] - (E[f_{\mathbf{i}3} \mid \mathbf{w}_{\mathbf{i}}])^2 \\
&= \frac{1}{4}(E[\frac{r_{\mathbf{i}}^2}{\pi_{\mathbf{i}}^2}(I(y_{i1} \leq y_{j0}) - g(\mathbf{w}_{\mathbf{i}}, \gamma))^2 | \mathbf{w}_{\mathbf{i}}] + E[\frac{r_{\mathbf{j}}^2}{\pi_{\mathbf{j}}^2}(I(y_{j1} \leq y_{i0}) - g(\mathbf{w}_{\mathbf{i}^c}, \gamma))^2 | \mathbf{w}_{\mathbf{i}}]) \\
&= \frac{1}{4} E\left[\frac{r_{\mathbf{i}}^2}{\pi_{\mathbf{i}}^2} g\left(\mathbf{w}_{\mathbf{i}}, \boldsymbol{\gamma}\right)\left(1 - g\left(\mathbf{w}_{\mathbf{i}}, \boldsymbol{\gamma}\right)\right) + \frac{r_{\mathbf{j}}^2}{\pi_{\mathbf{j}}^2} g\left(\mathbf{w}_{\mathbf{i}^c}, \boldsymbol{\gamma}\right)\left(1 - g\left(\mathbf{w}_{\mathbf{i}^c}, \boldsymbol{\gamma}\right)\right) | \mathbf{w}_{\mathbf{i}}\right] \\
&= \frac{1}{4}\left[\frac{1}{\pi_i(1-\pi_j)} g\left(\mathbf{w}_{\mathbf{i}}, \boldsymbol{\gamma}\right)\left(1 - g\left(\mathbf{w}_{\mathbf{i}}, \boldsymbol{\gamma}\right)\right) + \frac{1}{\pi_j(1-\pi_i)} g\left(\mathbf{w}_{\mathbf{i}^c}, \boldsymbol{\gamma}\right)\left(1 - g\left(\mathbf{w}_{\mathbf{i}^c}, \boldsymbol{\gamma}\right)\right)\right].
\end{aligned}$$

2. Proof of Theorem 1.

Without loss of generality, consider the normalized quantity $\binom{n}{2}^{-1}\mathbf{U}_n = \binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} \mathbf{D}_{\mathbf{i}} \mathbf{V}_{\mathbf{i}}^{-1} \mathbf{S}_{\mathbf{i}}$ and continue to denote the normalized quantity as $\mathbf{U}_n$. By a Taylor series expansion, we have:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \left(-\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{U}_n\right)^{-\top} \{\sqrt{n}\mathbf{U}_n - \left(\frac{\partial}{\partial \boldsymbol{\alpha}} \mathbf{U}_n\right)^{\top} \sqrt{n}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right)\} + \mathbf{o}_p(1),$$

where

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{U}_n = \mathbf{B}^{\top} + \mathbf{o}_p(1), \quad \frac{\partial}{\partial \boldsymbol{\alpha}} \mathbf{U}_n = \mathbf{o}_p(1),$$

It follows from properties of multivariate U-statistics (Kowalski and Tu, 2007) that:

$$\sqrt{n}\mathbf{U}_n = \sqrt{n}\frac{2}{n}\sum_{i=1}^{n} E\left(\mathbf{U}_{n,\mathbf{i}} \mid \mathbf{y}_i, \mathbf{x}_i\right) + \mathbf{o}_p(1) = \sqrt{n}\frac{2}{n}\sum_{i=1}^{n} \mathbf{v}_i + \mathbf{o}_p(1).$$

It follows that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = -\mathbf{B}^{-\top}\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left(2\mathbf{v}_i\right) + \mathbf{o}_p(1) \to_d N\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right).$$

,where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is given in (3.17).