Spring 5-13-2023

Contributions to Causal Inference in Observational Studies

Jenny Park
Southern Methodist University, jennyp@smu.edu

Daniel F. Heitjan
dheitjan@smu.edu

Christy Boling Turer
christy.turer@utsouthwestern.edu

CONTRIBUTIONS TO CAUSAL INFERENCE IN OBSERVATIONAL STUDIES

Approved by:

Daniel F. Heitjan, PhD
Professor and Chair, Department of Statistical Science, SMU

Jing Cao, PhD
Professor, Department of Statistical Science, SMU

Lynne Stokes, PhD
Professor, Department of Statistical Science, SMU

Christy Boling Turer, MD
Associate Professor, Department of Interal Medicine, Department of Pediatrics, UT Southwestern Medical Center

CONTRIBUTIONS TO CAUSAL INFERENCE IN OBSERVATIONAL STUDIES

A Dissertation Presented to the Graduate Faculty of the

Dedman College

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Biostatistics

by

Jenny Park

BM, Piano Performance, Texas State University
MM, Piano Performance, University of Texas at Austin

May 13, 2023

Acknowledgments

I would like to thank my advisors Dr. Heitjan, for bringing me into the Biostatistics program and for his ongoing support during my five years in the program, and Dr. Turer for the research opportunities and mentorship.

I would also like to give special thanks to the committee members Dr. Stokes, for all her time spent outside the classroom to help me when I was struggling, and Dr. Cao for her humorous and passionate teaching style that kept me awake in class.

Last but not least, I would like to acknowledge my husband and my mother who have been my most loyal supporters in this journey.

Park, Jenny

BM, Piano Performance, Texas State University
MM, Piano Performance, University of Texas at Austin

Contributions to Causal Inference in Observational Studies

Advisor: Daniel F. Heitjan, PhD
Doctor of Philosophy degree conferred May 13, 2023
Dissertation completed April 26, 2023

The electronic health record (EHR) is a digital version of the patient chart. All clinically relevant patient information can be accessed from the EHR by professionals involved in the patient's care. For researchers, the EHR is a rich, convenient source for data to address a vast range of medical research questions.

In observational studies with EHR data, it is common to define the treatment/exposure status as a binary indicator reflecting whether patient was documented to receive a particular medication or procedure. The outcome can be any type of information on patient status documented in the EHR after the treatment has taken place.

The EHR, although not designed primarily for research, can serve as a platform for observational studies in clinical medicine. An advantage of the EHR is that it can document treatments unequivocally, provided the treatment – medication or procedure – appears in the record. For example, in a study in which treatment is the route of medication (intravenous=treated, oral=control), the EHR makes it clear which route was used. This does not, however, relieve the investigator from the responsibility of defining and measuring confounding variables, and properly adjusting for them in comparative analyses.

In Chapter 1, we demonstrate the use of longitudinal EHR data in an evaluation of the effects of treatment of 12,754 children with overweight/obesity in greater Dallas. Our objective in this study is to estimate the causal effect of clinician attention to elevated body

mass index (BMI), measured at up to 10 timepoints per child, on subsequent weight change. To account for bias from confounding, we use the propensity score stratification method, applied longitudinally at each timepoint. We specify the propensity score model to include baseline covariates, current values of time-varying covariates, and treatment status at the most recent visit.

An alternative method of causal inference when treatments are applied longitudinally in an observational study relies on the marginal structural model (MSM). When estimating an MSM, one eliminates confounding bias by constructing a series of propensity score models for treatment at each time, then weighting the subjects based on these scores. The MSM has the interpretation of a causal model for the effect of the series of treatments on the outcome.

Although MSMs are in wide use, there has been relatively little evaluation of the properties of model estimates in small samples. One can conduct a simulation study to assess properties such as the suitability of asymptotic approximations to moderate samples, best methods for computing the standard errors, choice of the weighting method, and robustness to incorrect assumptions about the MSM or the underlying propensity score model. Several simulation methods have been proposed, each with its pros and cons. In Chapter 2, we introduce a new, simplified simulation method that addresses the limitations of the existing methods. We demonstrate the use of our method in a Monte Carlo study to assess the properties of an estimated MSM involving treatment at two timepoints.

An oft-cited concern with MSMs is the sensitivity of model estimates to large weights. This issue arises in particular when there are multiple timepoints. As the number of timepoints increases, an individual's propensity score can become very small, while the estimation weights – defined as the inverse of the propensity score – becomes correspondingly large. Having a few subjects with large weights can result in an unstable estimate. In Chapter 3, we use the novel simulation method that we introduced in Chapter 2 to conduct a Monte Carlo assessment of the impact of large weights on the validity of MSM estimates. Finally,

we estimate a series of MSMs for the child obesity example from Chapter 1 and interpret the results in light of our simulation findings.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I dedicate this dissertation to my father Honju Park (1960-2004).

Chapter 1

The causal effect of clinician attention to high BMI in children across time

1.1. Background and significance Although childhood rates of overweight and obesity continue to rise, time allocated to address this complex issue during pediatric visits has not. In a median timespan of 20 minutes, pediatricians measure child weight, height, and vital signs, compute body mass index (BMI) to screen for overweight/obesity, perform an exam, communicate with children and their parents about a host of items, and prescribe vaccines, laboratory studies, and specialist referrals [1,2]. The American Academy of Pediatrics recommends annual BMI screening, healthy weight/lifestyle communication, and evaluation of health risks of overweight/obesity starting at age two years [3,4].

The longitudinal impact on relative BMI of pediatrician BMI communication during primary-care visits is unclear. Turer *et al* (2019) examined the association of pediatrician communication of high BMI and relative BMI improvement in a retrospective analysis of electronic health record (EHR) data from 6–12-year-olds followed in primary care [5]. Evidence that a clinician addressed weight management (termed "attention to BMI" or"BMI attention") was determined using billing codes for visit diagnoses and problem lists, and orders for laboratory tests, medication prescriptions, patient education, and referrals. The authors reported that, among 6–12-year-old children with overweight or obesity, high-BMI/comorbid-disease-risk communication was associated with a 20% increased likelihood of improvement in BMI at the next visit (adjusted odds ratio 1.2, 95% confidence interval (1.09, 1.28)).

Nevertheless, the causal impact of BMI communication on longitudinal BMI improvement remains unclear. Turer *et al*'s study applied random-effects regression analysis of repeated measures to evaluate the association of BMI attention and relative BMI improvement in

6–12-year-olds [5]. The analysis was not suited to examine the longitudinal effect of BMI communication on relative BMI over time nor whether the association was causal. Also lacking were data for children ages 2–5 and 13–18 years. Causal inference requires controlling for bias, commonly using methods based on the propensity score [6, 7]. The present study sought to estimate the longitudinal causal effect of attention to high BMI on relative BMI improvement in children across the guideline-recommended age spectrum of 2–18 years. Using methods designed for causal analysis, we estimated the causal impact of BMI attention on relative BMI improvement across time, accounting for prior BMI attention.

1.2. Materials and methods

1.2.1. Study cohort We conducted a retrospective study using 2009–2016 clinical practice data from children followed in pediatric primary care practices that shared a networked EHR hosted through Children's Medical Center Dallas. The University of Texas Southwestern Medical Center IRB approved the study with a waiver of informed consent. Primary care visits for children represent an opportunity to intervene on overweight and obesity in childhood. Population-based data indicate that 88% of children have a pediatric "usual source of care" and complete three outpatient visits/year [8]. We selected children with ≥ 2 primary-care visits at ages 2–18 years with (a) height and weight measures, (b) ≥ 60 days between the first and last visits, (c) BMI ≥ 25 or BMI centile ≥ 85 [9]; and ≥ 1 well-child/health-maintenance visit. We excluded children with diagnoses/conditions that might affect clinician-patient BMI communication or the validity of BMI measurement, including amputation, presence of a feeding tube, type 1 diabetes, pervasive development delay, and other chronic metabolic, congenital, or oncologic condition. Though not part of the initial study inclusion/exclusion criteria, children were required to have five separate height data points, as noted in the statistical methods section.

1.2.2. Data structure and measures

Visit and time structure In contrast to the 2019 study, here we followed children longitudinally with multiple (up to 10) high-BMI visits at which treatment (attention to BMI) could have been administered. We assessed the effect of attention to BMI at one "high-BMI visit" on BMI at a subsequent visit occurring from 1 to 30 months later. In contrast to the uniformity of follow-up intervals in randomized trials, we permitted a wide range of time intervals for each of the subsequent visits. US-based population data show no difference in pediatric outpatient-visit frequency by BMI despite recommendations for 3–6-month follow-up to readdress high BMI (frequency of visits/year for a child with a healthy weight is 2.9, vs. 2.8 for overweight, and 3.0 for a child with obesity) [8].

Attention to BMI We used an electronic phenotyping approach to define a binary attention-to-BMI variable that equaled 1 if there was EHR evidence that a clinician addressed BMI at a visit with a 2–18-year-old child with overweight/obesity [unpublished data, 2022]. The electronic phenotype sought evidence of guideline-recommended weight-management clinical practices using specific text associated with numeric International Classification of Diseases (ICD) codes (from billing and problem-list codes), education codes for primary-care obesity counseling, and referrals to nutrition, weight-management, and bariatric surgery. Evidence was used as a proxy for a clinician identifying high BMI (BMI $\geq$ 25 or centile $\geq$ 85) in children $\geq$ 2 years, communicating regarding high BMI/health risks, and providing lifestyle communication in primary care or via referral to nutrition, weight management, or weight-loss surgery. We previously reported that the electronic phenotype was 84.7% sensitive and 99.6% specific for detecting clinician attention to high BMI [unpublished data, 2022].

Covariates We identified clinic and patient covariates that were measured/documented in the EHR and relevant to BMI change. Then, we selected ones that were associated with both

treatment (BMI communication) and outcome (BMI change) [10,11]. Instrumental variables in the former study excluded from the present study were clinic location and visit type of the high-BMI visit, because we determined that these variables affected the outcome through treatment but did not affect the outcome in the absence of treatment. Finally, we categorized the selected covariates as fixed or time-varying (Table 1.1) for timepoint $t = 1, \ldots, 10$. When $t = 1$, covariates from timepoint $t - 1$ are set to 0. Fixed variables were measured at visit 1 and did not change over time. Time-varying variables were measured at every timepoint. Baseline variables were time-varying variables observed at timepoint 1.

Table 1.1: Covariates at time of visit t (when child eligible for BMI attention) in the causal analysis.

Type of covariate	Variable
Fixed, at visit 1	Sex
	Race/ethnicity
Time-varying,	Percent of BMI 95th percentile ($\%\text{BMI}_{\text{p95}}$)
observed at visit t.	Obesity category at visit t
	Age, continuous
variables from $t-1$	Age, categorical
When $t = 1$,	Interval length from visit $t - 1$
were set to 0.	Treatment status (attention to BMI) at visit $t - 1$
The value at $t = 1$	Reassessment status of treatment received at visit $t - 1$
is the.	Prescription medication associated with weight gain
baseline value.	Prescription medication associated with weight loss
	Well-child visit in past 12 months
	Sick visit in past 12 months
	Visit occurred between April and July

Outcome The primary study outcome was the change in percent of BMI 95^{th} percentile ($\%\text{BMI}_{\text{p95}}$) from baseline, which was used as a proxy for improvement in child adiposity.

Studies document that $\%BMI_{p95}$ more accurately reflects within- and between-child changes in adiposity compared to use of BMI z score or BMI percentile, particularly in analyses that include children across the BMI spectrum (from mild overweight to class 3 obesity) [12, 13]. We assessed the effect of treatment (attention to BMI) given at visit t, by subtracting a child's post-treatment $\%BMI_{p95}$ (observed 1–30 months after BMI attention at visit t) from the baseline $\%BMI_{p95}$ (observed at visit 1). We denoted this change in $\%BMI_{p95}$ as $\Delta\%BMI_{p95}$.

Height cleaning Some EHR height data are missing or implausible [14]. This presents a challenge because computation of BMI depends on height (BMI = weight [in kg]/height [in meters] squared). We addressed this issue using the following steps:

1. Imputed missing heights by converting height centiles at nearby visits to height at the visit missing height.
2. Removed implausible height values using methods published by Daymont *et al* [14].
3. Used ≥ 5 heights per child to fit separate monotone (nondecreasing) spline models.
4. Replaced observed and missing heights with model-fitted values.

1.2.3. Causal inference Causal inference — estimation of the effect of an intervention on an outcome — is straightforward in a randomized study because treatment assignment is unconfounded by design. In an observational study such as the present study, the mechanism by which treatment is assigned is unknown and likely to depend on factors that also affect treatment and outcome (for example, the severity of a child's obesity). Because our interest is in whether physician attention to BMI causes future weight change, we have applied methods designed specifically to adjust for confounders and estimate this causal effect.

1.2.4. Propensity score stratification analysis A common method to estimate a causal effect from observational data is propensity score (PS) stratification analysis [6]. In a longitudinal analysis like the present study, one conducts the following analysis at each timepoint:

1. Estimate each subject's PS (probability of receiving treatment at a visit) by constructing a logistic regression model. The observed treatment serves as the outcome and is regressed on all known confounders associated with both treatment and outcome at the visit [10, 11]. In the present study, the treatment is BMI attention, and known confounders are those listed in Table 1.1.

2. Sort children by their estimated PS values and group them into 10 strata (so that, within each stratum, treated and control groups possess similar distributions of child characteristics for unbiased causal effect estimation) [15, 16].

3. Apply criteria to verify balance across strata [7, 17]. Criteria used to determine if covariate balance was achieved across strata were: (1) an absolute weighted (by stratum size) standardized difference of covariates ≥ 0.25, and, (2) a treated-to-control weighted (by stratum size) variance ratio between 0.5 and 2 for continuous variables (e.g., age).

4. Average the treatment effects across strata (weighted by stratum size) to form a single estimate of the causal effect of attention to BMI [6, 17],

$$(\Delta\%\text{BMI}_{\text{p95, treated}} - \Delta\%\text{BMI}_{\text{p95, control}})_{\text{PS-adjusted}}.$$

We performed the above analysis at each of the ten timepoints. Separately, we computed the raw unadjusted effects of attention to BMI,

$$(\Delta\%\text{BMI}_{\text{p95, treated}} - \Delta\%\text{BMI}_{\text{p95, control}})_{\text{unadjusted}}.$$

An average treatment effect less than 0 denoted BMI attention improved $\Delta\%\text{BMI}_{\text{p95}}$ relative to $\Delta\%\text{BMI}_{\text{p95}}$ among controls. An effect greater than 0 denoted $\Delta\%\text{BMI}_{\text{p95}}$ worsened for those receiving BMI attention compared to controls. We conducted the analysis in R version 4.2.0.

1.2.5. Sensitivity analysis PS analysis is valid for causal inference assuming no unmeasured confounders. We conducted a sensitivity analysis [18] to test the impact of failing to include an unmeasured binary confounder, denoted U, by exploring varying degrees of U's influence on treatment assignment and outcome at each timepoint.

1.3. Results

Population Of 17,397 children aged 2–18 years with overweight or obesity followed in primary care, 13,036 (74.9%) had five more height datapoints. Of these, 12,574 (72.3%) had a subsequent visit to evaluate the impact of BMI communication on BMI improvement. The baseline BMI category of the 12,574 children was 65.0% with overweight, 29.9% with obesity class 1, and 5.2% with obesity class 2 or greater.

Covariate balance At each of the 10 timepoints and across strata, we achieved a balanced distribution of patient characteristics. At timepoint 1, for example (Table 1.2), the between-group difference in mean child age decreased from 1.5 years before PS stratification (std. diff, 0.39) to 0 years after PS stratification (std. diff, 0.01). Over half of the sample was younger than age six years or 12 years or older.

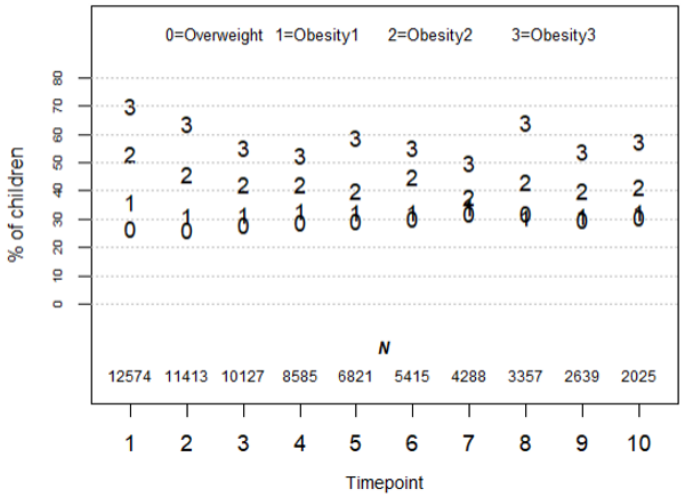
BMI attention by obesity class The proportion of children receiving BMI attention varied by BMI category (Figure 1.1). Children with obesity class 3 had the highest rates of attention starting at 70% for timepoint 1, with half or more of the children receiving BMI

Table 1.2: Covariate balance between control and treated groups before and after propensity score (PS) stratification (balance for $t = 1$ shown as an example; balance achieved at all 10 timepoints)

	Before PS stratification			After PS stratification		
Characteristic ($N = 12,574$)	Contr ($N = 8,653$)	Interv ($N = 3,921$)	Std. diff	Contr ($N = 8,653$)	Interv ($N = 3,921$)	Std. diff
Age in years	6.9	8.4	0.39	7.50	7.50	0.01
Age category (%)			0.41			0.11
2–5	47.6	29.5		42.5	40.2	
6–8	22.8	24.6		22.6	25.7	
9–11	16.3	24.7		18.1	20.2	
12–18	13.2	21.2		16.8	13.9	
Sex, female (%)	48.0	51.0	0.06	48.9	48.8	0.00
Race/ethnicity (%)			0.12			0.01
Black	19.2	22.5		20.1	20.6	
Hispanic	61.1	60.5		61.0	60.6	
White	12.8	9.6		11.9	11.7	
Other	6.9	7.3		7.0	7.1	
Percent of BMI_{p95}	101.4	110.0	0.55	104.0	104.3	0.06
Obesity class (overweight = 0)	1.3	1.5	0.32	1.4	1.4	0.01
Med with weight loss effect	1.6	1.6	0.00	1.6	1.6	0.00
Med with weight gain effect	11.0	10.3	0.03	10.9	11.1	0.01
Visit month (April-July)	28.7	31.6	0.06	29.8	30.3	0.01
Well child in past 12 mo.	20.4	11.4	0.25	17.7	17.3	0.01
Sick visit in past 12 mo.	36.8	27.8	0.19	34.3	35.0	0.02

attention at every timepoint. In contrast, among children with overweight, the proportion receiving attention remained less than one third. Yet, most children in the study received BMI attention at some point, including 100% with obesity class 3 and 85% with overweight.

Figure 1.1: Percent receiving BMI attention at each timepoint by BMI category, from timepoint 1–10.



BMI category	1	2	3	4	5	6	7	8	9	10	*% no BMI attention, any timepoint*
Obesity 3	70%	64%	55%	53%	59%	55%	50%	64%	54%	58%	0%
Obesity 2	53%	46%	43%	42%	40%	45%	38%	43%	40%	42%	5%
Obesity 1	36%	31%	32%	33%	33%	33%	34%	31%	31%	33%	12%
Overweight	27%	27%	28%	29%	30%	30%	32%	32%	30%	31%	15%

Causal effect of BMI attention on $\Delta\%\text{BMI}_{p95}$ With PS stratification, BMI attention significantly improved $\Delta\%\text{BMI}_{p95}$, relative to no attention, at timepoints 1 and 8 and yielded non-significant improvement in $\Delta\%\text{BMI}_{p95}$ for all but timepoint 3 (Figure 1.2). Without PS stratification, one might have concluded that BMI attention improved $\Delta\%\text{BMI}_{p95}$ at timepoint 1, worsened $\Delta\%\text{BMI}_{p95}$ at timepoints 3–4, and yielded non-significant $\Delta\%\text{BMI}_{p95}$ worsening at timepoints 5–10.

Sensitivity analysis results To illustrate the results of the sensitivity analysis using a clinical example, assume high parental BMI was an unmeasured confounder (denoted U) that affected both the likelihood of BMI attention and ΔBMI_{p95} of the child but was not included

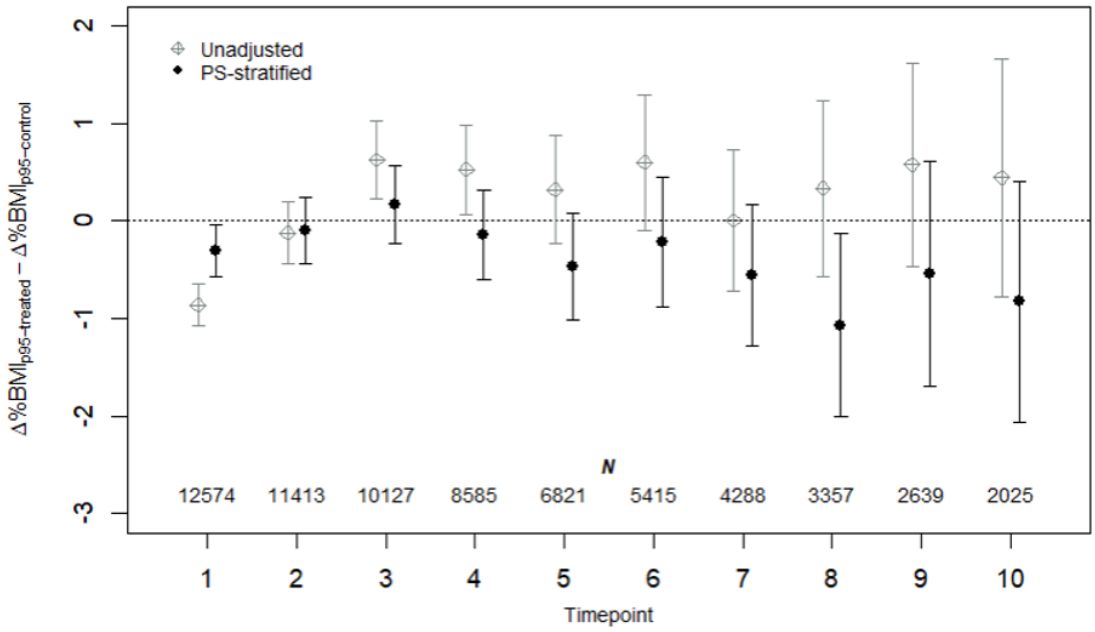
in the PS analysis (Table 1.3). Let $U = 1$ denote high parental BMI. Column 1 presents two scenarios in which U doubles or triples the odds of BMI communication. Columns 2–3 indicate the impact of U on $\Delta\text{BMI}_{\text{p95}}$ among controls vs. treated. The final column presents the impact of varying the prevalence of U. When $U = 1$ and the prevalence of U is low (10%), all point estimates fall within the 95% confidence intervals (CIs) of our study results. Once prevalence reaches $\geq 50\%$, sensitivity estimates remain within the 95% CI only when U affects the outcome equally and in the same direction for treated and controls. A similar pattern occurs at timepoints 2–5 with more results robust to U's presence compared to results at timepoint 1 (data not shown). By timepoints 6–10, all estimates fall within the 95% CI of study results. In sum, study results are largely robust to U at time 1–5, and completely robust to U by timepoints 6–10.

Table 1.3: Sensitivity analysis examining impact of identifying an unmeasured confounder at $t = 1$. Shading denotes point estimates outside 95% CI of study results at $t = 1$.

Effect of $U = 1$:			Point estimates resulting from varying prevalence of U		
On treatment	On outcome for:				
	controls	treated	10%	50%	90%
$U = 1$ doubles the odds of receiving BMI atten.	-0.3	-0.3	-0.29 (-0.55, -0.03)	-0.26 (-0.52, 0.00)	-0.29 (-0.55, -0.03)
		0.3	-0.38 (-0.64, -0.12)	-0.63(-0.87, -0.37)	-0.85 (-1.11, -0.60)
	0.3	-0.3	-0.24 (-0.50, 0.02)	0.01 (-0.25, 0.27)	0.24 (-0.02, 0.50)
		0.3	-0.33 (-0.59, -0.07)	-0.36 (-0.62, -0.10)	-0.32 (-0.58, -0.07)
$U = 1$ triples the odds of receiving BMI atten.	-0.3	-0.3	-0.28 (-0.53, -0.02)	-0.23 (-0.49, 0.03)	-0.28 (-0.54, -0.03)
		0.3	-0.38 (-0.64, -0.12)	-0.64 (-0.90, -0.38)	-0.86 (-1.12, -0.60)
	0.3	-0.3	-0.23 (-0.49, 0.02)	0.02 (-0.24, 0.28)	0.24 (-0.02, 0.50)
		0.3	-0.34 (-0.60, -0.08)	-0.39 (-0.65, -0.13)	-0.33 (-0.59, -0.07)

1.4. Discussion We observed that the effect of pediatricians addressing high BMI in 2–18-year-olds during primary-care visits is that $\Delta\%\text{BMI}_{\text{p95}}$ improves early (at timepoint 1)

Figure 1.2: Comparison of results from unadjusted and PS-stratification methods.



Time point	**1**	2	3	4	5	6	7	**8**	9	10
Estimate	**-0.31**	-0.10	0.17	-0.14	-0.47	-0.21	-0.56	**-1.07**	-0.54	-0.83
95th CI	**-0.57, -0.04**	-0.43, 0.24	-0.22, 0.56	-0.60, 0.31	-1.01, 0.07	-0.88, 0.46	-1.29, 0.17	**-2.01, -0.13**	-1.69, 0.61	-2.06, 0.41

and later (at timepoint 8) with non-significant improvements at most other timepoints. The presence of a significant effect of BMI attention on $\Delta\%\text{BMI}_{\text{p95}}$ at any timepoint suggests that pediatrician attention to BMI can improve overweight/obesity. The absence of a significant effect of BMI attention at most timepoints, even in large samples, highlights the need for increasingly potent tools to improve the weight of children with obesity. This study provides the first explicitly causal estimates of the effect of BMI attention on $\Delta\%\text{BMI}_{\text{p95}}$ across time using observational data from an EHR.

The non-uniform pattern of BMI effects differs from Turer *et al*'s 2019 finding that BMI attention was associated with a 20% increase in the likelihood of BMI improvement [5]. Whereas the 2019 study used regression to account for clinic and patient (but not time) effects, we applied longitudinal causal methods with PS stratification. This allowed us to minimize the imbalance of patient characteristics between arms and thereby reduce confounding bias.

PS methods have several notable advantages over regression-based covariate adjustment [6,19]: First, one can conduct a PS analysis without repeatedly "touching" the outcome value, thereby eliminating the risk of introducing bias through outcome model selection. Second, correlation among covariates, which causes instability of treatment effects in regression models, is irrelevant in a PS analysis, where the causal effect estimation depends on predicted PS values rather than model coefficients. Third, PS methods are less reliant on tenuous model assumptions about the dependence of outcomes on covariates. Finally, the fact that our longitudinal PS analysis demonstrates an enhanced effect of BMI attention on the reduction of $\Delta\%\text{BMI}_{\text{p95}}$ suggests that we have properly adjusted for bias by indication, which we anticipated would be substantial in this study.

Strengths Our use of PS methods created excellent balance in participant characteristics between arms, thereby substantially reducing bias in treatment effect estimates. Our study

benefitted from access to ample longitudinal data, with each timepoint including data from at least 2,000 children. Our study population was diverse, including 60% Hispanics and 20% African Americans (Table 1.1).

Limitations A limitation of PS stratification is its assumption of no unmeasured confounders. We addressed this concern by conducting a sensitivity analysis that demonstrated robustness of our findings to all but the most extreme confounding. A limitation of any analysis taking BMI as an outcome is the dependence of BMI on height, which is often missing or inaccurately recorded in EHR. We addressed this concern by imputing missing and implausible height data. Finally, the 84.7% sensitivity of the electronic phenotyping approach implies that $\sim 15\%$ of the treated are misclassified as controls. Because such misclassification would be expected to attenuate treatment effects, it is conceivable that the causal effect of attention to BMI is stronger than we have estimated it to be.

1.5. Conclusion Attention to BMI has a modest causal effect in the direction of reducing $\Delta\%\text{BMI}_{\text{p95}}$. The magnitude of this effect varies across time.

Chapter 2

Generating data from a marginal structural model

2.1. Introduction The marginal structural models (MSMs) comprise "a class of causal models for the estimation, from observational data, of the causal effect of a time-dependent exposure in the presence of time-dependent covariates that may be simultaneously confounders and intermediate variables" [20]. MSMs do not contain covariates because they are models for causal effects on the entire population [20]. MSM parameters are the marginal treatment effects. Thus, the response of an MSM is the expected outcome of the population modeled on the treatment status or dosage only. When using observational data to estimate an MSM, covariate-induced bias is eliminated by weighting the data by inverse-probability-of-treatment weighting (IPTW). The weighted data, or *pseudo-population*, has two properties: (1) the treatment and covariates are independent and (2) the expected outcome of one group (treated or control) in the pseudo-population equals the standardized outcome mean of the same group of the actual population [20, 21]. MSM estimates obtained using IPTW are consistent and valid for causal inference.

To empirically assess the validity of MSM estimates, one can conduct a simulation study in which one generates data under a specified MSM. A common, intuitive approach to data generation is in the order of covariate X, treatment status A (1 for treatment received, 0 for control), and the outcome Y as a function of X and A, modeled as a generalized linear model with linear predictor $\alpha + \gamma X + \beta A$. After weighting the data by IPTW, the treatment effect β is estimated by regressing Y on A only, with the linear predictor modeled as $\alpha^* + \beta^* A$.

A problem with this approach is that β (the conditional effect of A on Y given X) and β^* (the marginal effect of A on Y) are in general nonnegligibly different, regardless of whether

X and A are independent [22, 23]. This "bias" [22, 23] between β and β^* was proven by Gail *et al* [22] to occur when Y takes a form other than linear or exponential on X and A. From this point onward, we refer to the "bias" as the *noncollapsibility* of covariates. Noncollapsibility arises from the inclusion (or exclusion) of X; $\hat{\beta}$ is a biased estimate of β^* and $\hat{\beta}^*$ is a biased estimate of β. Because of noncollapsibility, when one generates data from X and A, the actual marginal treatment effect may differ substantially from the desired marginal treatment effect. The generated data will generally follow a saturated model rather than the desired MSM.

Others have proposed methods of data generation to create a desired MSM [24–31]. However, the existing methods are highly specific to the type of data to be generated (e.g, survival) [24–27, 31], or are computationally burdensome [28–30].

In this chapter, we introduce a new method of generating data with a known marginal treatment effect. Our method differs from the existing methods in that it is generalizable to any type of outcome and is easy to implement. The key feature of the new method is data generation directly from a specified MSM. Therefore, the order of data generation is potential outcomes $Y \rightarrow X$ instead of the more natural order of $X \rightarrow Y$. Using the new method, we conducted Monte Carlo simulations to assess the the performance of estimated MSMs by examining the (i) bias of the estimate, (ii) standard error (SE) and the coverage probability (CP), (iii) Type I error, defined as the probability of falsely finding a significant interaction term when interaction is not present, and (iv) power, defined as the probability of correctly finding a significant interaction term.

2.2. Overview of the marginal structural model

The key features of the MSM are captured in its name. The term *marginal* refers to the fact that MSMs estimate the marginal distribution of potential outcomes (as opposed to a joint distribution of the potential outcomes), and the term *structural* is borrowed from

the econometric and social sciences literature, in which potential variables are referred to as structural [20]. Table 2.1 illustrates the potential outcomes of N subjects in a longitudinal setting with two timepoints. Under this scenario, there are four possible sequences of treatment (11, 10, 01, 00), and thus, four potential outcomes for subject i $(Y_i^{11}, Y_i^{10}, Y_i^{01}, Y_i^{00})$, corresponding to the treatment sequences. The MSM models $\mathbb{E}[Y^{11}, Y^{10}, Y^{01}, Y^{00}]$.

Table 2.1: Under two timepoints, each subject has four potential outcomes.

i	Y_i^{11}	Y_i^{10}	Y_i^{01}	Y_i^{00}
1	Y_1^{11}	Y_1^{10}	Y_1^{01}	Y_1^{00}
2	Y_2^{11}	Y_2^{10}	Y_2^{01}	Y_2^{00}
3	Y_3^{11}	Y_3^{10}	Y_3^{01}	Y_3^{00}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	Y_N^{11}	Y_N^{10}	Y_N^{01}	Y_N^{00}

The implementation of an MSM consists of three steps: (1) model specification, (2) calculation of the propensity scores, and (3) estimation. We define A_t as a binary variable that equals 1 for treatment received and 0 for control, for $t = 1, \ldots, T$. The final outcome Y is observed at timepoint T, after observing the covariate history up to time T, $\bar{X}_T = (X_1, \ldots, X_T)$, and the treatment history up to time T, $\bar{A}_T = (A_1, \ldots, A_T)$.

Step 1: Model specification The model specification depends on the type of treatment effect that one wishes to estimate. For continuous Y, we might specify model (2.1) if we are interested in estimating time-specific treatments.

$$\mathbb{E}[Y] = \beta_0 + \sum_{t=1}^{T} \beta_t A_t, \tag{2.1}$$

where $\beta_1, \ldots, \beta_T$ are the treatment effects at timepoints 1 to T. Model (2.1) contains the main effects only, but one might fit a more complex model by including interaction terms. Alternatively, model (2.1) can be simplified further to assume a constant treatment effect (β_c) over time:

$$\mathbb{E}[Y] = \beta_0 + \beta_c \sum_{t=1}^{T} A_t. \tag{2.2}$$

Models (2.1) and (2.2) do not contain covariates because they model the marginal distribution of the outcome given treatment status. In the observed dataset, the bias induced from covariates is adjusted by weighting the subjects.

Step 2: Probability of observed treatment history Given a dataset, we can mimic the population from which the data were sampled (in other words, create a pseudo-population) by weighting subjects by the inverse of their probabilities of the observed treatment history, conditional on $\bar{X}_T$:

$$\Pr[\bar{A}_T = \bar{a}_T | \bar{X}_T = \bar{x}_T] = \prod_{t=1}^{T} \Pr[A_t = a_t | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{X}_t = \bar{x}_t]. \tag{2.3}$$

We estimate $\Pr[A_t = 1 | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{X}_t = \bar{x}_t]$ from a logistic regression model, setting $A_{t-1} = 0$ when $t = 1$. Then, we define the weight as

$$w = \frac{1}{\prod_{t=1}^{T} \widehat{\Pr}[A_t = a_t | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{X}_t = \bar{x}_t]}.$$

Because w can be highly unstable as T increases, it is generally recommended to use the stabilized weight (w^*) [20] defined as

$$w^* = w \cdot \Pr[\bar{A}_T = \bar{a}_T], \tag{2.4}$$

where $\Pr[\bar{A}_T = \bar{a}_T]$ is the marginal probability of the observed treatment history. The marginal probability is estimated by the observed proportion of subjects with the specific treatment history. Thus, the stabilized weight of subject i with treatment sequence $\bar{a}_{iT}$ is estimated by

$$\widehat{w}_i^* = w_i \cdot \frac{\sum_{j=1}^{N} I(\bar{a}_{jT} = \bar{a}_{iT})}{N}, \tag{2.5}$$

where I is an indicator function that equals 1, if $\bar{a}_{jT} = \bar{a}_{iT}$, and 0 otherwise.

Step 3: Estimation The final step is to estimate the MSM parameters (specified in Step 1) by fitting a weighted regression using w^* as weights. Estimates obtained from the weighted regression are the estimated causal treatment effects.

2.3. Noncollapsibility Known to occur when the outcome is modeled other than linear or exponential on the covariates and treatment effect, noncollapsibility is a "bias" between the conditional treatment effect and marginal treatment effect that arises due to the inclusion or exclusion of covariates [22,23]. In this section, we conduct simple simulations by generating normal, binary and survival data for $N = 1000, 10,000, 10,0000, 1,000,000$. First, we demonstrate the *collapsibility* of covariates using normal data, followed by the demonstration of noncollapsibility in binary and survival data.

Normal: collapsibility We generated continuous, normal outcome data as

$$X \sim \mathcal{N}(0, 1)$$

$$\Pr[A = 1] = 0.5$$

$$Y = X + \beta A + \mathcal{N}(0, 0.2),$$

where $\beta = 2$ is the group treatment effect for the treated. For each N, β was estimated by regressing Y on A only. No weighting was required because there was no confounding by X. Table 2.2 shows that $\hat{\beta}$, which can also be obtained by using the averages of individual outcomes in the treated and control, is approximately equal to β for all N.

Table 2.2: Illustration of collapsibility in normal data.

N	$\hat{\beta}$	95% Confidence interval	β
1000	1.98	1.86, 2.11	2
10,000	1.99	1.95, 2.03	
100,000	1.99	1.98, 2.01	
1,000,000	2.00	2.00, 2.01	

Binary We generated binary outcomes as

$$X \sim \mathcal{N}(0, 1)$$

$$\Pr[A = 1] = 0.5$$

$$\mathbb{E}[Y|X, A] = \text{expit}\left(X + \beta A\right),$$

where $\exp(\beta) = 2$ is the odds ratio of $Y = 1$ when $A = 1$. Ignoring noncollapsibility, a naive estimate of β ($\hat{\beta}_{\text{naive}}$) was obtained by using a logistic regression to model Y on A only. No weighting was required because there was no confounding by X on A. Table 2.3 shows $\exp(\hat{\beta}_{\text{naive}})$ across N. Noncollapsibility is evident in that $\exp(\hat{\beta}_{\text{naive}})$ does not approximate $\exp(\beta)$ whether N is small or as large as one million, and the confidence interval deviates further from $\exp(\beta)$ as the standard error decreases.

Table 2.3: Illustration of noncollapsibility in binary data.

N	$\exp(\hat{\beta}_{\text{naive}})$	95% Confidence interval	$\exp(\beta)$
1000	1.69	(1.31, 2.18)	2
10,000	1.71	(1.58, 1.85)	
100,000	1.75	(1.70, 1.79)	
1,000,000	1.78	(1.76, 1.79)	

Survival We generated survival data from a Cox-Weibull distribution as

$$X \sim \mathcal{N}(0,1)$$

$$\Pr[A=1] = 0.5$$

$$\mathbb{E}[Y|X,A] = \left[\frac{-\log(U)}{\lambda \exp(X+\beta A)}\right]^{1/\eta},$$

derived by Bender et al [32], where $U \sim \mathcal{U}(0,1)$. Following the examples in the studies of Austin *et al*, we set $\lambda = 0.00002$ and $\eta = 2$ [30, 33–36]. We set $\exp(\beta) = 2$ as the hazard ratio for the treatment status A. All event times Y were assumed to be observed.

Ignoring noncollapsibility, a naive estimate of β ($\hat{\beta}_{\text{naive}}$) was obtained by using a Cox regression to model Y on A only (Table 2.4). No weighting was required because there was no confounding by X on A. As in the binary data, the bias of the estimated measure $\exp(\hat{\beta}_{\text{naive}})$ did not diminish with increasing N.

2.4. Method We describe the new method of generating data from a specified MSM for a longitudinal setting with T timepoints and binary treatment A. Here, we define Y_i as the set of all potential outcomes of subject i that correspond to all possible treatment sequences under T timepoints.

Table 2.4: Illustration of noncollapsibility in survival data.

N	$\exp(\hat{\beta}_{\text{naive}})$	95% Confidence interval	$\exp(\beta)$
1000	1.52	1.39, 1.64	2
10,000	1.49	1.45, 1.53	
100,000	1.55	1.54, 1.56	
1,000,000	1.55	1.54, 1.55	

Step 1: Generate potential outcomes According to the designated MSM, generate the vector Y_i (1×2^T) of potential outcomes under all possible treatment sequences for each subject i in $i = 1, \ldots, N$.

Step 2: Generate baseline covariates For subject i, generate the baseline covariates, X_{i1}, as a function of the subject's potential outcomes Y_i. For continuous X, $X_{i1} = (x_{i11}, \ldots, x_{i1p})$ can be defined as

$$X_{i1} = H \cdot Y_i^{\intercal} + \epsilon, \tag{2.6}$$

where H is a $p \times 2^T$ matrix and ϵ is an error term. By varying the values H, one can control the degree of confoundedness of X_1 in Y. Setting all elements of H to 0 implies that Y and X_1 are independent.

Step 3: Define propensity score models For $t = 1, \ldots, T$, define the propensity score model at timepoint t, π_t, conditional on the covariate history up to time t and the treatment history up to time $t-1$ (if $t = 1$, $\bar{A}_{t-1} = 0$):

$$\pi_t = \Pr[A_t = 1 | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{X}_t = \bar{x}_t].$$

Use a logistic regression to estimate π_1. Then, generate generate A_1 using $\hat{\pi}_1$ as the probability of $A_1 = 1$. In R, A_1 is generated by using rbinom() with $\hat{\pi}_1$ as the probability.

Carry out Steps 4, 5 for $j = 2, ..., T$.

Step 4: Generate time-varying covariates Generate X_j as a function of $\bar{X}_{j-1}$ and $\bar{A}_{j-1}$.

$$X_j = \sum_{k=1}^{j-1} v_k^\intercal X_k + \sum_{k=1}^{j-1} u_k A_k + \mathcal{N}(0, \sigma), \tag{2.7}$$

where v_k is a vector that controls the degree of influence of X_k on X_j, u is a scalar that controls the degree of influence of A_k on X_j, and σ is the standard deviation of an error term. Setting $v_k = 0$ implies independence between X_k and the current covariate X_j; setting $u_k = 0$ implies independence between A_k and X_j.

Step 5: Generate treatment Generate A_j using $\hat{\pi}_j$ as the probability of $A_j = 1$. In R, A_j is generated by using rbinom() with $\hat{\pi}_j$ as the probability.

Step 6: Observe one outcome per subject After carrying out Steps 4, 5 for $j = 2, \ldots, T$, each subject now has a realized treatment history $\bar{a}_T = (a_1, \ldots, a_T)$. We then observe the outcome (of the potential outcomes) that corresponds to the realized treatment sequence. In an example setting with $T = 2$ timepoints, if a subject's realized treatment sequence is $\bar{a}_T = (1, 1)$, we observe the potential outcome Y^{11}, but not the other potential outcomes Y^{10}, Y^{01}, Y^{00}. Table 2.5 demonstrates an example of potential outcomes vs. observed outcome for data with binary outcomes.

Step 7: MSM estimation Use the X, the observed treatment sequence $\bar{a}_T$, and the observed outcome $Y^{\bar{a}_T}$ to estimate an MSM.

2.4.1. Binary data To apply the method to binary data, only step 1 needs to be modified.

Table 2.5: An example population of N subjects with their four potential outcomes and one observed outcome (bold).

i	(a_1, a_2)	Y^{11}	Y^{10}	Y^{01}	Y^{00}
1	(1, 0)	1	**1**	1	1
2	(0, 1)	1	0	**0**	1
3	(0, 0)	1	1	1	**1**
⋮	⋮	⋮	⋮	⋮	⋮
N	(0, 1)	1	0	**0**	0

Step 1 For binary data, the MSM can be specified as, but not limited to, equation (2.1) or equation (2.2). Use $\mathbb{E}[Y_T|\bar{A}_T = \bar{a}_T]$ as probabilities to generate binary potential outcomes for $\bar{A}_T = \bar{a}_T$.

2.4.2. Survival data The method can be applied to survival data in the context of Cox proportional hazards model by modifying steps 1, 7.

Step 1 Generate potential survival times S. For demonstration, we again use Bender *et al*'s formula from Section 2.3 for generating survival times from a Cox-Weibull distribution [32]:

$$S = \left[\frac{-\log(U)}{\lambda \exp\left(g(\bar{A}_T;\beta)\right)}\right]^{1/\eta},$$

where $U \sim \mathcal{U}(0,1)$. Following the studies of Austin *et al*, one can set $\eta = 2$ and $\lambda = 0.00002$ [30, 33–35].

The exponentiated function $g(\bar{A}_T;\beta)$ is the linear form of an MSM that defines the treatment effect on the outcome. As with any MSM, $g(\bar{A}_T;\beta)$ can be simple, saturated, or have a single common treatment effect across T timepoints.

After generating S, one can apply censoring by setting a threshold so that survival times exceeding the threshold are censored. The threshold can be uniform or depend on $\bar{A}_T$. Alternatively, one can generate a binary censored status (1 if censored) randomly or by using probabilities defined by $\bar{A}_T$.

Step 7 Using $g(\bar{A}_T; \beta)$ from Step 1, fit a Cox proportional hazards model using the weights at timepoint T:

$$h(T) = h_0(T)\exp\left(g(\bar{A}_T; \beta)\right),$$

where $h(T)$ and $h_0(T)$ are the hazard and baseline hazard at timepoint T.

2.5. Simulation Data were generated as described below for $T = 2$.

Specifying MSMs Motivated by obesity studies [5, unpublished data, 2022], we generated data in which the intervention was clinician attention to high body mass index (BMI) and the outcome was the binary status of weight improvement. Potential outcomes Y were generated under two models:

$$\text{Model A: } \Pr[Y = 1] = \text{expit}\left(\beta_0 + \beta_1 A_1 + \beta_2 A_2\right)$$

$$\text{Model B: } \Pr[Y = 1] = \text{expit}\left(\beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_3 A_1 A_2\right)$$

The selection between the two models depended on the outcome measure to be assessed (Table 2.7). The chosen values of β of the MSM reflect the study's clinical findings:

- $\beta_0 = \log(0.2)$: Without any attention to BMI, the odds of weight improvement is low (OR= 0.2).
- $\beta_1 = \log(3)$: Attention to BMI is effective at timepoint 1 (OR= 3).

- $\beta_2 = \log(2)$: The effect is diminished at timepoint 2 (OR= 2).
- $\beta_3 = -0.5\log(3)$ (Model B only): More intervention is not much better than one intervention at timepoint 1; treatment effects are not additive.

Baseline covariates For our simulated data, we induced confounding by generating the scalar baseline covariate X_{i1} with varying values of H (equation (2.6), $p = 1$), for $i = 1, \ldots, N$. We induced four levels of confounding:

1. $H = (0, 0, 0, 0)$: No confounding — X_1 and Y are independent.
2. $H = (-2, 0, 0, 0)$: The observed proportion of subjects with $Y^{11} = 1$ underrepresents the true proportion. Those with $Y^{11} = 1$ are assigned lower values of X_1, X_2, lowering their p_1, p_2. As a result, those with $Y^{11} = 1$ are less likely to receive $A = (1, 1)$. Instead, those with $Y^{11} = 0$ are more likely to receive $A = (1, 1)$.
3. $H = (0, 0, 0, -2)$: The observed proportion of subjects with $Y^{00} = 1$ overrepresents the true proportion. Those with $Y^{00} = 1$ are less likely to receive $A = (1, 1)$ (in other words, more likely to receive $A = (0, 0)$). The magnitude of confounding levels 2 and 3 are considered equal.
4. $H = (-2, 0, 0, -2)$: Both 2 and 3. This level of confounding is the highest.

For all confounding levels, H is a 1×2^2 matrix and $\epsilon = \mathcal{N}(0, 0.5)$ (equation 2.6).

Table 2.6 illustrates the observed proportions of subjects with $Y^{a_1 a_2} = 1$ for the four confounding levels using a sample of $N = 10,000$ subjects. The numbers in the first row for the population are the $\Pr[Y^{a_1 a_2} = 1]$ determined by $\beta_0, \beta_1, \beta_2, \beta_3$.

Table 2.6: The population's probabilities of $Y^{a_1a_2} = 1$ and the observed means of $Y^{a_1a_2}$ in a sample of $N = 10,000$.

Source	Conf	Y^{11}	Y^{10}	Y^{01}	Y^{00}
Population	NA	0.42	0.39	0.30	0.16
Sample	1	0.42	0.37	0.29	0.17
	2	**0.05**	0.40	0.29	0.17
	3	0.41	0.37	0.29	**0.43**
	4	**0.05**	0.38	0.29	**0.28**

Propensity score models The propensity score models at timepoints 1 and 2 were defined as below:

$$\pi_1 = \text{expit}(X_1)$$

$$\pi_2 = \text{expit}(0.4X_1 + 0.6X_2 + 0.5A_1).$$

A_1, A_2 were generated by using rbinom() in R, setting π_1, π_2 as the probabilities.

The correct propensity score models were used to compute the inverse of probability of treatment.

Time-varying covariates X_2 were generated by setting $v_1 = 1$, $u_1 = -2$, $\sigma = 0.5$ (Equation (2.7)):

$$X_2 = X_1 + (-2A_1 + b) + \mathcal{N}(0, 0.5),$$

where $b = 1$. The additive term b was included so that when $A_1 = 1$, the effect of A_1 on X_2 is -1, and when $A_1 = 0$, the effect is 1.

Fitting MSMs We also used Models A and B as the MSM to be estimated. The selection between the two models depended on the outcome measure to be assessed. Table 2.7 outlines the combinations of the specified and fitted MSMs and the corresponding result measures.

Table 2.7: Result measures assessed under Scenarios 1, 2, and 3.

Scenario	True model	Fitted model	Results
1	A	A	Bias, SE, CP
2	B	B	Bias, SE, CP, Power
3	A	B	Type I error

Sample sizes Within each scenario, three sample sizes ($N = 1000, 5000, 10,000$) were used to generate data. Within each N, four confounding levels were used. Thus, for each scenario, there were 12 unique combinations of N and confounding level. At each combination, we generated 1000 samples using R version 4.2.0.

The correct propensity score models were used to compute the inverse of probability of treatment.

2.6. Results

Bias We assessed the bias of estimates ($\frac{1}{1000}\sum_{m=1}^{1000}(\hat{\beta}_j - \beta_j)$) under Scenarios 1 ($j = 1, 2$) and 2 ($j = 1, 2, 3$). Figure 2.1 shows that bias decreased as N increased, but it increased as the level of confounding worsened. The effect of confounding on bias was greater as the N decreased. The sensitivity of bias with respect to confounding levels suggests that IPTW is not completely robust to the degrees of confounding. This was consistent with the study results of Austin and Stuart [30] which showed, using survival data, that strong confounding

in the treatment mechanism resulted in biased estimates of the hazard ratio, even when the propensity score model was correctly specified (as was the case in our study). In Scenario 2, the estimate of β_3 (compared to β_1, β_2) incurred the most bias across N and confounding.

Standard error and coverage probability The standard error (SE) and coverage probability (CP) of MSM coefficients were evaluated under Scenarios 1 and 2. To obtain standard errors, we used two methods of variance estimation: the conventional method based on maximum likelihood estimation, and the bootstrap method using 250 samples per iteration. Figures 2.2 and 2.3 show the SE of each coefficient under both scenarios, stratified by N.

Overall, the bootstrap SEs were larger than the conventional SEs. In Scenario 1 (Figure 2.2), the bootstrap SEs were larger at confounding 2, 4. In Scenario 2 (Figure 2.3), at confounding 2 and 4, the bootstrap SEs of $\hat{\beta}_3$ were notably larger than the SEs of $\hat{\beta}_1$, $\hat{\beta}_2$, likely because, at levels 2 and 4, confounding was induced to distort the observed effect of interaction.

The bootstrap CP performed better than the conventional CP. Figures 2.4 and 2.5 show that, across N, the CPs using the bootstrap method were around 0.95, while the conventional method tended to undercover, consistent with the findings of Austin (2016) using survival outcome data [29]. Compared to bootstrap, the conventional CPs were sensitive to the confounding levels.

The superior performance of the bootstrap method is especially highlighted in the CPs of $\hat{\beta}_3$ (Figure 2.5). While the bootstrap CPs hovered around 0.95, the conventional CPs of $\hat{\beta}_3$ performed poorly (as low as 0.46 at confounding 4).

The undercoverage of the conventional CPs suggests that the model-based SEs underestimate the true variability.

Figure 2.1: Bias under Scenarios 1 and 2, stratified by N and confounding levels.

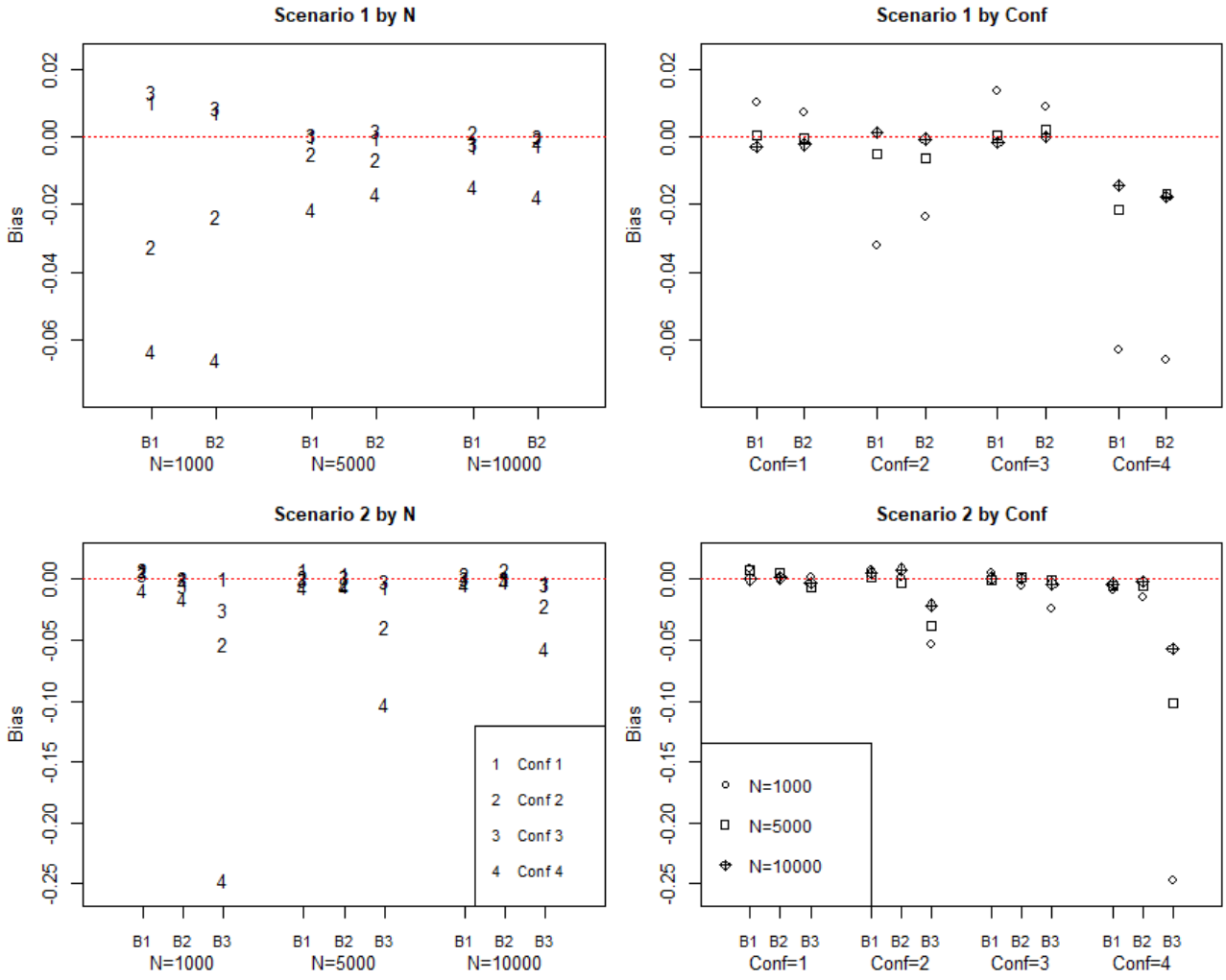


Figure 2.2: Standard errors in Scenarios 1, stratified by N and confounding levels.

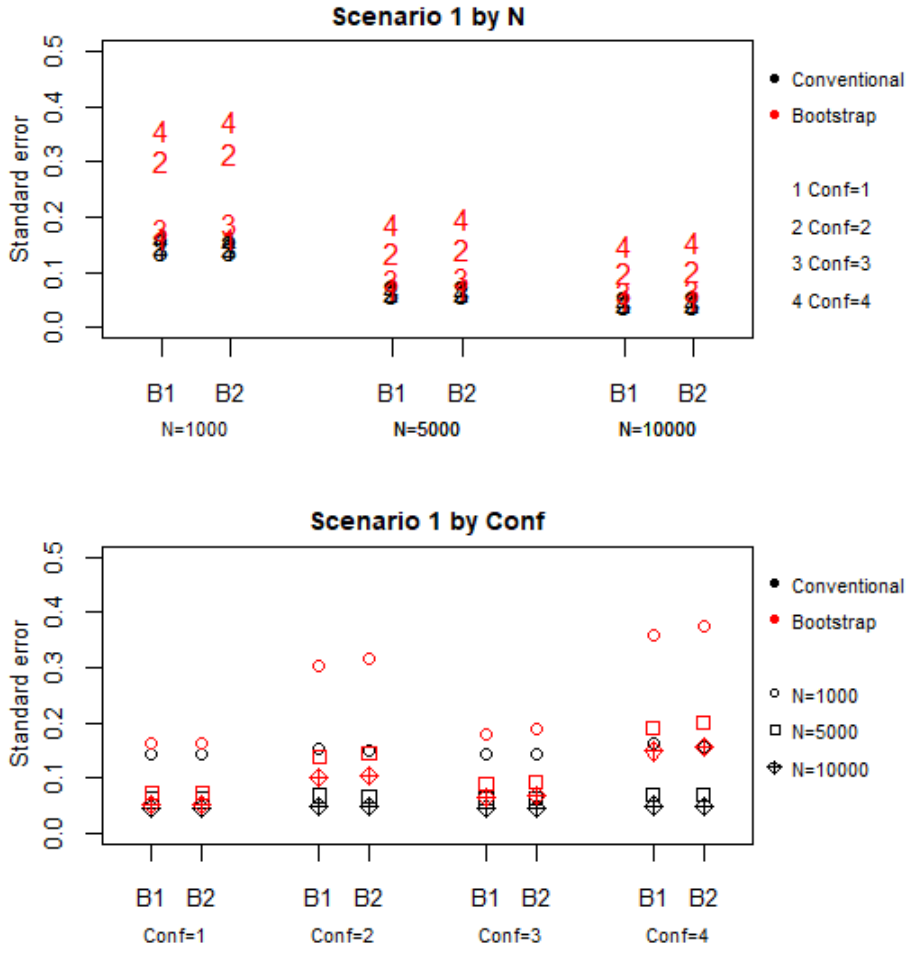


Figure 2.3: Standard errors in Scenarios 2, stratified by N and confounding levels.

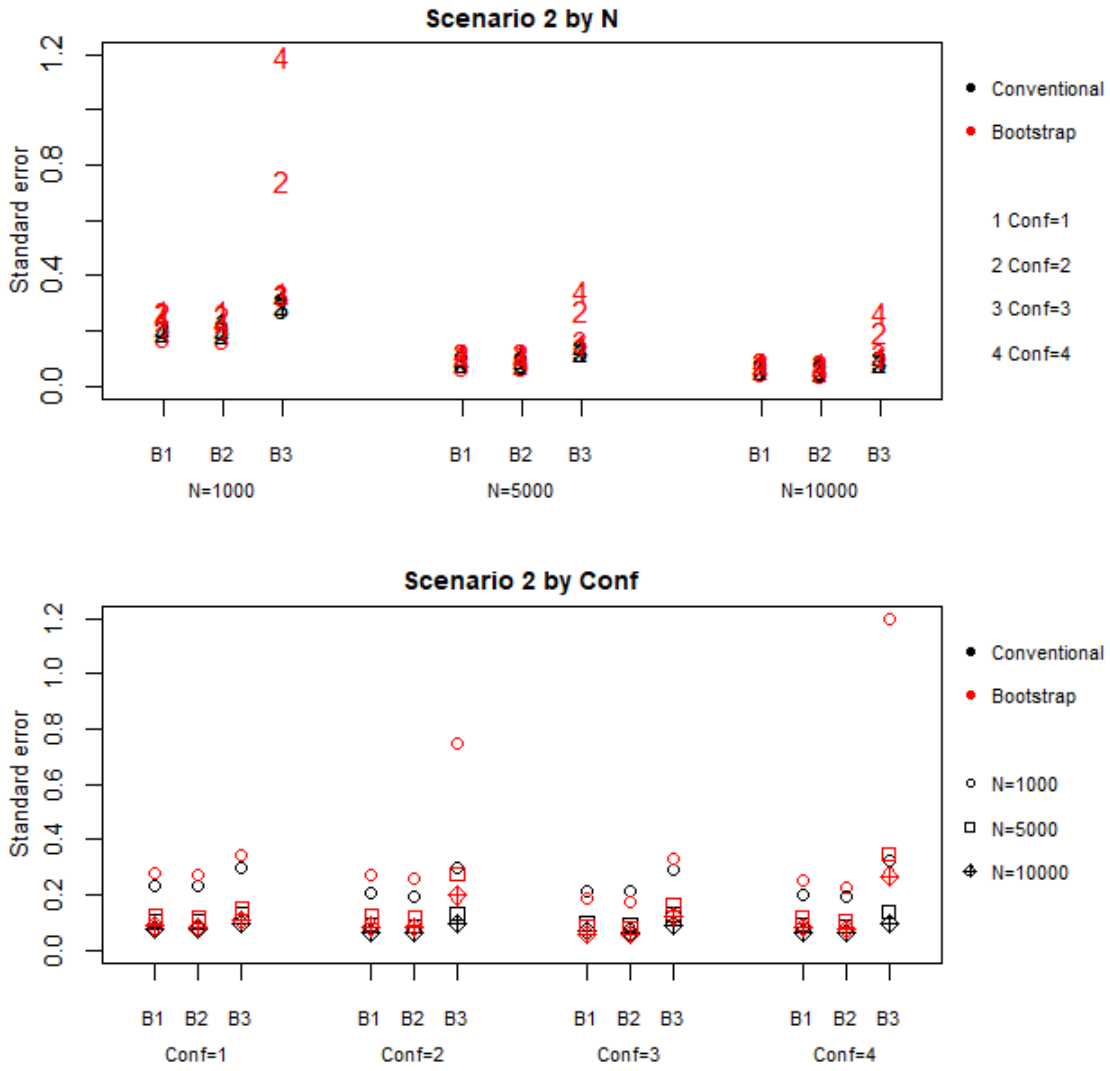


Figure 2.4: Coverage probability in Scenarios 1, stratified by N and confounding levels.

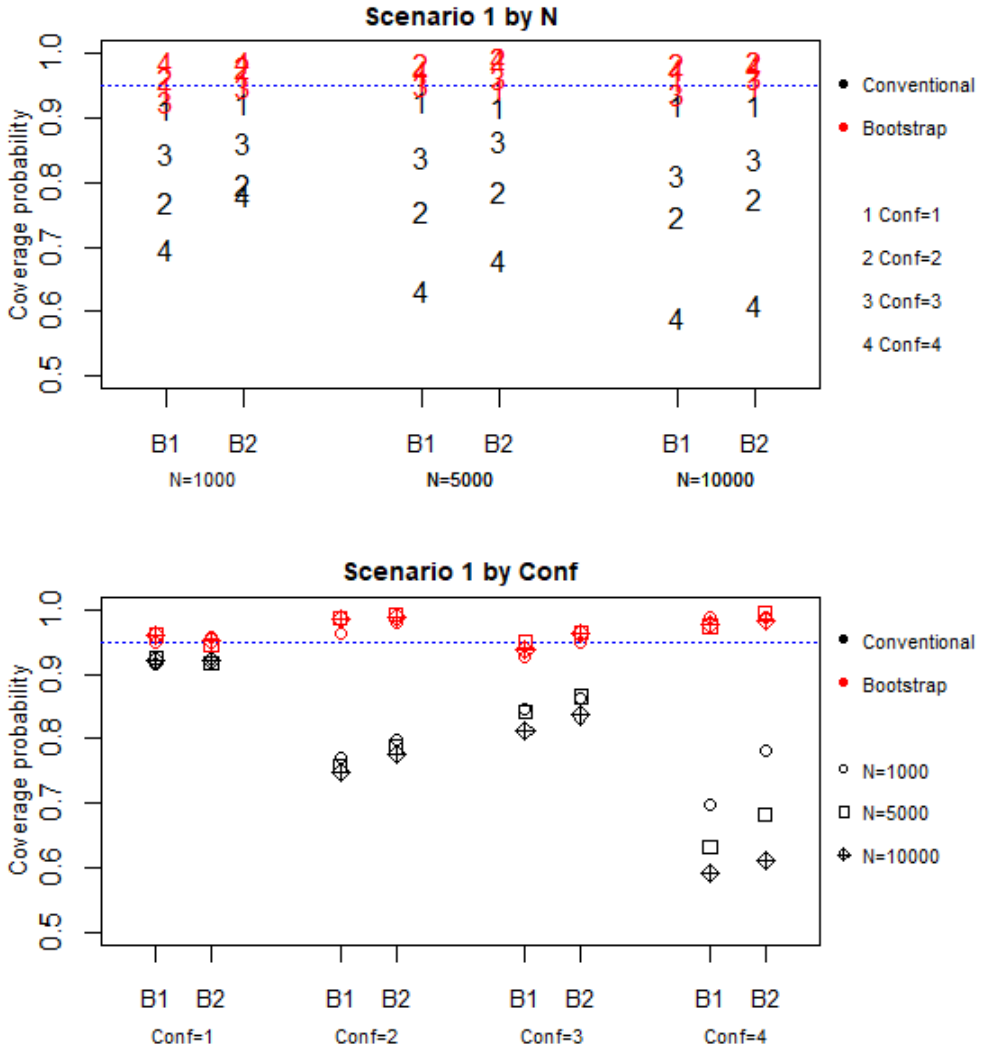
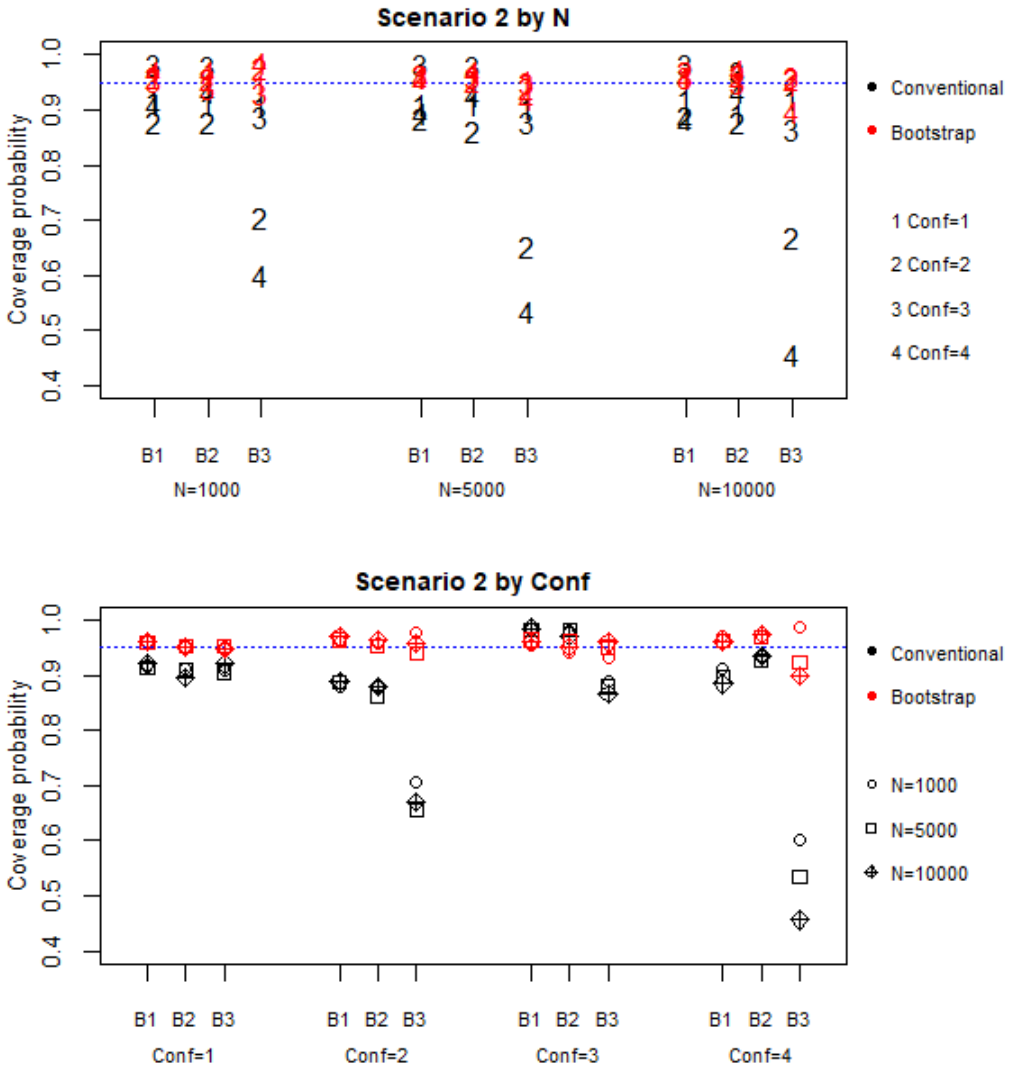


Figure 2.5: Coverage probability in Scenarios 2, stratified by N and confounding levels.



Type I error We defined Type I error as the probability of falsely finding a significant interaction term, at level 0.05, in the fitted MSM. Thus, we assessed Type I error under Scenario 3, in which the true model used to generate data did not have an interaction term, but the fitted model falsely had an interaction term. When determining the statistical significance of the interaction term (level 0.05), we used the bootstrap method of variance estimation, as it was shown to be superior to the conventional method. Figure 2.6 shows the Type I error across N and the confounding levels. There was no discernible pattern with respect to N or the confounding levels. Overall, the Type I errors ranged from 0.024 to 0.064.

Power We defined power as the probability of correctly obtaining a significant interaction, at level 0.05, from the fitted MSM. Thus, we assessed power under Scenario 2, in which both the true and fitted models contained an interaction term. At each N, the reference power was obtained by conducting 1000 Monte Carlo simulations without confounding, using equal weights for all subjects (0.480, 0.986, 1.000 in the order of increasing N). Figure 2.7 shows that, across N, the power of all four confounding levels resembles the pattern of the reference power across N, with confounding levels 2,4 lagging behind the other levels.

A plausible reason for the lower power at confounding 2, 4 is that, at those confounding levels, confounding was induced to distort the observed effect of the interaction term β_3 (See *Basleine covariates* in Section 2.5, Table 3.2). If confounding is not completely adjusted by IPTW, the estimated interaction effect $\hat{\beta}_3$ may not be substantial enough to be statistically significant.

2.7. Discussion We introduced a simple method of generating data directly from a specified MSM. Using the new method, we conducted Monte Carlo simulations and assessed some properties of the MSM using binary data and the correct propensity score models. The

Figure 2.6: Type I error in Scenario 3, stratified by N and confounding levels.

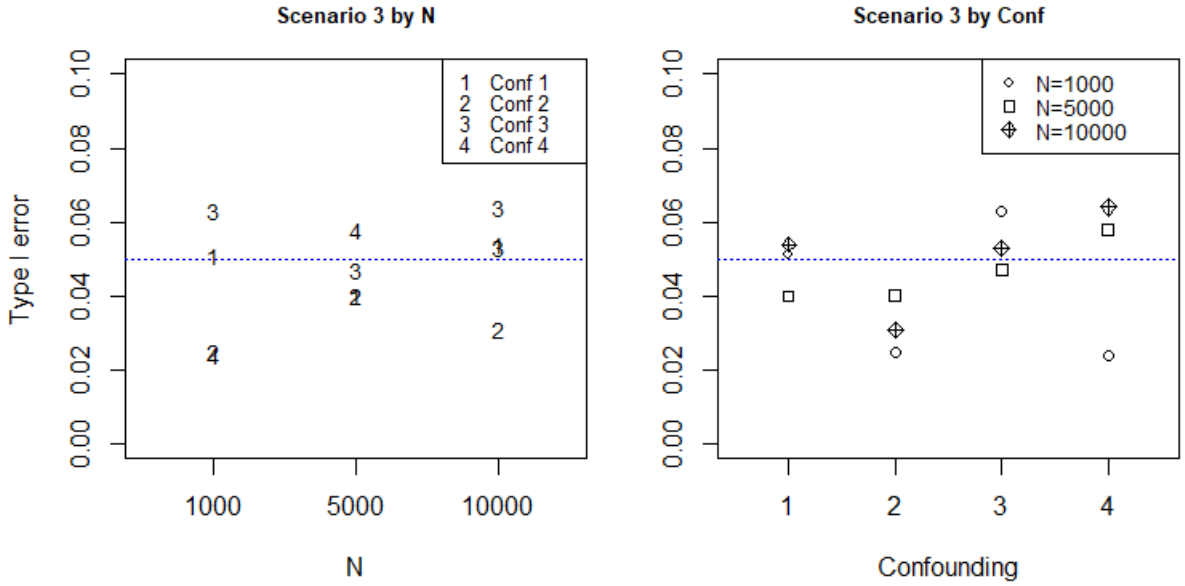
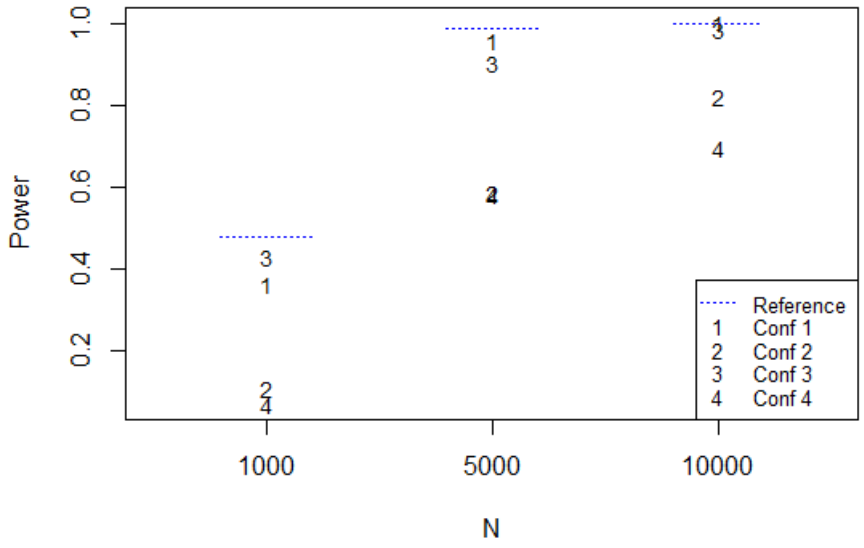


Figure 2.7: Power in Scenario 2, stratified by N.



results highlighted three findings. First, using binary data, the increasing magnitude of bias of the MSM estimates with respect to the level of confounding showed that IPTW is not robust to the degree of confounding, consistent with the findings of Austin and Stuart using survival data [30]. Second, also consistent with Austin's study [29], the bootstrap method of variance estimation outperformed the conventional (maximum-likelihood based) method. The bootstrap standard errors were larger, but the resulting coverage probability approximated the nominal value at all confounding levels. This signifies that one can generally make a valid inference using the bootstrap estimation of variance. Lastly, the estimated interaction terms were less stable than the estimates of the main effects when confounding was induced to affect the observed outcomes of those who received treatment at both timepoints. The relative instability was reflected in lower power. For investigators, the low power indicates that the decision to include interaction terms in the MSM should not be solely based on the significance of the estimated interaction parameter.

Our method differs from other methods in that the potential outcomes are generated directly from the specified MSM, without covariates, thus avoiding the noncollapsibility of covariates. Our method is not the only way to avoid noncollapsibility for all types of data. Alternatively, one can specify the marginal treatment effect, and use an iterative process to determine values of the conditional treatment effects that would induce the specified marginal effect [29, 30]. The advantage of our method is that it omits the iterative process and generates potential outcomes directly from the specified MSM.

We limited the simulation study to two timepoints ($T = 2$). As T increases, the probability of treatment becomes smaller, resulting in large weights even if stabilized weighting is used. When a handful of subjects with large weights drives the study results, weight truncation is a widely explored alternative. However, if the propensity score model is correct, large weights obtained using the propensity score should be valid. The next step of this study is to evaluate the validity of the MSM procedure when the weights are large.

Chapter 3

Assessing the impact of large weights on MSM estimates under multiple timepoints

3.1. Introduction In the preceding chapter we introduced a new method for generating data directly from a designated marginal structural model (MSM). Using this method, we demonstrated the behavior of MSM estimates in a simulation study assuming two timepoints. In this chapter, we further examine the behavior of MSM estimates, addressing some issues that arise in practice when applying more complex MSMs.

Generally, as the number of timepoints T increases, the estimated probabilities of treatment can become very small, and correspondingly the subject weights (computed as the inverses of the probabilities of treatment) can become large [20]. One does expect to see some discrepancy in weights, reflecting the value of weighting as a means to eliminate bias caused by confounding. But when discrepancies are extreme, investigators may be concerned about sensitivity of estimates to the outcome values of small numbers of highly influential observations, as well as sensitivity to the specification of the propensity score model [37].

Weight-trimming methods can reduce the impact of a few extremely large weights [38–42]. However, it is known that weight trimming may not always improve estimate accuracy [43], and there remains a lack of guidance on the optimal level of weight trimming [37, 43]. Moreover, the *ad hoc* approach of weight trimming is effectively an admission of failure of the procedure used to estimate the propensity scores [43]; thus, assuming that the propensity score model is correct – or nearly so – one would prefer to avoid trimming.

In this chapter, we deployed our new simulation method to assess the validity of MSM estimates obtained without weight trimming under scenarios with multiple treatment time-

points. We used two types of weights to obtain MSM estimates: Inverse-probability-of-treatment weights (IPTW) and stabilized weights (SW), the latter being generally recommended in the literature [20, 38]. Our study aims to compare the two weighting schemes using MSMs with binary outcomes. We moreover assess the sensitivity of estimates to misspecification of the propensity score model at a range of sample sizes (N = 200, 500, 1000, and 10,000) and numbers of treatment timepoints (T = 2, 5, 8, and 10).

3.2. Method The method of generating data from an MSM without noncollapsibility (Chapter 2, Section 2.4) is briefly outlined.

Step 1: Generate potential outcomes According to the designated MSM, generate the vector Y_i (1×2^T) of potential outcomes under all possible treatment sequences for each subject i in $i = 1, \ldots, N$.

Step 2: Generate baseline covariates For each subject, generate the baseline covariates, X_1, as a function of the subject's potential outcomes. For continuous X, X_{i1} can be defined as

$$X_{i1} = H \cdot Y_i^{\intercal} + \epsilon,$$

where H is a $p \times 2^T$ matrix and ϵ is an error term.

Step 3: Define propensity score models For $t = 1, \ldots, T$, define the propensity score model at timepoint t, π_t, conditional the covariate history up to time t and the treatment history up to time $t-1$ (if $t = 1$, $\bar{A}_{t-1} = 0$.).

$$\pi_t = \Pr[A_t = 1 | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{X}_t = \bar{x}_t].$$

Generate A_1 using π_1 as the probability of $A_1 = 1$.

Carry out Steps 4, 5 for $j = 2, \ldots, T$.

Step 4: Generate time-varying covariates Generate X_j as a function of $\bar{X}_{j-1}$ and $\bar{A}_{j-1}$.

$$X_j = \sum_{k=1}^{j-1} v_k X_k + \sum_{k=1}^{j-1} u_k A_k + \mathcal{N}(0, \sigma),$$

where v_k is a vector that controls the degree of influence of X_k on X_j, u is a vector that controls the degree of influence of A_k on X_j, and σ is the standard deviation of an error term.

Step 5: Generate treatment Generate binary A_j using π_j as the probability of $A_j = 1$.

Step 6: Observe one outcome per subject For each subject, observe the potential outcome that corresponds to the treatment history realized from generating $A_1, \ldots, A_T$ having used $\pi_1, \ldots, \pi_T$ as the probabilities.

Step 7: MSM estimation

3.3. Simulation For $T = 2, 5, 8, 10$, potential outcomes Y were generated from an MSM with a common treatment effect:

$$\Pr[Y = 1] = \text{expit}\left(\beta_0 + \beta_c \sum_{t=1}^{T} A_t\right), \tag{3.1}$$

setting $\beta_0 = \log(0.7)$, $\beta_c = \log(1.2)$.

Baseline scalar covariate X_{i1} was generated by defining the 1×2^T matrix as $H = (-1, 0, \ldots, 0, -1)$, setting the first and last elements to -1 and the rest to 0:

$$X_{i1} = H \cdot Y_i^{\intercal} + \epsilon,$$

where $\epsilon = \mathcal{N}(0, 0.5)$.

The subsequent time-varying covariates were generated by setting $v_{j-1} = 1$, $u_{j-1} = -2$, $\sigma = 0.5$ (Equation (2.7)):

$$X_j = X_{j-1} + (-2A_{j-1} + b) + \mathcal{N}(0, 0.5),$$

for $j = 2, \ldots, T$. The additive term $b = 1$ was included so that when $A_{j-1} = 1$, the effect of A_{j-1} on X_j is -1, and when $A_{j-1} = 0$, the effect is 1.

We considered two scenarios of propensity score models (Table 3.1). The first scenario only includes first-degree polynomials and the second scenario contains a third-degree polynomial.

At each timepoint, treatment status was generated by using rbinom() in R, setting the propensity score as the probability.

For both scenarios, propensity scores were estimated by using the logistic regression to model A as:

$$A_1 \sim X_1$$

$$A_j \sim X_{j-1} + X_j + A_{j-1},$$

for $j = 2, \ldots, T$. Thus, the propensity score model was correctly specified under Scenario 1 and misspecified under Scenario 2.

Under each scenario, data were weighted by IPTW and SW. Using the weighted data, we estimated model (3.1) to obtaini $\hat{\beta}_c$.

We conducted 1000 iterations using sample sizes $N = 200, 1000, 5000, 10,000$ and timepoints $T = 2, 5, 8, 10$. Following the results from Chapter 2, we used the bootstrap method of variance estimation for $\hat{\beta}_c$.

Table 3.1: Two propensity score models were considered.

Scenario	True propensity score models
1	$\pi_1 = \text{expit}(X_1)$ $\pi_j = \text{expit}(0.6X_j + 0.4X_{j-1} + 0.5A_{j-1}), j = 2, \ldots, T.$
2	$\pi_1 = \text{expit}(X_1)$ $\pi_j = \text{expit}(0.6X_j + 0.2X_{j-1} + 0.2X_{j-1}^3 + 0.5A_{j-1}), j = 2, \ldots, T.$

3.4. Results Results are shown for four combinations of weighting and type of propensity score models: (1) IPTW with the correct propensity score models, (2) SW with the correct propensity score models, (3) IPTW with misspecified propensity score models (denoted IPTW-mis), and (4) SW with misspecified propensity score models (denoted SW-mis).

Distribution The Q-Q plots (Figure 3.1) were plotted by normalizing $\hat{\beta}_{cm}$ of iteration m as $\frac{\hat{\beta}_{cm}-\beta_c}{s}$, where $s = \frac{1}{999}\sum_{m=1}^{1000}(\hat{\beta}_{cm} - \bar{\hat{\beta}}_c)^2$, $\bar{\hat{\beta}}_c = \frac{1}{1000}\sum_{m=1}^{1000}\hat{\beta}_{cm}$. The distribution of $\hat{\beta}_c$ obtained using IPTW is well-approximated by the normal distribution. The distribution becomes more thick-tailed as T and N increase. For instance, the tails are thicker at $T = 10, N = 10,000$ than at $T = 8, N = 10,000$. The same pattern applies to $\hat{\beta}_c$ obtained using SW, IPTW-mis, and SW-mis (Appendix A).

Estimate error The error of $\hat{\beta}_{cm}$ from iteration m, defined as $\hat{\beta}_{cm} - \beta_c$, was examined using boxplots (Figures 3.2, 3.3). Outliers (shown as empty circles) were defined as being outside 1.5 interquartile range from the first or third quartile. The median estimate errors approximated 0 across T, except at $T = 2, N = 200$. Overall, the error distributions of SW varied less with fewer outliers than distributions of IPTW. Furthermore, the distributions of IPTW-mis varied less with fewer outliers than the distributions of IPTW, and SW-mis less than SW.

Figure 3.1: Distribution of $\hat{\beta}_c$ obtained using IPTW. A similar pattern was observed for SW, IPTW-mis, and SW-mis.

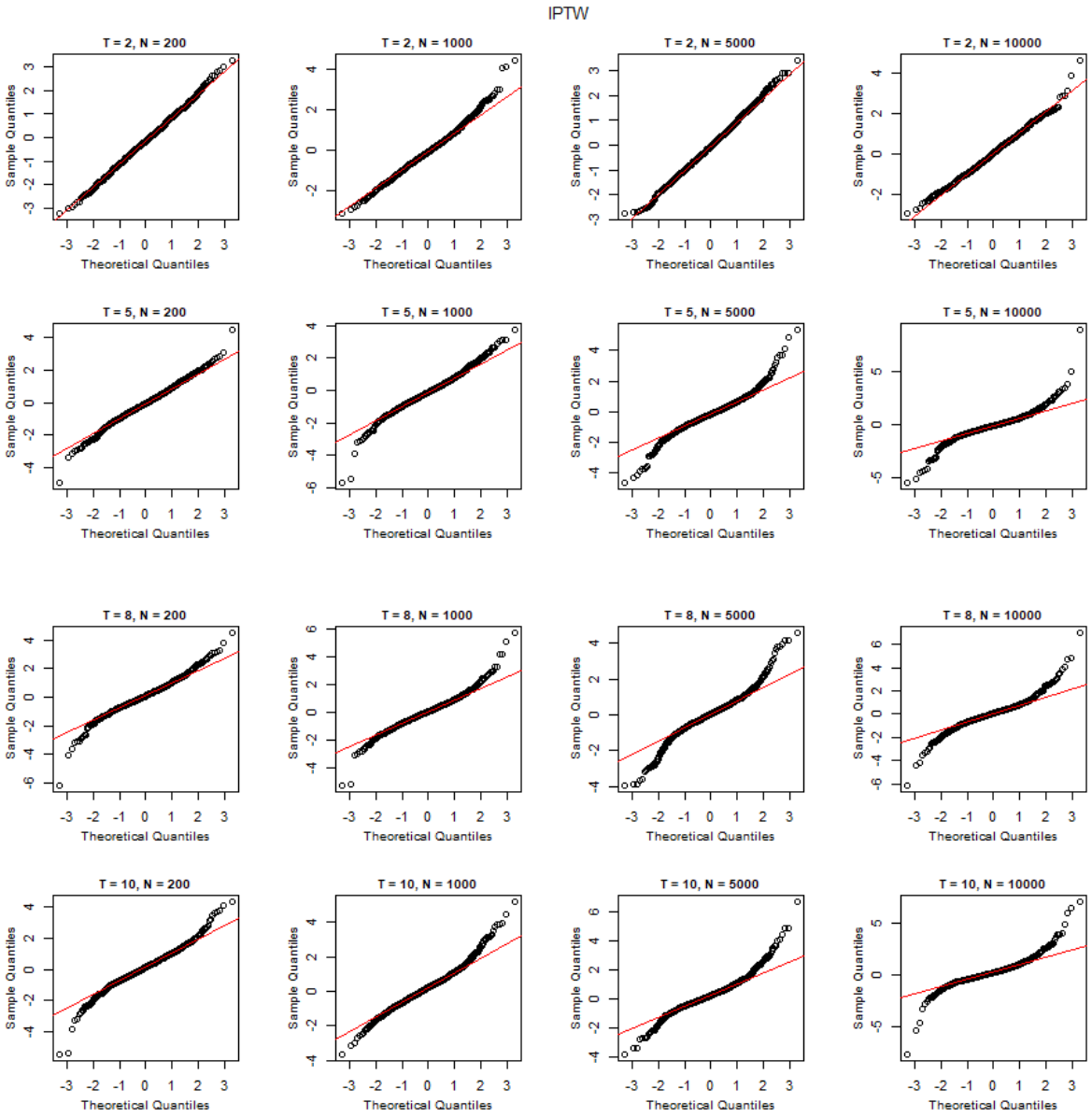


Figure 3.2: Distributions of estimate error at $T = 2, 5$.

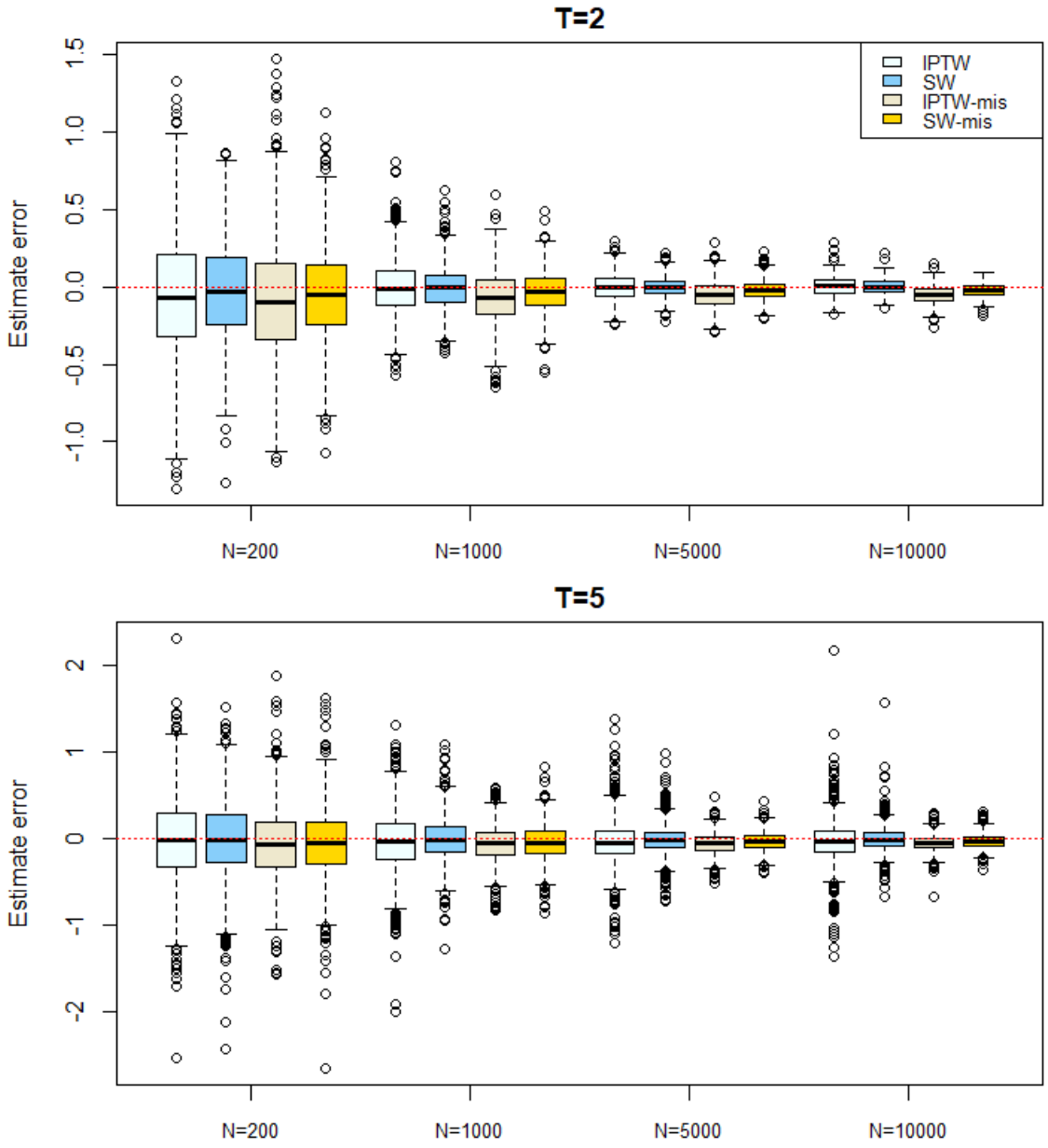
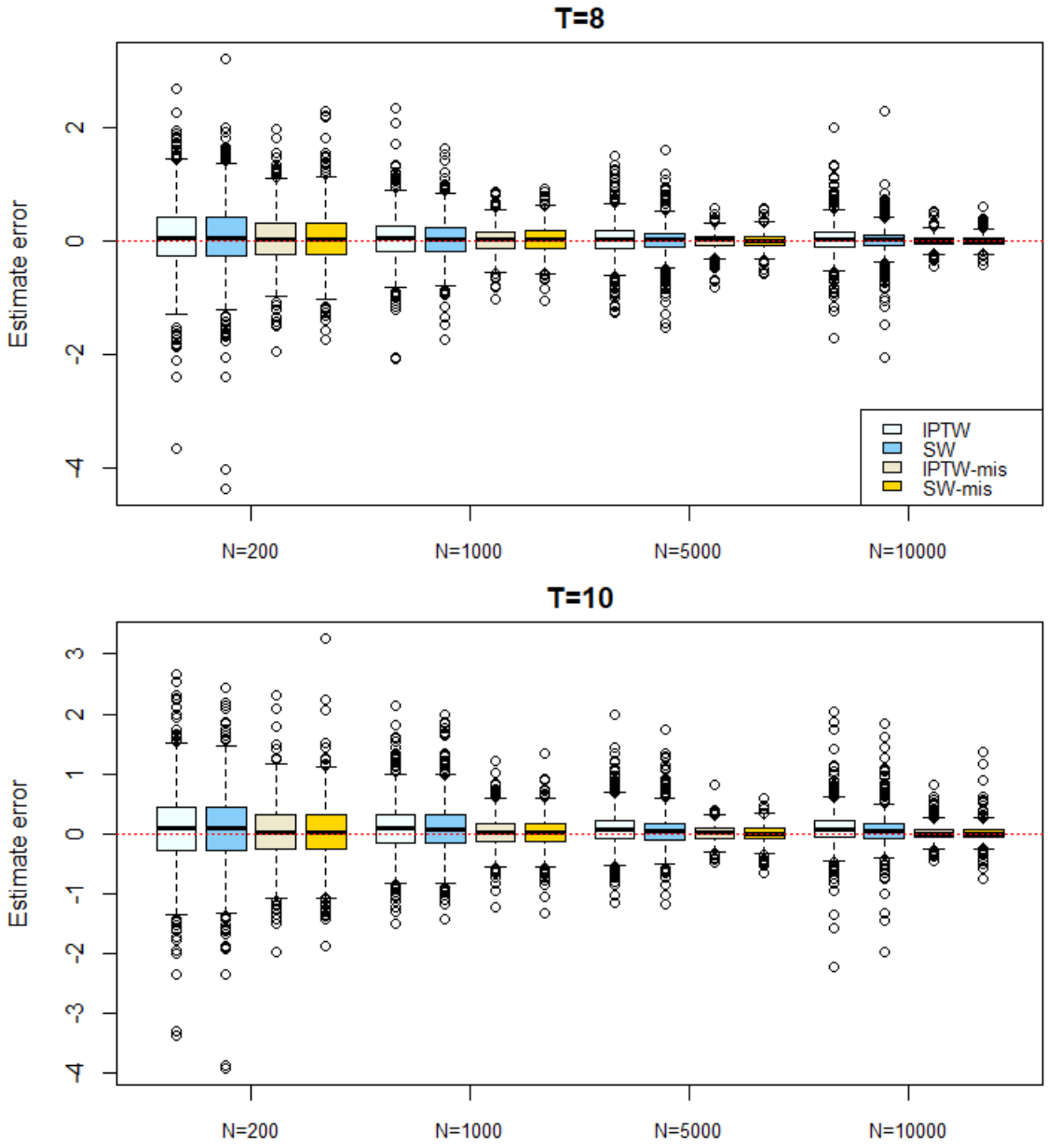


Figure 3.3: Distributions of estimate error at $T = 8, 10$.



% Largest weight % Largest weight (%LW) is defined as the relative percentage of the largest weight to the sum of all weights in the sample. It measures the influence of the subject with the largest weight on MSM estimates. Figure 3.4 shows that both T and N affect %LW. The median %LW increased with T and decreased with N, indicating that the weights are more equally distributed in larger samples with fewer timepoints. In IPTW-mis and SW-mis, the median %LW were lower than in IPTW and SW and the effect of T on %LW lessened as N increased.

In geneneral, %LW is positively correlated with absolute estimate error ($|\hat{\beta}_c - \beta_c|$), more so in IPTW (Figure 3.5) and SW (Figure 3.6) than in IPTW-mis (Figure 3.7) and SW-mis (Figure 3.8). For IPTW and SW, the correlation increased with T and N (except at $T = 2$), reaching 0.8 and 0.82 respectively at $T = 10, N = 10,000$. On the other hand, IPTW-mis and SW-mis showed weaker correlations, reaching 0.7 and 0.78, respectively, at $T = 10, N = 10,000$. The higher correlation can be attributed to the larger largest %LW in IPTW. At $T = 10, N = 10,000$, the largest %LW for IPTW and SW were 98.4% and 89.4%, compared to 44.5% for IPTW-mis and 57.7% for SW-mis.

Standard error The standard errors (SEs) increased with T. At $T = 2$, SW and SW-mis had smaller SEs than IPTW and IPTW-mis (Figure 3.9). At the other timepoints however, IPTW-mis and SW-mis had smaller SEs than IPTW and SW, indicating that MSM estimates obtained with IPTW-mis and SW-mis were more stable than estimates obtained by IPTW and SW.

Coverage probability The coverage probability (CP) was generally higher for SW and SW-mis than IPTW and IPTW-mis (Figure 3.10). The CPs of SW and SW-mis ranged from 0.90 to 0.97, whereas the CPs of IPTW and IPTW-mis ranged from 0.83 to 0.94.

Figure 3.4: Median % largest weights.

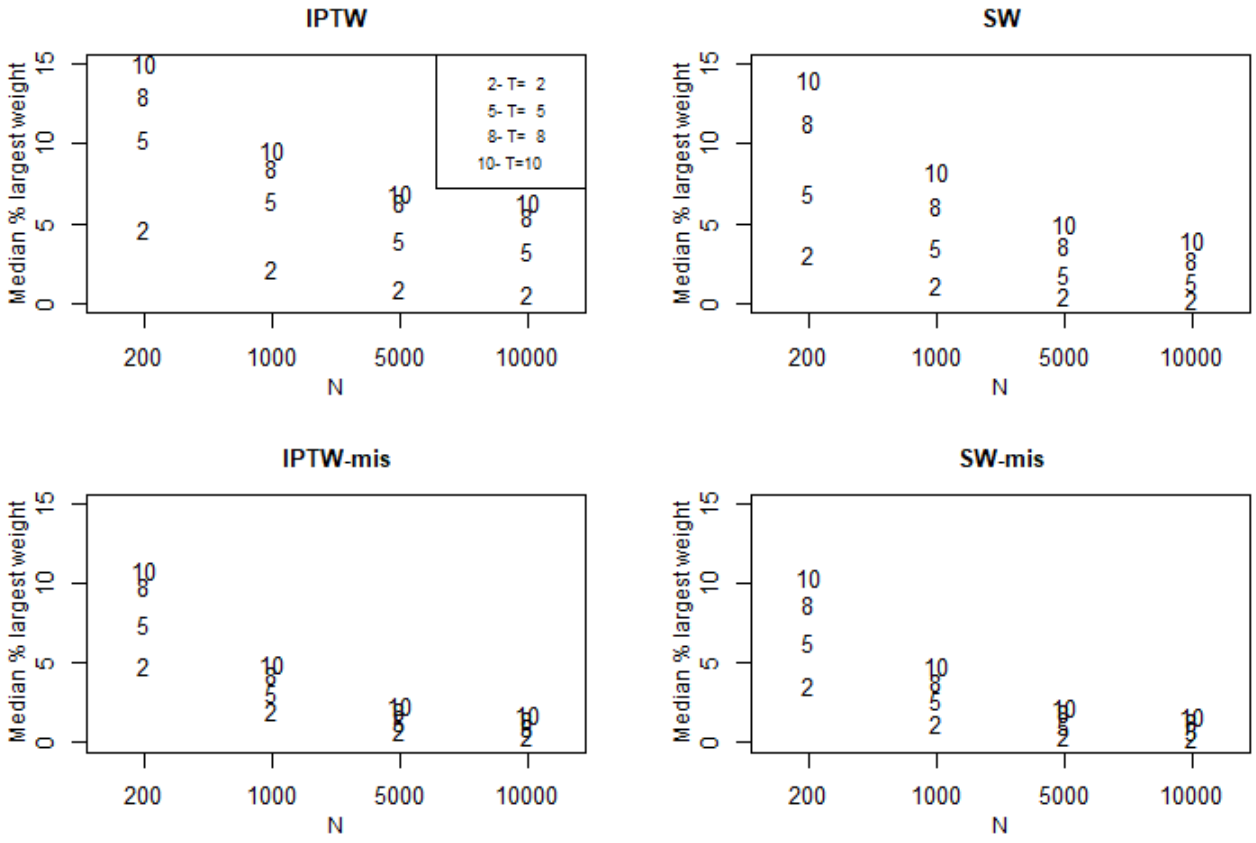


Figure 3.5: Correlation between %LW and absolute estimate error for IPTW. Green is the median %LW and red is the fitted regression.

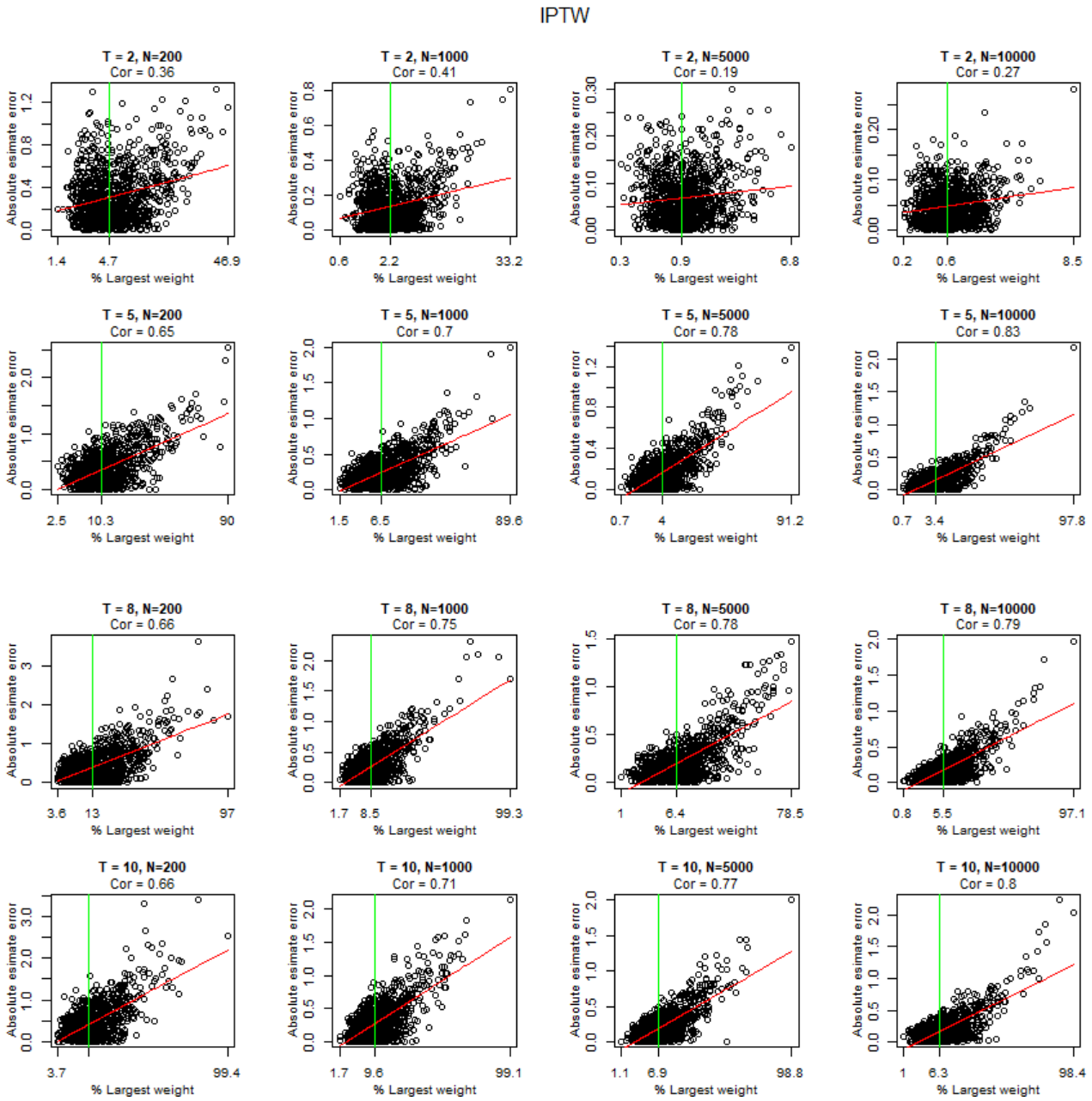


Figure 3.6: Correlation between %LW and absolute estimate error for SW. Green is the median %LW and red is the fitted regression.

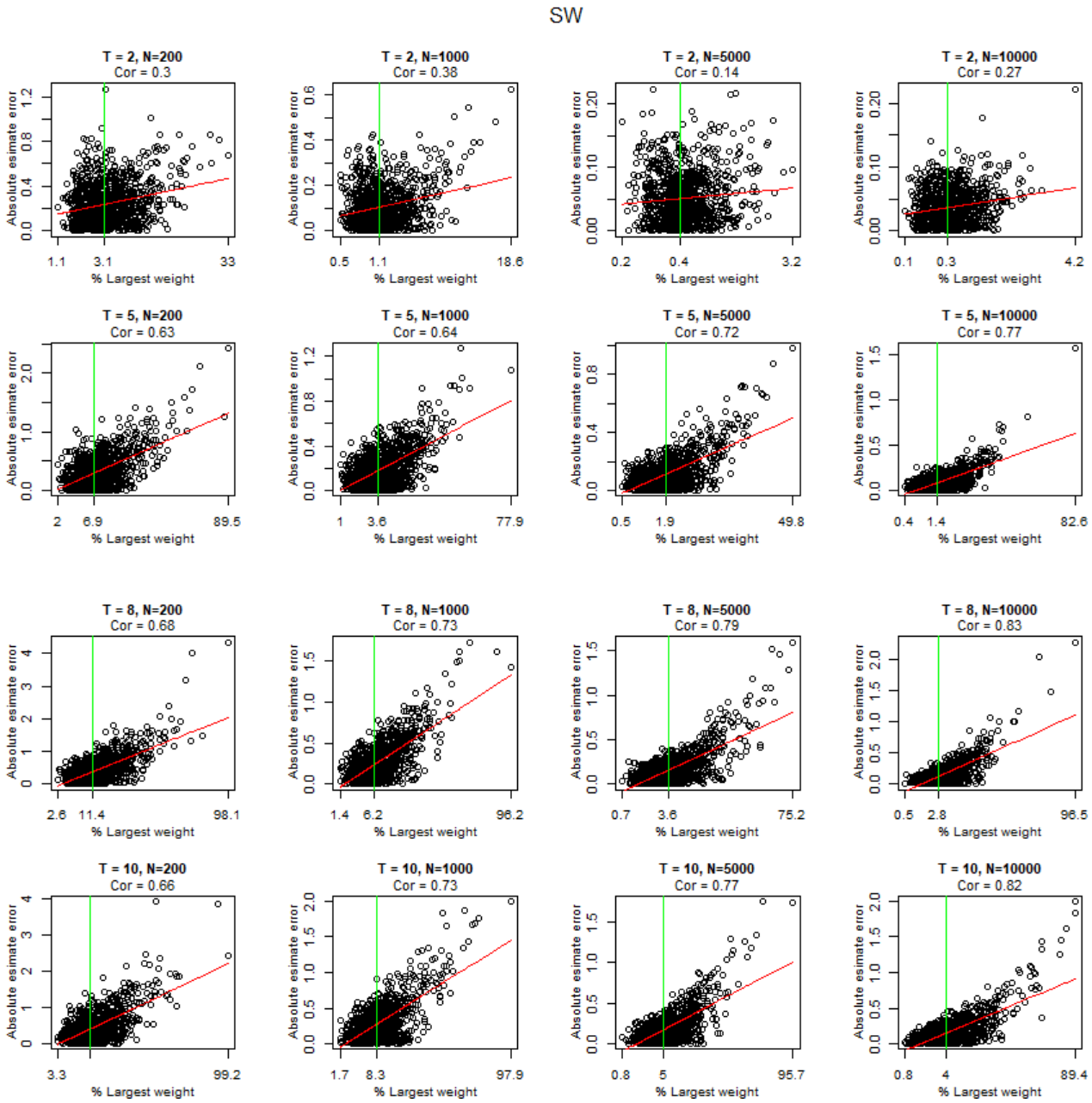


Figure 3.7: Correlation between %LW and absolute estimate error for IPTW-mis. Green is the median %LW and red is the fitted regression.

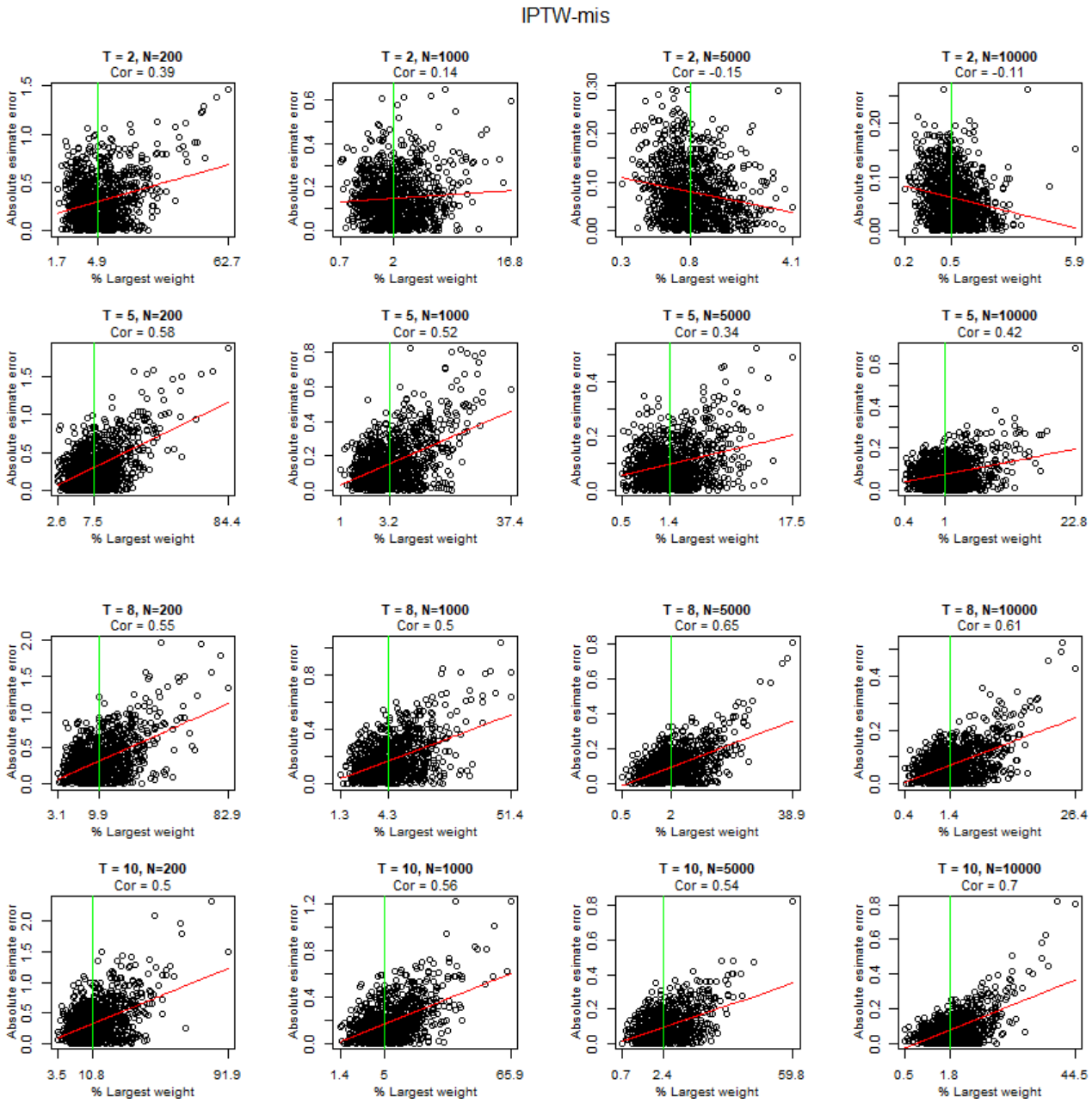


Figure 3.8: Correlation between %LW and absolute estimate error for SW-mis. Green is the median %LW and red is the fitted regression.

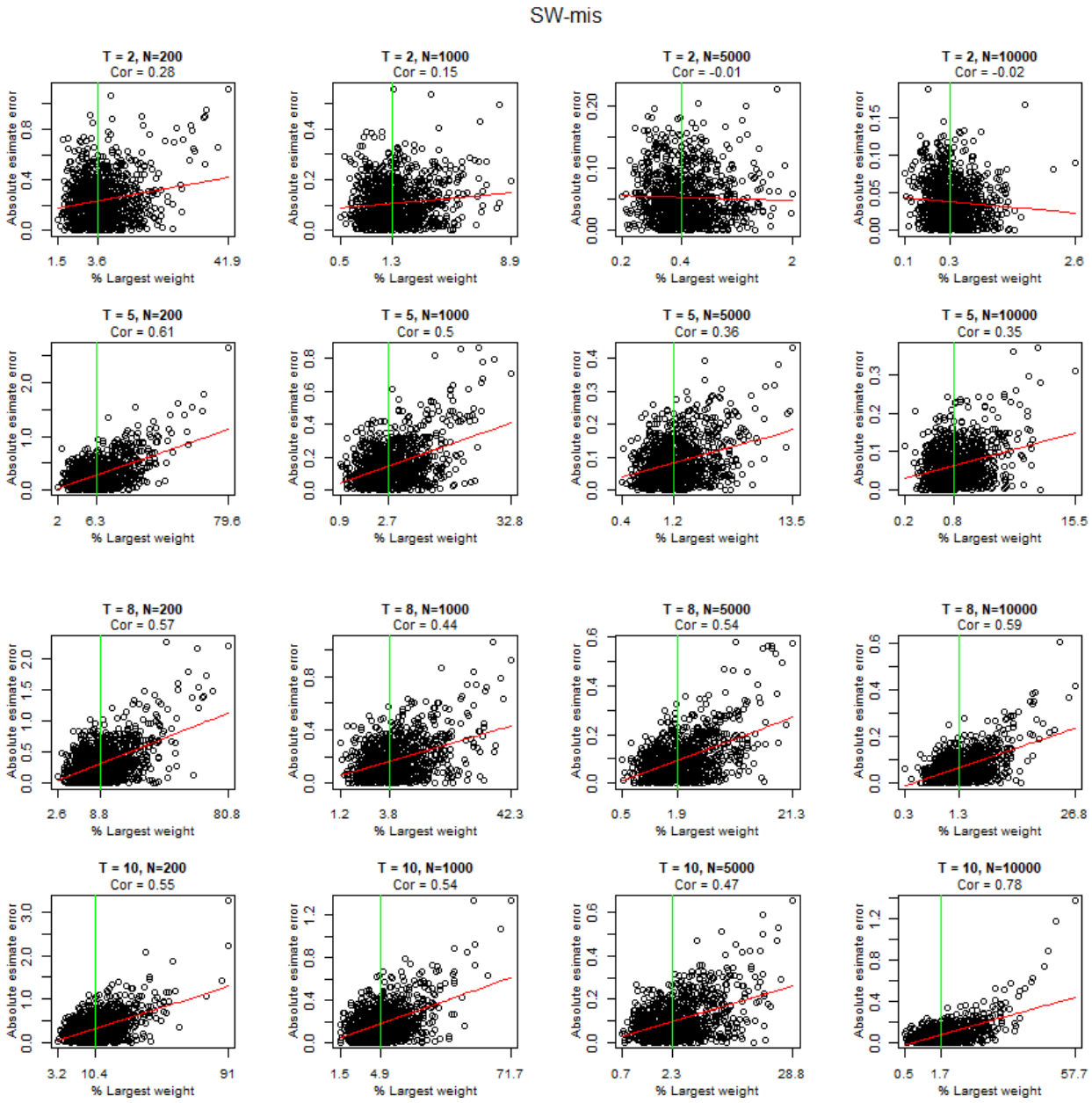


Figure 3.9: Standard errors

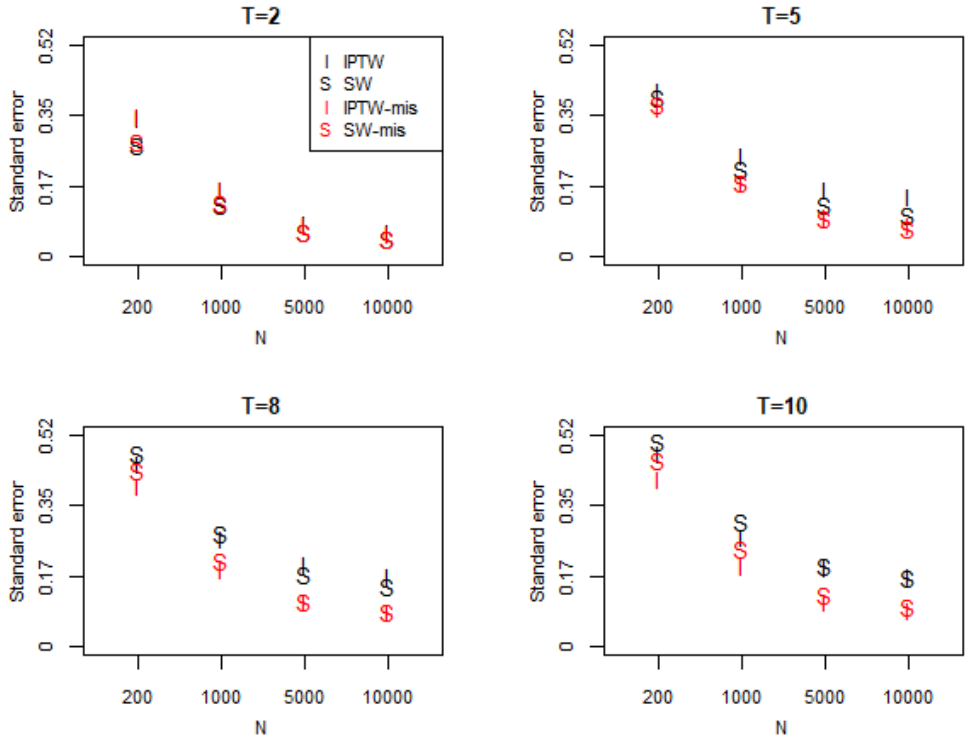
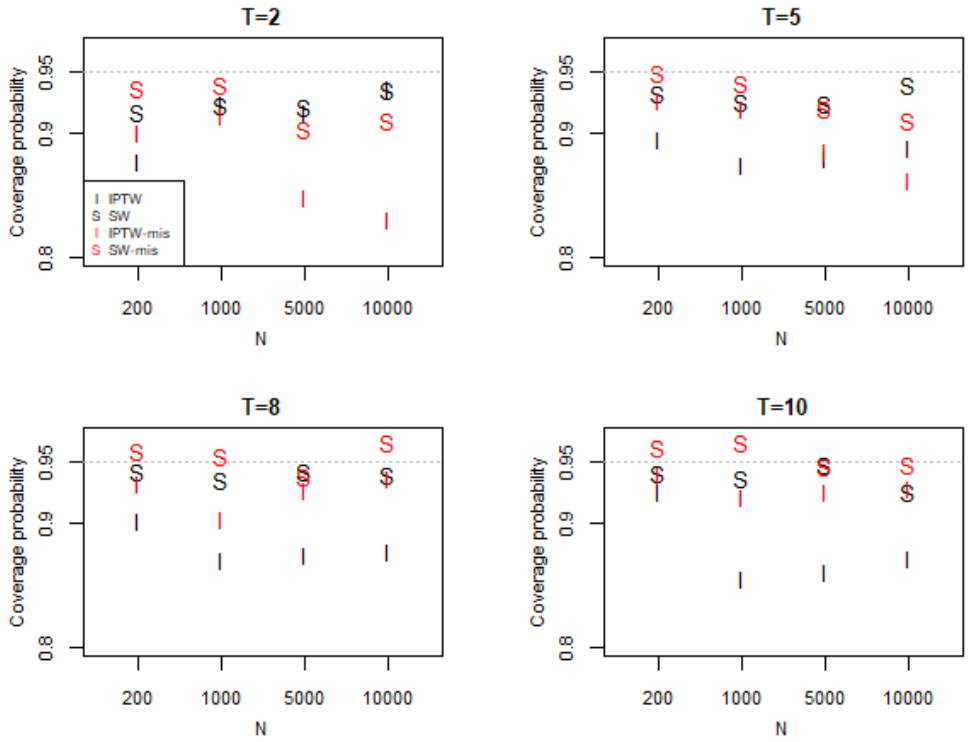


Figure 3.10: Coverage probability



3.5. Real-data example Using the child obesity data from Chapter 1 [5, unpublished data, 2022], we illustrate the interpretation of the results presented in the previous section. The electronic health record (EHR) data consists of primary-care visits of 12,754 children with overweight or obesity in greater Dallas. The treatment variable in this dataset is the clinician's attention to the child's high body mass index (BMI). The binary status of attention to BMI is determined by an electronic phenotype approach. The electronic phenotype sought evidence of guideline-recommended weight-management clinical practices using specific text associated with numeric International Classification of Diseases (ICD) codes (from billing and problem-list codes), education codes for primary-care obesity counseling, and referrals to nutrition, weight-management, and bariatric surgery. The binary outcome, observed 1 month to 2.5 years after attention to BMI, is the status of change in percent of BMI 95^{th} percentile ($\%\text{BMI}_{\text{p95}}$) from the baseline $\%\text{BMI}_{\text{p95}}$, equaling 1 if $\%\text{BMI}_{\text{p95}}$ decreased by more than 0.5%. At timepoint t, covariates used to estimate propensity scores p_t included age, sex, race, $\%\text{BMI}_{\text{p95}}$ (observed at time t), medication associated with weight gain, medication associated with weight loss, whether the child had a well-child care visit in the past 12 months, whether the child had a sick visit in the past 12 months, whether the current visit was in the summer, and the treatment status at the previous timepoint $(t-1)$.

We took subsets of the data by the number of timepoints, $T = 2, 5, 8, 10$. In each subset, we estimated an MSM that modeled the outcome with a common treatment effect for treatments at all timepoionts: $\mathbb{E}[Y] = \text{expit}\left(\beta_0 + \beta_c \sum_{t=1}^{T} A_t\right)$. **Table 3.2** shows the odds ratio (OR), 95% confidence interval (CI), and % largest weight (%LW) for each subset. The confidence intervals were constructed using the bootstrap method variance estimation (250 samples). Here, we assume that the propensity score models used for this data are misspecified. Based on the findings in Section 3.4, we focus on estimates obtained using SW-mis.

The estimated OR 1.11 at $T = 2$ can be considered more reliable than at other timepoints due to the advantages of a small number of timepoints and a large sample size ($N > 10,000$). In Section 3.4, we saw that, for $T = 2, N = 10,000$, the distribution of estimate error was close to 0 with few outliers. However, because we also saw in Section 3.4 that the coverage probabilities at $T = 2, N = 10,000$ did not exceed 0.95, it is possible that the 95% CI using SW, $(1.02, 1.18)$, should be wider.

The estimated ORs at other timepoints should be accepted with more caution not only because we saw that there were more outlying estimate errors at $T = 5, 8, 10$ (Section 3.4), but also because the %LWs in Table 3.2 are large compared to the largest %LWs of the corresponding (approximately) T and N in Figure 3.8. For SW-mis, the reported correlations between %LW and absolute estimate error are not large, but they are also not negligible.

Table 3.2: Estimated odds ratios of improvement in $\%BMI_{p95}$ for children who received BMI attention. Propensity score models are assumed to be misspecified.

T	N	IPTW-mis			SW-mis		
		OR	95% CI	%LW	OR	95% CI	%LW
2	11413	1.11	(1.04, 1.20)	0.33%	1.10	(1.02, 1.18)	0.27%
5	6821	1.17	(1.01, 1.34)	1.78%	1.23	(0.98, 1.55)	13.52%
8	3357	1.25	(0.90, 1.74)	15.5%	1.25	(0.87, 1.81)	31.24%
10	2025	0.83	(0.62, 1.10)	16.6%	0.82	(0.61, 1.12)	39.49%

3.6. Discussion This simulation study assessed the the validity of MSM estimates obtained under multiple timepoints using the correct and misspecified propensity score models and inducing mild confounding in the generated data. Data were weighted using two methods of weights: inverse-probability-of-treated weights (IPTW) and stabilized weights (SW).

General findings MSM estimates were affected by the number of timepoints. The distribution of normalized estimates (against the true effect) were approximately normal under $T = 2$ but became increasingly thick-tailed under $T = 5, 8,$ and 10. The thick-tailed distribution of estimates went hand-in-hand with estimate error. In later timepoints, the occurrences of large estimate errors did not diminish with increasing sample sizes. %LW also increased with with respect to T and N. As %LW increased, the correlation between %LW and estimate error also increased. The increasing positive correlation with respect to T and N suggests that in a large dataset with multiple timepoints, the presence of large weights carries more potential for an inaccurate estimate than in a smaller sample with less timepoints.

IPTW, IPTW-mis vs. SW, SW-mis SW is generally recommended over IPTW [20,38]. Using binary data, our study results confirmed this recommendation, consistent with the results of Xu *et al*'s study using risk ratios as the outcome [44]. The advantage of SW manifested most prominently in the coverage probability. The coverage probabilities of SW and SW-mis were always higher than IPTW and IPTW-mis.

SW vs. SW-mis There were notable differences between SW and SW-mis. Estimates obtained using SW-mis were closer to the true parameter than estimates obtained using SW. For SW-mis, %LW were smaller and there was less correlation between %LW and estimate error. The standard errors were also smaller but the coverage probabilities were higher.

The reason for the superior performance of SW over IPTW, and of SW-mis over SW may lie in the magnitude of extreme weights. Compared to IPTW, SW is less extreme. In our study, weights produced under misspecified propensity score models were less extreme because the misspecified propensity score model did not have the third-degree polynomial term present in the true propensity score model. Thus, the resulting propensity scores under

misspecified models varied less than the true propensity scores and yielded less extreme weights.

A possible alternative to SW is the overlap weights (OW) introduced by Li *et al* [39]. Unlike IPTW and SW, the overlap weights are bounded. The next plausible simulation study would be to assess MSM estimates obtained by SW and OW under varying propensity score models.

Appendix A

Appendix of Chapter 3

Figure 1.1: Distribution of $\hat{\beta}_c$ obtained using SW.

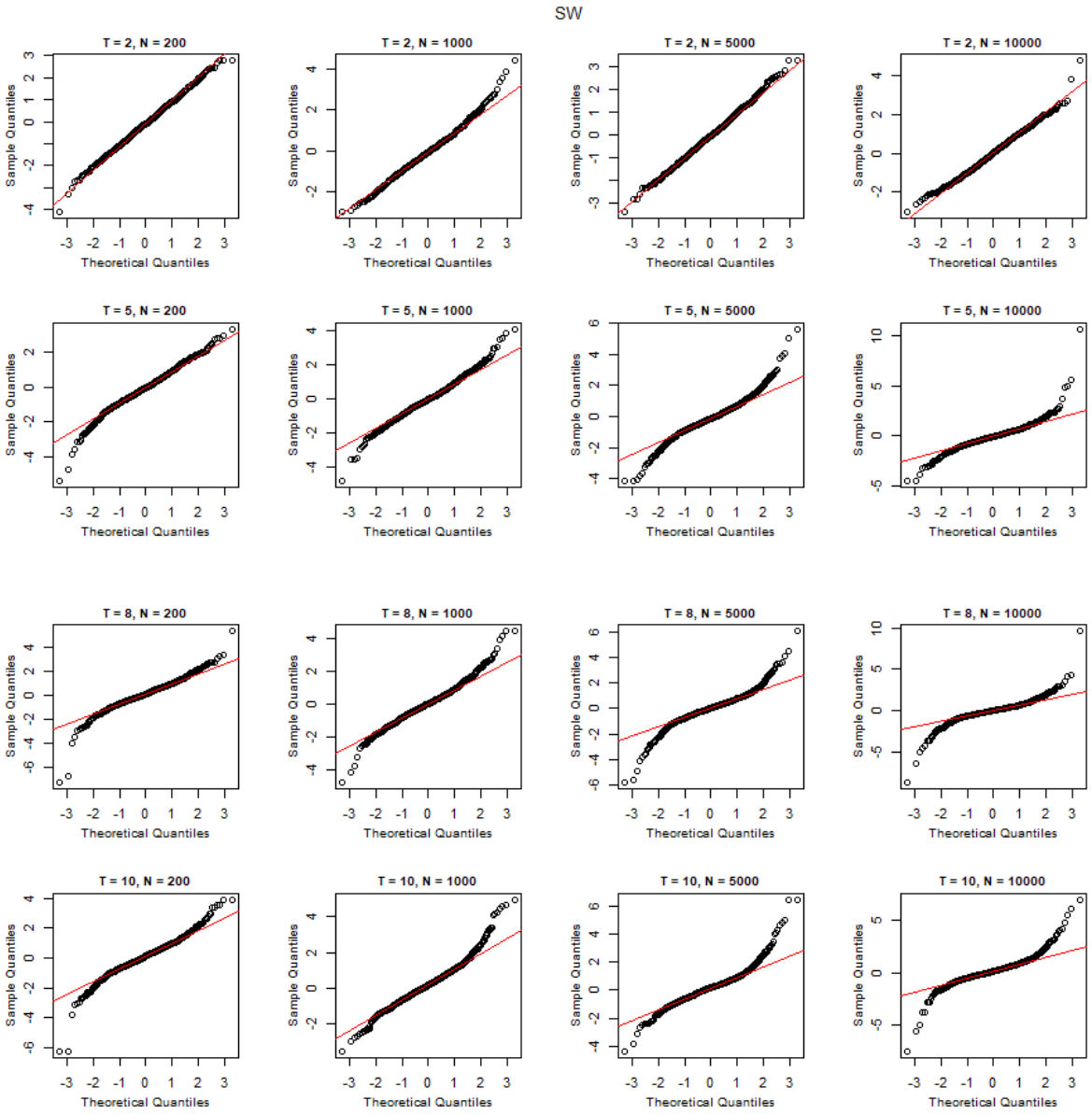


Figure 1.2: Distribution of $\hat{\beta}_c$ obtained using IPTW-mis.

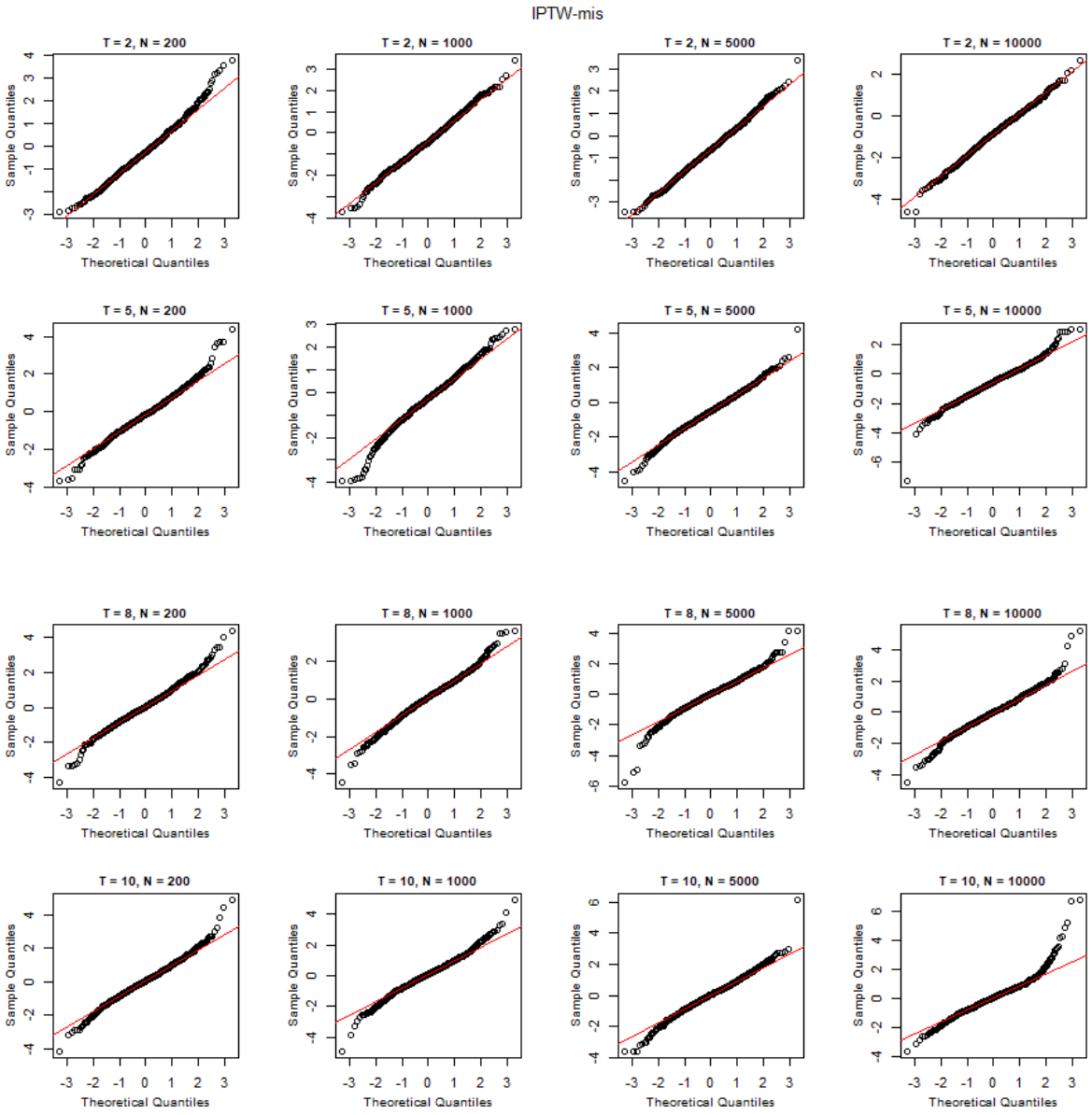
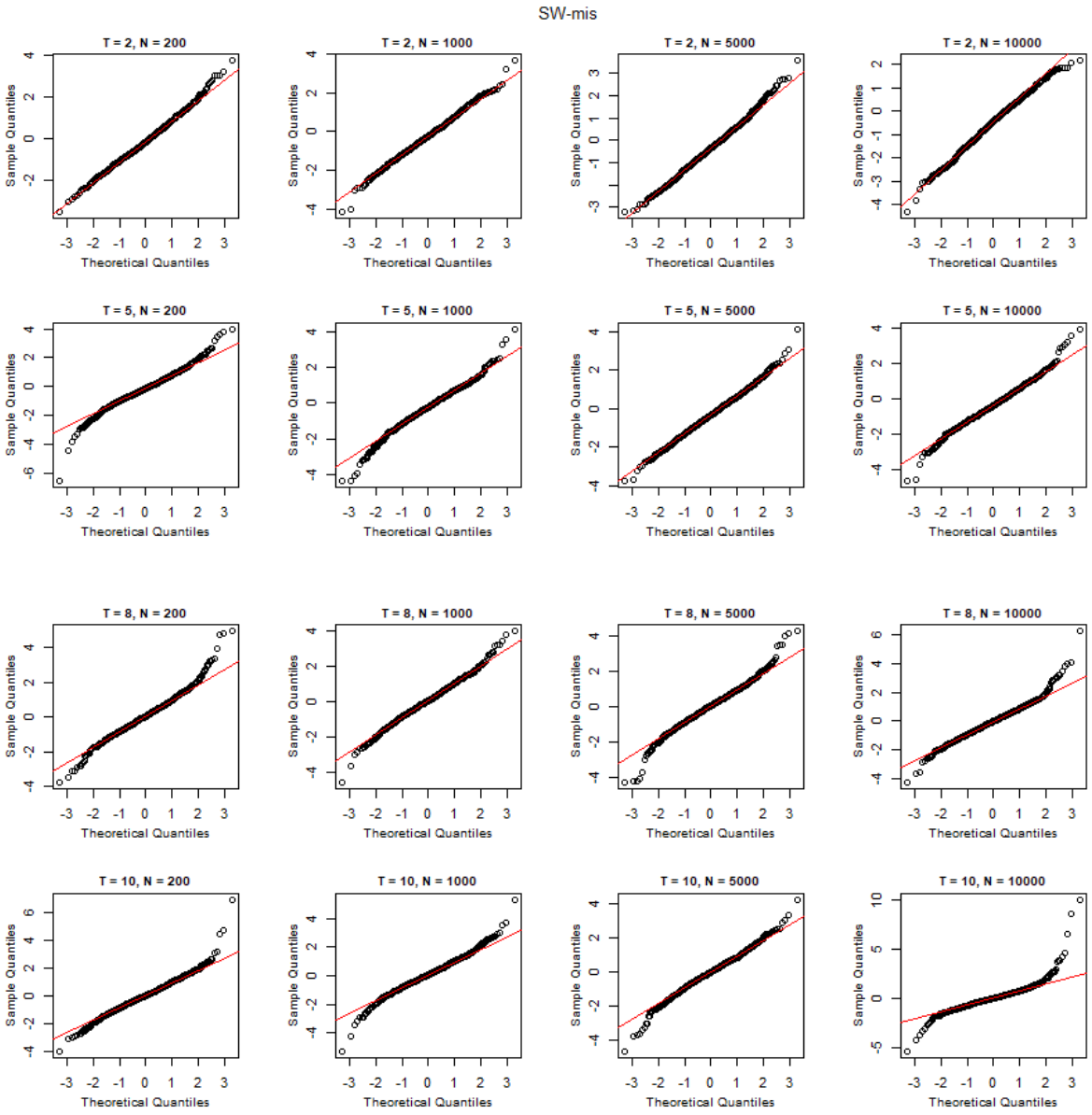


Figure 1.3: Distribution of $\hat{\beta}_c$ obtained using SW-mis.



BIBLIOGRAPHY

[1] C. Norlin, M. Crawford, C. Bell et al., Delivery of well-child care: A look inside the door, *Acad Pediatr* **11** (2011) 18–26. 1

[2] K. Manning, A. Ariza, T. Massimino et al., Health supervision visits of very young children: Time addressing 3 key topics, *Clin Pediatr (Phila)* **48** (2009) 931. 1

[3] S. Barlow and E. Committee, Expert committee recommendations regarding the prevention, assessment, and treatment of child and adolescent overweight and obesity: summary report, *Pediatrics* **120** (2007) Suppl 4:S164–92. 1

[4] B. Dennison, Bright futures and NHLBI integrated pediatric cardiovascular health guidelines, *Pediatr Ann.* (2012) e31–6. 1

[5] C. B. Turer, S. E. Barlow, D. B. Sarwer, B. Adamson, J. Sanders et al., Association of clinician behaviors and weight change in school-aged children, *American Journal of Preventive Medicine* **000** (2019) 1–10. 1, 2, 12, 24, 52

[6] G. Imbens and D. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences.* New York, NY: Cambridge University Press, 2015:257-306. 2, 6, 12

[7] D. Rubin, Using propensity scores to help design observational studies: Application to the tobacco litigation, *Health Services and Outcomes Research Methodology* **2** (2001) 169. 2, 6

[8] C. Turer, H. Lin and G. Flores, Health status, emotional/behavioral problems, health care use, and expenditures in overweight/obese US children/adolescents, *Acad Pediatr.* **13** (2013) 251–258. 2, 3

[9] Centers for Disease Control and Prevention, "Defining childhood weight status." https://www.cdc.gov/obesity/basics/childhood-defining.html, Aug., 2022. 2

[10] K. Kainz, N. Greifer, A. Givens et al., Improving causal inference: Recommendations for covariate selection and balance in propensity score methods, *Journal of the Society for Social Work and Research* **8** (2017) 279–303. 4, 6

[11] J. Pearl, Invited commentary: Understanding bias amplification, *American Journal of Epidemiology* **174** (2011) 1223–1227. 4, 6

[12] D. Freedman and G. Berenson, Tracking of BMI z scores for severe obesity, *Pediatrics (Evanston)* **140** (2017) e20171072. 5

[13] D. Freedman, N. Butte, E. Taveras et al., BMI z-Scores are a poor indicator of adiposity among 2- to 19-year-olds with very high BMIs, NHANES 1999-2000 to 2013-2014, *Obesity (Silver Spring, MD)* **25** (2017) 739–746. 5

[14] C. Daymont, M. Ross, L. Russell et al., Automated identification of implausible values in growth data from pediatric electronic health records, *J Am Med Infom Assoc.* **24** (2017) 1080–1087. 5

[15] J. Lunceford and M. Davidian, Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study, *Statistics in Medicine* **23** (2004) 2937–2960. 6

[16] M. Neuhäuser, M. Theilmann and G. Ruxton, The number of strata in propensity score stratification for a binary outcome, *Archives of Medical Science* **14** (2018) 695–700. 6

[17] SAS Institute Inc, *SAS/STAT 15.2 User's Guide*. Cary, NC: SAS Institute Inc, 2020:8178-8295. 6

[18] P. Rosenbaum and D. Rubin, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome, *J R Stat Soc Series B Stat Methodol* **45** (1983) 212–218. 7

[19] J. Pearl, Causal inference in statistics: An overview, *Statist Surv* (2009) 96–146. 12

[20] J. M. Robins, M. A. Hernán and B. Brumback, Marginal structural models and causal inference in epidemiology, *Epidemiology* **11** (2000) 550–560. 14, 16, 17, 37, 38, 54

[21] M. A. Hernán and R. J. M, *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020. 14

[22] M. H. Gail, S. Wieand and S. Piantadosi, Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates, *Biometrika* **71** (1984) 431–444. 15, 18

[23] W. W. Hauck, S. Anderson and S. Marcus, Should we adjust for covariates in nonlinear regression analyses of randomized trials?, *Controlled Clinical Trials* **19** (1998) 249–256. 15, 18

[24] J. G. Young, M. A. Hernán, S. Picciotto and J. M. Robins, Simulation from structural survival models under complex time-varying data structure, *JSM Proceedings, section on statistics in epidemiology, American Statistical Association, Denver, CO* (2008) . 15

[25] J. G. Young, M. A. Hernán, S. Picciotto and J. M. Robins, Relation between three classes of structural models for the effect of time-varying exposure on survival, *Lifetime Data Anal.* **16** (2010) 71–84. 15

[26] G. Vourli and G. Touloumi, Performance of the marginal structural models under various scenarios of incomplete maker's values: A simulation study, *Biometrical Journal* **2** (2015) 254–270. 15

[27] C. Lusivika-Nzinga, H. Selinger-Leneman, S. Grabar, D. Costagliola and F. Carrat, Performance of the marginal structural cox model for estimating individual and joined effects of treatments given in combination, *BMC Medical Research Methodology* **17** (2017) . 15

[28] P. C. Austin and J. Stafford, The performance of two data-generation processes for data with specified marginal treatment odds ratios, *Communications in Statistics–Simulation and Computation* **37** (2008) 1039–1051. 15

[29] P. C. Austin, Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis, *Statistics in Medicine* **35** (2016) 5642–5655. 15, 28, 36

[30] P. C. Austin and E. A. Stuart, The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect treatment on survival outcomes, *Statistical Methods in Medical Research* **26** (2017) 1654–1670. 15, 20, 23, 27, 36

[31] W. G. Havercroft and V. Didilez, Simulating from marginal structural models with time-dependent confounding, *Statistics in Medicine* **31** (2012) 4190–4206. 15

[32] R. Bender, T. Augustin and M. Blettner, Generating survival times to simulate Cox proportional hazard models, *Statistics in Medicine* **24** (2005) 1713–1723. 20, 23

[33] P. C. Austin, The performance of different propensity score methods for estimating marginal hazard ratios, *Statistics in Medicine* **32** (2013) 2837–2849. 20, 23

[34] P. C. Austin and T. Schuster, The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: A simulation study, *Statistical Methods in Medical Research* **25** (2016) 2214–2237. 20, 23

[35] P. C. Austin and D. S. Small, The use of bootstrapping when using propensity-score matching without replacement: a simulation study, *Statistics in Medicine* **33** (2014) 4306–4319. 20, 23

[36] P. Austin, P. Grootendorst, S. Normand and G. Anderson, Conditioning on the proopensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study, *Statistics in Medicine* **26** (2007) 754–768. 20

[37] E. A. Stuart, Matching methods for causal inference: a review and a look forward, *Institute of Mathematical Statistics* **25** (2010) 1–21. 37

[38] S. R. Cole and M. A. Hernán, Constructing inverse probability weights for marginal structural models, *American Journal of Epidemiology* **168** (2008) 656–664. 37, 38, 54

[39] F. Li, L. Thomas and F. Li, Addressing extreme propensity scores via the overlap weights, *American Journal of Epidemiology* **188** (2019) 250–257. 37, 55

[40] F. Li, K. Morgan and A. Zaslavsky, Balancing covariates via propensity score weighting, *Journal of American Statistical Association* **113** (2018) 390–440. 37

[41] F. Potter, The effect of weight trimming on nonlinear survey estimates, *Proceedings of the Section on Survey Research Methods of American Statistical Association, San Francisco, CA* (1993) e18174. 37

[42] D. Scharfstein, A. Ritnitzky and J. Robins, Adjusting for non-ignorable drop-out using semiparametric non-response models, *Journal of American Statistical Association* **94** (1999) 1096–1120. 37

[43] B. K. Lee, J. Lessler and E. A. Stuart, Weight trimming and propensity score weighting, *PLOS One* **6** (2011) e18174. 37

[44] S. Xu, C. Ross, M. A. Raebel, S. Shetterly, C. Blanchette and D. Smith, Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals, *Vaue in Health* **13** (2010) 273–277. 54

[45] W. Cochran and D. Rubin, Controlling bias in observational studies: A review, *Sankhya Series A* **35** (1973) 417–446.

[46] M. Pang, J. S. Kaufman and P. R. W, Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models, *Statistical Methods in Medical Research* **25** (2016) 1925–1937.

[47] A. S. Whittmore, Collapsibility of multidimensional contingency tables, *J R Stat Soc Series B Stat Methodol* **40** (1978) 328–340.

[48] G. R. Ducharme and Y. LePage, Testing collapsibility in contingency tables, *J R Stat Soc Series B Stat Methodol* **48** (1986) 197–205.

[49] S. Greenland and R. M. Mickey, Closed-form and dually consistent methods for inference on collapsibility in 2*2K and 2*J*K tables, *J R Stat Soc Series C Appl Stat* **37** (1988) 335–343.

[50] S. Greenland, J. M. Robins and J. Pearl, Confounding and collapsibility in causal inference, *Stat Sci* **14** (1999) 29–46.

[51] S. C. Newman, *Biostatistical methods in epidemiology*. Wiley, New York, 2001.

[52] S. Greenland and H. Morgenstern, Confounding health research, *Annu. Rev. Public Health* **22** (2001) 189–212.

[53] S. Greenland, Noncollapsibility, confounding, and sparse-data bias. Part 2: What should researchers make of persistent controversies about the odds ratio?, *Journal of Clinical Epidemiology* **139** (2012) 264–268.

[54] G. Lefebvre, J. A. C. Delaney and R. W. Platt, The impact of mis-specification of the treatment models on estimates from a marginal structural model, *Statistics in Medicine* **27** (2008) 3629–3642.

[55] D. Talbot, J. Atherton, A. M. Rossi, S. L. Bacon and L. G, A cautionary note concerning the use of stabilized weights in marginal structural models, *Statistics in Medicine* **34** (2015) 812–823.

[56] S. Greenland and H. Morgenstern, Confounding in health research, *Annu Rev Public Health* **22** (2001) 189–212.

[57] S. Greenland, Interpretation and choice of effect measures in epidemiologic analyses, *American Journal of Epidemiology* **125** (1987) 761–768.