Fall 12-22-2010

Causal Inference with Two-Stage Logistic Regression - Accuracy, Precision, and Application

Bing Cai
School of Medicine, bingcai@hotmail.com

Causal Inference with Two-Stage Logistic Regression - Accuracy, Precision, and Application

Abstract
Two-stage predictor substitution (2SPS) and the two-stage residual inclusion (2SRI) are two approaches to instrumental variable (IV) analysis. While 2SPS and 2SRI with linear models are well-studied methods of causal inference, the properties of 2SPS and 2SRI for logistic binary outcomes have not been thoroughly studied. We study the bias and variance properties of 2SPS and 2SRI for a logistic outcome model so that we can apply these IV approaches to the causal inference of binary outcomes. We also propose and implement an extension of generalized structure mean model originally developed for a randomized trial. We first present closed form expressions of asymptotic bias for the causal odds ratio from both 2SPS and 2SRI approaches. Our closed form bias results show that the 2SPS logistic regression generates asymptotically biased estimates of this causal odds ratio when there is no unmeasured confounding and that this bias increases with increasing unmeasured confounding. The 2SRI logistic regression is asymptotically unbiased when there is no unmeasured confounding, but when there is unmeasured confounding, there is bias and it increases with increasing unmeasured confounding. In the second part, we propose the sandwich variance estimator of logistic regression of both 2SPS and 2SRI approaches and the variance estimator is adjusted for the fact that the estimates from the first stage regression is included as covariates in the second stage regression. The simulation results show that the adjusted estimates are consistent with the observed variance while the naive estimates without the adjustments are biased. This study also shows that the 2SRI method has a larger variance than the 2SPS method. Lastly, we compare the 2SPS and 2SRI logistic regression with the generalized structure mean model (GSMM). Our simulation results show that the GSMM is an unbiased estimator of complier-average causal effect (CACE) and has the least variance among the three approaches. We apply these three methods to the analysis of the GPRD database on antidiabetic effect of bezafibrate.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Epidemiology & Biostatistics

First Advisor
Thomas R. Ten Have

Second Advisor
Dylan S. Small

Keywords
Causal Inference, Two-stage logistic regression, bias, variance

CAUSAL INFERENCE WITH TWO-STAGE LOGISTIC REGRESSION -ACCURACY, PRECISION, AND APPLICATION

Bing Cai

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2010

Supervisor of Dissertation

*Signature*_____

Thomas R. Ten Have

Professor of Biostatistics

Co-Supervisor

*Signature*_____

Dylan S. Small

Associate Professor of Statistics

Graduate Group Chairperson

*Signature*_____

Daniel F. Heitjan, Professor of Biostatistics

Dissertation Committee

Peter Groeneveld, Assistant Professor of Medicine

Sean Hennessy, Associate Professor of Epidemiology

Dylan S. Small, Associate Professor of Statistics

Thomas R. Ten Have, Professor of Biostatistics

Peter W. Yang, Assistant Professor of Biostatistics

In the memory of my grandparents and my father.

In the memory of Professor Harry Guess.

Harry was our admired leader and he is always my role model of epidemiologist. He recommended me to this Ph.D. program of biostatistics at University of Pennsylvanian in 2004. I wish he knows in spirit that I have successfully finished my Ph.D. Study.

Acknowledgments

For the last six to seven years, finishing my Ph. D while working full time has been a very arduous journey. I would have never made it to the end without the support of my mentors, friends, colleagues, and family.

My deep gratitude to the following people:

First, to Dr. Thomas Ten Have and Dr. Dylan Small for their combined effort to advise my dissertation. Not only did I benefit from their advisory, but also from their discussions. When I listened to them debate about certain topics, I realized that their interaction is the essence of amazing scientific discovery. I will never forget this experience studying under the supervision of Dr. Ten Have and Dr. Small with countless meetings including those on sunny days in summer and raining and windy days in winter, when one of them went to the other's office for our weekly meeting.

To Dr. Sean Hennessy, who provided important advices to me with his insight in pharmacology and pharmacoepidemiology.

To Dr. Peter Yang and Dr. Peter Groeneveld who provided me with statistical and clinical input respectively.

To Dr. James Flory and Dr. Kevin Haynes for their support of the GPRD data.

Many thanks to my colleagues and friends, especially to Drs. Guanghan Liu, Thomas Rhodes, Agnes Baffoe-Bonnie, Douglas Watson, Jay Pearson, Edward Bortnichak and Nancy Santanello for their support and advice, and to Merck Co., Inc. for the financial support.

Finally, my deepest thanks to my wife Jing and my son Brian for their love and support throughout the whole process.

ABSTRACT

CAUSAL INFERENCE WITH TWO-STAGE LOGISTIC REGRESSION

-ACCURACY, PRECISION AND APPLICATION

Bing Cai

Thesis supervisors: Thomas R. Ten Have, MPH, Ph.D. and Dylan S. Small, Ph.D.

Two-stage predictor substitution (2SPS) and the two-stage residual inclusion (2SRI) are two approaches to instrumental variable (IV) analysis. While 2SPS and 2SRI with linear models are well-studied methods of causal inference, the properties of 2SPS and 2SRI for logistic binary outcomes have not been thoroughly studied. We study the bias and variance properties of 2SPS and 2SRI for a logistic outcome model so that we can apply these IV approaches to the causal inference of binary outcomes. We also propose and implement an extension of generalized structure mean model originally developed for a randomized trial. We first present closed form expressions of asymptotic bias for the causal odds ratio from both 2SPS and 2SRI approaches. Our closed form bias results show that the 2SPS logistic regression generates asymptotically biased estimates of this causal odds ratio when there is no unmeasured confounding and that this bias increases with increasing unmeasured confounding. The 2SRI logistic regression is asymptotically unbiased when there is no unmeasured confounding, but when there is unmeasured confounding, there is bias and it increases with increasing unmeasured confounding. In the second part, we propose the sandwich variance estimator of logistic regression of both 2SPS and 2SRI approaches and the variance estimator is adjusted for the fact that the estimates from the first stage

regression is included as covariates in the second stage regression. The simulation results show that the adjusted estimates are consistent with the observed variance while the naive estimates without the adjustments are biased. This study also shows that the 2SRI method has a larger variance than the 2SPS method. Lastly, we compare the 2SPS and 2SRI logistic regression with the generalized structure mean model (GSMM). Our simulation results show that the GSMM is an unbiased estimator of complier-average causal effect (CACE) and has the least variance among the three approaches. We apply these three methods to the analysis of the GPRD database on antidiabetic effect of bezafibrate.

Contents

1 Introduction **1**

2 Bias of Causal Inference for the Odds Ratio Using Two-Stage Instrumental Variable Methods **10**

2.1 Introduction 10

2.2 Assumption and Notation 14

2.3 Bias of Two-Stage Predictor Substitution (2SPS) 17

2.3.1 Probability limit of the estimator 17

2.3.2 Bias analysis 19

2.4 Bias of Two-Stage Residual Inclusion (2SRI) 21

2.4.1 Closed form expression for the probability limit of the estimator 22

2.4.2 Bias analysis 25

2.5 Simulation 26

2.5.1 Simulation algorithm 26

2.5.2 Simulation results 28

2.6 Discussion 29

2.7 Appendix 38

3 Variance Estimate of Causal Odds Ratio with Instrumental Variable Two-Stage Logistic Regression **45**

3.1 Introduction 45

3.2 Notation and parameter setting 49

3.3 Variance estimate of 2 stage logistic regression 50

3.3.1 Variance estimate of 2SPS 50

3.3.2 Variance estimator for the 2SRI approach 53

3.4 Simulations 54

3.5 Result 55

3.6 Discussion 58

3.7 Appendix: The adjusted variance estimate is equal to the heteroskedasticity robust variance estimate for the simple linear case. 68

4 Different Approaches of Instrumental Variable Analysis of Antidiabetic Effect of Bezafibrate **72**

4.1 Introduction 72

4.2 Method 76

4.3 Assumptions and notations 76

4.3.1 Generalized Structural Mean Models 78

4.3.2 Two stage logistic regression 81

4.3.3 Simulations 82

4.3.4 Bezafibrate Data from the GPRD Database 84

4.3.5 IV analysis of the Bezafibrate Data 85

4.4 Results 86

4.4.1 Simulation results 86

4.4.2 IV analysis of bezafibrate data 88

4.5 Discussion 90

5 Conclusion **98**

List of Tables

2.1 Comparison of simulation results and analytic results when there are no always-takers. 36

2.2 Comparison of simulation results and analytic results when there are always-takers. 37

3.1 Comparison of adjusted variance estimates with nave estimats and the variance estimated by bootstrap for the percentage difference from the sample variance of simulation: 2SPS approach with small sample size. 60

3.2 Comparison adjusted variance estimates with nave estimates and the variance estimated by bootstrap for the percentage difference from the sample variance of simulation: 2SPS approach with large sample size. 61

3.3 Comparison of adjusted variance estimates with nave estimates and the variance estimated by bootstrap for the percentage difference from the sample variance of simulation: 2SRI approach with small sample size. 62

3.4 Comparison adjusted variance estimates with nave estimates and the variance estimated by bootstrap for the percentage difference from the sample variance of simulation: 2SRI approach with large sample size. 63

3.5 Comparison of width and coverage of 95% confidence intervals for the true log odds ratio between 2SPS and 2SRI approaches with small sample size. 64

3.6 Comparison of width and coverage of 95% confidence intervals for the true log odds ratio between 2SPS and 2SRI approaches with large sample size. 65

4.1 Simulation results of GSMM estimator without always-takers. 94

4.2 Simulation results of GSMM estimator with always-takers. 95

4.3 Comparing bias, variance and MSE of 2SRI, 2SPS and GMSS. 96

4.4 Correlation of the IV and the exposure. 96

4.5 Rate of outcome associated with exposure and IV. 97

4.6 Comparison of results of causal log OR by different approaches. . . . 97

4.7 Association of IV and Covariate. 97

List of Figures

2.1 Plot of bias on magnitude of confounding δ with 2SPS approach: $\rho_A = 0$, $\rho_C = 0.8$, $\omega_C^1 = 0.6$, $\omega_C^0 = 0.3$. 32

2.2 Plot of bias on magnitude of confounding δ with 2SPS approach: $\rho_A = 0$, $\rho_C = 0.5$, $\omega_C^1 = 0.6$, $\omega_C^0 = 0.3$. 32

2.3 Plot of bias on magnitude of confounding δ with 2SPS approach: $\rho_A = 0$, $\rho_C = 0.5$, $\omega_C^1 = 0.06$, $\omega_C^0 = 0.03$. 33

2.4 Plot of bias on magnitude of confounding δ with 2SPS approach: $\rho_A = 0$, $\rho_C = 0.5$, $\omega_C^1 = 0.006$, $\omega_C^0 = 0.003$. 33

2.5 Plot of bias on magnitude of confounding δ with 2SRI approach: $\rho_A = 0$, $\rho_C = 0.8$, $\omega_C^1 = 0.6$, $\omega_C^0 = 0.3$. 34

2.6 Plot of bias on magnitude of confounding δ with 2SRI approach: $\rho_A = 0$, $\rho_C = 0.5$, $\omega_C^1 = 0.6$, $\omega_C^0 = 0.3$. 34

2.7 Plot of bias on magnitude of confounding δ with 2SRI approach: $\rho_A = 0$, $\rho_C = 0.5$, $\omega_C^1 = 0.06$, $\omega_C^0 = 0.03$. 35

2.8 Plot of bias on magnitude of confounding δ with 2SRI approach: $\rho_A = 0$, $\rho_C = 0.5$, $\omega_C^1 = 0.006$, $\omega_C^0 = 0.003$. 35

3.1 Comparison of variance between 2SPS and 2SRI with sample size N=500. 66

3.2 Comparison of MSE between 2SPS and 2SRI with sample size N=500. 66

3.3 Comparison of variance between 2SPS and 2SRI with Sample Size N=5000. 67

3.4 Comparison of MSE between 2SPS and 2SRI with Sample Size N=5000. 67

Chapter 1

Introduction

In randomize clinical trial, if there are non-compliers, the treatment effect estimated by comparing patients who take the study drug with patients who take placebo or comparison medication is biased, because non-compliance may be associated with the outcome. The bias causing by some factors that are associated with both treatment and outcome is called confounding. Intend-to-treat (ITT) analysis which compares groups defined by the treatment assignment, instead of treatment received, is unbiased, but it can only estimate effect of treatment assignment instead of effect of treatment, even though the effect of treatment assignment is sometimes of biologic and public health interest. In epidemiology research, confounding is bigger issue than in clinical trial because there is no randomization, thus the confounding bias may be caused by many observed or unobserved factors. For many studies, even identifying potential confounding factors is difficult. For this reason, developing a method to control different kinds of confounding is an important task in epidemiology research. Traditional methods for controlling confounding bias include matching

(1; 2), stratification, standardization and multivariate analysis. In the recent years, new methods have been applied to epidemiology research, which include propensity score methods (3; 4; 5), inverse probability of treatment weighting to estimate marginal structure model (6; 7; 8; 9; 10), case-crossover design (11; 12; 13; 14), case time-control design (15), self-controled case series method (16; 17), and instrumental variable analysis. Among these novel methods, instrumental variable (IV) analysis is a potentially important tool for controlling measured and unmeasured confounding in both clinical trials and nonrandomized observational studies. The IV method has been used in econometrics for many years as an important tool to address endogeneity, which means there is a correlation between the parameter or variable and the error term (confounding is one reason for endogeneity). In recent years, this method has been applied to clinical trials and epidemiology research. An IV is a variable that meet the following three criteria: a) it is associated with treatment; b) it has no direct causal effect on the outcome; and c) it is independent of all (unmeasured) confounders of the treatment-outcome relationship. For randomized trials, the IV is a randomized treatment assignment, but for observational studies it needs to be a carefully selected to meet the above assumptions. The IV approach is usually implemented by two-stage regression, which includes two-stage predictor substitution (2SPS) and two-stage residual inclusion (2SRI). Under the 2SPS approach, predicted treatment from the first stage model replaces observed treatment as the principal covariate in the second stage model relating outcome to treatment. Under the 2SRI method, predicted and observed treatment are used to compute a residual that is included as a covariate in the second stage model where the principal covariate is observed

treatment. The IV analysis not only provides a tool for controlling measured and unmeasured confounding, but also has interpretability in causal inference. The causal association of intervention and outcome can be qualitatively analyzed by graphical models (18; 19), and more importantly, can be quantitatively analyzed by the counterfactual or potential-outcome framework originating from Neyman and Fisher in the early 20th century (20). Counterfactual or potential-outcome means what outcome would be for an individual if this individual received different intervention. The difference of this potential outcome for each individual represents the causal effect of the intervention on outcome. In reality, this causal effect is difficult to identify because for each individual, we can only observed one of the two or more potential outcomes, unless there is some kind of study design under the specific assumption. For instance, in the cross-over design of clinical trial, under the assumption that there is no residual treatment effect and patients' condition and response to the treatment doesn't change over time, the causal effect with this potential outcome framework is identifiable. The case crossover design in epidemiology has the same virtue under specific assumptions. For general clinical trials without noncompliance, the patients in study treatment group and comparison group are 'exchangeable', there is no factor associated with treatment. The treatment effect based on treatment received is causal effect under the potential outcome framework, but it is not in the clinical trials with non-compliance. To identify the causal treatment effect in a two-arm randomized trial with non-compliance, Angrist, Imbens and Rubin develop a causal model with the following assumptions: 1) Stable unit treatment value assumption (SUTVA) (71; 105), which means that potential outcomes for each person is unrelated to the

treatment status of other individuals; this assumption also implies the consistency assumption, which means that the potential outcome of a certain treatment will be the same regardless of the treatment assignment mechanism (73); 2) Random assignment assumption, which means that the randomized assignment is unrelated to all confounders in randomized clinical trials, or it is unrelated to the unmeasured confounders (conditional on the measured confounders) of the treatment-outcome relationship in non-randomized studies; 3) Exclusion restriction, which means that any effect of treatment assignment on outcomes must be via an effect of treatment assignment on treatment received; 4) Nonzero average causal effect of treatment assignment on treatment received, which means that the treatment assignment should be associated with treatment received; and 5) Monotonicity, which means that there is no one who would do the opposite of his/her treatment assignment regardless of the actual assignment. They also classify patients into different strata based on the potential treatment receive with different treatment assignment. For patients whose potential treatment received is always the study treatment, they are called always-takers; for patients whose potential treatment received is always consistent with the treatment assignment, they are called compliers; for patients whose potential treatment received is always placebo or non-treatment, they are called never-takers; for patients whose potential treatment received is always opposite to the treatment assignment, they are called defiers. With the monotonicity assumption, there are no defiers. Comparing Angrist-Imbens-Rubin's model assumptions with the IV assumption if we take randomized assignments as IV, we can see that nonzero average causal effect of treatment assignment on treatment received makes the treatment assign-

ment meet the first IV criterion; exclusion restriction makes it meet the second IV criterion and random assignment assumption makes it meet the third criterion. By taking randomized treatment assignment as IV, they analytical proved that under the five assumption mentioned above, the treatment effect estimated with 2SPS linear regression is the average causal effect of receiving treatment among compliers, which is called the local average treatment effect (LATE) or the complier average causal effect (CACE). With linear models, it was proved that other types of estimators based on 2SRI or structural mean models estimated the same treatment effect, thus they can be interpreted similarly (63; 62). Since the outcome of interest in clinical trials and in non-randomized observation studies is often binary, the IV approach has been extended in different ways for inference based on odds ratios from logistic models, where the odds ratio is interpreted as the effect of treatment on outcome in compliers. Terza et al.(21) extended the two-stage IV approach for non-linear models including logistic regression model (two stage predictor substitution (2SPS)), where the predictor of treatment as a function of the instrumental variable replaces observed treatment in the treatment-outcome model. The two stage logistic regression IV approach has been applied to observational studies and compared with other method such as probit structural equation model and a generalized method of moment (GMM) instrumental variable approach(22; 69). Alternatively, Nagelkerke et al.(70) and Terza et al.(23) offered an approach where the treatment-outcome model includes a residual term from the treatment-instrumental variable model (two stage residual inclusion (2SRI)). While the 2SRI procedure is equivalent to the 2SPS approach under the linear model, this is not the case under the logistic model. Terza(23)

showed analytical and simulation-based differences under a true model where it was assumed the true confounders could be observed. Rassen at al.(22) compared the 2 stage logistic regression approach in the context of specific data analysis with other approaches, but they did not do a formal evaluation of these methods theoretically or with simulations, as their comparisons were based on feasibility of implementation. The 2SPS procedure was also used in binary regression measurement error models (24; 25; 26) and an approach similar to 2SRI was used for adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses (27; 28). Using the Bayesian logistic model estimated with Markov-Chain Monte Carlo techniques, Hirano et al. (29) estimated the log odds ratio for treatment on outcome in the compliers. Using an approach similar to the IV procedure, Goetghebeur et al.(65; 30; 96), Vanstellandt et al.(66), Robins and Rotnitzky (68) and Ten Have et al.(67) extended the structural mean model (SMM) for continuous outcomes in different ways to binary outcomes under logistic regression. These semi-parametric approaches estimate the same effect of treatment on outcome as do the IV approaches, but employ different estimation techniques involving estimating equations. Robins and Rotnitzky (68) proposed a multi-stage approach including an estimation step for the prediction of treatment as a function of the IV. Vansteelandt and Goetghebeur (66) offered a two-stage approach where the first stage models the association of outcome on exposure or treatment only among those with the treatment level of the IV and the causal effect is only among those who actually take the treatment. Ten Have et al.(67) presented an iteratively reweighted approach based on estimation under a linear structural mean model. Whereas the approaches by Vansteelandt and Goet-

ghebeur (66), Robins and Rotnitzky (68) are not biased if the association model is specified correctly, the other structural mean model approaches are biased in general at least away from the null association with the bias increasing with the magnitude of confounding. Our research question in the first part of the dissertation is whether the two-stage logistic regression, either of 2SPS or 2SRI, has the interpretation causal log odds ratio of complier average causal effect. In other words, our study objective is to analyze the bias of IV methods of two-stage logistic regression as an estimator of causal log odd ratio. For this purpose, we set up the parameter according to the principal stratification framework and focus on evaluating analytically and empirically with simulations the bias of the 2SPS and 2SRI approaches following the results of Angrist et al. (43) for the additive model. Terza et al. (23) assessed the bias of these two approaches but with respect to a different treatment effect (conditional on the true confounder). There is a need for evaluation in terms of the effect of treatment in compliers, which is the focus of much of the IV literature on the clinical trials (43; 32; 33; 47; 49). To apply the 2SPS and 2SRI to the causal inference of binary outcomes, we not only need to evaluate the performance of the point estimates, but also need to develop methods to correctly estimate variance of the estimates. For the linear two stage methods, the naive variance estimate obtained with ordinary least squares under the second stage regression model of treatment on outcome is not correct, since it does not adjust for the variability of the predicted treatment as a covariate in this model. The variance estimator for the 2SPS IV estimator is based on a heteroskedasticity-robust or sandwich estimator of variance involving cross products of the predicted treatment vector and a scalar dispersion factor based on

the observed treatment factor (34). For the 2SRI approach, we have not found any published research on the variance estimator, but the estimate needs to be adjusted in a similar way, as the 2SRI and 2SPS approaches yield the same estimate of treatment effect for the linear case (35). In the second part of my dissertation, we will derive the sandwich variance estimator for the for the two-stage logistic regression using a similar approach described by Wooldridge (34) to account for the fact that the second stage model include parameters estimates obtained from the first stage. This approach was also used by Zeger and Liang to get asymptotically unbinased variance estimators of generalized estimate equation model (GEE) for longitudinal data analysis (89; 90). We will also do simulations to compare the naive variance estimate and our adjusted estimates with the observed variance of estimates of the treatment effect. We will also compare variance, mean standard error (MSE) and 95% confidence interval coverage of 2SPS and 2SRI by simulation. In the third part of my dissertation, we will compare both 2SPS and 2SRI logistic regression with the generalized structure mean mode (GSMM) proposed by Vansteelandt and Geotghebeur as an extension of the structure mean model (SMM) (96; 68; 97; 104) to the logistic regression for binary outcomes under the randomized clinical trial (RCT) setting when patients assigned to the placebo group can not access the study treatment. In order to apply GSMM to observational study, we first modified the R program so that the GSMM IV method is extended from the RCT setting to observation studies when the placebo group can access the study treatment. Then we can use this program to perform simulations to evaluated performance of SMM IV method under the principal stratification framework. Flory et al did a retrospective cohort study using

the General Practice Research Database (GPRD) to compare instance of diabetes between bezafibrate users and other fibrate users. Their results suggest a significant protective effect of bezafibrate against diabetes (95). As the last part of my dissertation, we will apply different IV approaches to the analysis of the same data to test if there is further evidence of causal effect of bezafibrate against diabetes.

Chapter 2

Bias of Causal Inference for the Odds Ratio Using Two-Stage Instrumental Variable Methods

2.1 Introduction

Instrumental variable (IV) methods are used to estimate effects of receiving treatment or exposure to risk factor on outcome when there is unmeasured confounding in medical research, such as in clinical trials under non-adherence to treatment (40) or observational studies (41; 42). We present closed form expressions of asymptotic bias for the causal odds ratio from two-stage logistic regressions, which is an extension of the conventional IV method for continuous outcomes to a binary outcome.

In the following discussion, we use "treatment" to represent either treatment received or exposure to a risk factor. An IV has the following properties: a) it is

associated with treatment; b) it has no direct causal effect on the outcome; and c) it is independent of all (unmeasured) confounders of the treatment-outcome relationship (41; 43; 45; 46). Note that in randomized trials, the randomized treatment assignment IV is independent of all confounders because it is randomized. In an observational study, the IV could be associated with measured confounders as long as it is independent of all unmeasured confounders of the treatment-outcome relationship conditional on the measured confounders, and the measured confounders are controlled for in the analysis (45). Under these conditions, IV analysis of the treatment-outcome relationship controls for measured and unmeasured confounding (43; 47; 48; 49).

In the context of randomized trials, the IV analysis has been used to adjust for all measured and unmeasured confounding due to treatment non-compliance when estimating the effect of actually receiving treatment. Such confounding factors impact outcome while causing treatment non-compliance or switching from one treatment to another. While intent-to-treat (ITT) inference comparing randomized groups but ignoring treatment non-compliance is protected against such unmeasured confounding, this inference pertains to the effect of prescribing or assigning treatment in the population with the same rate and pattern of non-compliance in the particular trial. Using randomized treatment as an IV, IV inference for the effect of receiving treatment is not dependent on the rate of compliance in the trial except that lower compliance leads to higher variability (50). This IV inference aims to estimate the effect of actually receiving treatment, which is useful for individual patient decisions and for predicting the effect of making the treatment available to populations in which the

rate of compliance might differ from the trial (51; 52).

Besides clinical trials, IV methods are used in observational studies, such as data-based evaluations of the effect of medication on clinical or adverse outcomes. IVs such as physician's prescribing preference (101; 54; 55; 111; 57), clinic or hospital (58),or geographic region (59; 93; 61) have been used to adjust for confounders of the intervention-outcome relationship.

For the additive effect of treatment, Angrist, Imbens and Rubin (43) considered five assumptions for a setting with a proposed IV that are explained in detail in Section 2. Briefly, the key assumptions are that the proposed IV is associated with treatment, is independent of unmeasured confounders given the measured confounders and that the IV only affects outcome through treatment received and there are no defiers. With these assumptions, they used principal stratification (44) to motivate interpretation of the IV estimand. Under the principal stratification framework, the population is divided into sub-classes based on potential treatment receipt that would occur under each level of the instrument variable. In the context of randomized trials with non-compliance, the principal strata are defined as compliers, who adhere to the assignment of treatment but do not take it when not assigned to it; always-takers and never-takers, who respectively always or never take treatment regardless of assignment; and defiers, who only take treatment when not assigned to it. They proved that the probability limit of the two-stage least squares estimator, the usual IV estimator, is the average causal effect of receiving treatment among compliers, which is called the local average treatment effect (LATE) or the complier average causal effect (CACE). Under certain no-interaction assumptions, this effect pertains

to other sub-groups including anyone who takes the treatment or all patients. The estimands for other types of estimators based on structural mean models can be interpreted similarly (62; 63).

For binary outcomes, the IV approach has been extended in different ways for inference based on odds ratios under logistic models, where the odds ratio is interpreted as the effect of treatment on outcome in compliers. Those approaches include the Bayesian logistic model estimated with Markov-Chain Monte Carlo techniques (64), the structural mean model (SMM) (65; 66; 67), and a multi-stage approach including an estimation step for the prediction of treatment as a function of the IV (68).

Terza et al. (39) extended the two-stage IV approach for non-linear models including the logistic regression model (two-stage predictor substitution (2SPS)), where the predictor of treatment as a function of the instrumental variable replaces observed treatment in the treatment-outcome model. This two-stage logistic regression IV approach was applied to observational studies and compared with other IV methods such as the probit structural equation model and a generalized method of moment (GMM) IV approach (69). Alternatively, Nagelkerke et al. (70) and Terza et al. (39) offered an approach where the treatment-outcome model includes a residual term from the treatment-instrumental variable model (two-stage residual inclusion (2SRI)). The 2SRI procedure is equivalent to the 2SPS approach under the linear model, but this is not the case under the logistic model. Terza (39) showed analytical and simulation-based differences under a true model for the causal effect of treatment conditional on the unmeasured confounder.

Give the focus of much of the clinical trials literature on the causal effect of treatment in compliers, there is a need for assessment of the 2SPS and 2SRI two-stage logistic estimators with respect to this causal effect. To achieve this goal, we present analytical and simulation results for the bias of these two estimators under a causal logistic model expressed in terms of potential outcomes under the principal stratification framework, following the results of Angrist et al. (43) for the additive model. We also confirm our analytic result with simulations, and the simulations further reveal patterns of bias for different ranges of confoundings. Our bias evaluation is for a different context from that of Terza et al. (39), who focused on the causal odds ratio in the total population conditional on the unmeasured confounder, whereas we focus on the causal odds ratio among compliers.

2.2 Assumption and Notation

We have the same five assumptions as Angrist, Imbens and Rubin stated in their causal model (43): 1) Stable unit treatment value assumption (SUTVA) (71; 105), which means that potential outcomes for each person is unrelated to the treatment status of other individuals; this assumption also implies the consistency assumption, which means the potential outcome of a certain treatment will be the same regardless of the treatment assignment mechanism (73); 2) Random assignment assumption, which means that the IV is unrelated, as the randomized assignment, to all confounders in the randomized clinical trials, or it is unrelated to the unmeasured confounders (conditional on the measured confounders) of the treatment-outcome re-

lationship in observational studies; 3) Exclusion restriction, which means that any effect of treatment assignment on outcomes must be via an effect of treatment assignment on treatment received; 4) Nonzero average causal effect of treatment assignment on treatment received, which means that the treatment assignment should be associated with treatment received; and 5) Monotonicity, which means that there is no one who would do the opposite of his/her treatment assignment regardless of the actual assignment.

With the above five assumptions, we first define R and Z as the treatment assignment and treatment received variables, respectively. First, R=1 denotes that a patient is assigned to the study treatment, and R=0 means a patient is assigned to the other treatment (or non-treatment), thus R is the IV. Similarly, Z=1 means that a patient receives the study treatment, and Z=0 means that a patient receives the other treatment (or non-treatment). Additionally, $Y^{(1)}$ and $Y^{(0)}$ are the variables for potential outcomes. $Y^{(1)}$ indicates what the outcome for a patient would be if this patient were to take the study treatment, and $Y^{(0)}$ indicates what the outcome for this patient would be if he/she were to take the other treatment (or non-treatment). In contrast, Y is the observed outcome. Similarly, $Z^{(1)}$ and $Z^{(0)}$ are the variables for potential treatment. $Z^{(1)}$ indicates what treatment a patient would take if this patient were assigned to the study treatment, and $Z^{(0)}$ indicates what treatment this patient would take if he/she were assigned to the other treatment (or non-treatment). Based on the principal stratification and potential outcome framework, patients are defined as always-takers (AT) if $Z^{(1)} = 1$ and $Z^{(0)} = 1$; compliers (C) if $Z^{(1)} = 1$ and $Z^{(0)} = 0$; never-takers (NT) if $Z^{(1)} = 0$ and $Z^{(0)} = 0$; and defiers (DF) if $Z^{(1)} = 0$ and

$Z^{(0)} = 1$.

Accordingly, we define the following parameters in the principal stratification framework:

$$\omega_A^1 = \Pr\left(Y^{(1)} = 1|AT\right),$$

$$\omega_C^1 = \Pr\left(Y^{(1)} = 1|C\right),$$

$$\omega_N^1 = \Pr\left(Y^{(1)} = 1|NT\right),$$

$$\omega_A^0 = \Pr\left(Y^{(0)} = 1|AT\right),$$

$$\omega_C^0 = \Pr\left(Y^{(0)} = 1|C\right),$$

$$\omega_N^0 = \Pr\left(Y^{(0)} = 1|NT\right),$$

$$r = \Pr\left(R = 1\right),$$

$$\rho_A = \Pr(AT),$$

$$\rho_C = \Pr(C).$$

With our monotonicity assumption, there are no defiers (43), i.e., $Pr(DF) = 0$. Hence,

$$\Pr(NT) = \rho_N = 1 - \rho_A - \rho_C.$$

The causal log odds ratio for compliers is parameterized as:

$$\begin{aligned}\psi &= logit\left[\Pr\left(Y^{(1)} = 1|C\right)\right] - logit\left[\Pr\left(Y^{(0)} = 1|C\right)\right] \\ &= logit\left(\omega_C^1\right) - logit\left(\omega_C^0\right).\end{aligned}$$

The parameter ψ is the log of the odds ratio that compares the probability of a successful outcome if all compliers received the study treatment compared to if all compliers received the other treatment (or no treatment).

2.3 Bias of Two-Stage Predictor Substitution (2SPS)

In this section, we derive a closed form expression for the probability limit of the two-stage 2SPS logistic regression estimator based on the principal stratification framework and assumptions. We can then obtain closed form expressions for the bias, which is the difference between the expected value of the two-stage regression estimator and the causal log odds ratio.

2.3.1 Probability limit of the estimator

The first stage regression is the treatment received on the treatment assignment R as the IV. Let $D = E(Z|R)$ and $\hat{D}$ be an estimator of D (e.g., maximum likelihood) such that $\hat{D}$ converges in probability to D, $\hat{D} = \hat{E}(Z|R)$. Two-stage logistic regression estimates the causal log odds ratio with the coefficient for $\hat{D}$ in the logistic regression of Y on $\hat{D}$. Let $\hat{\xi}$ be an estimator (e.g., maximum likelihood) of the log odds ratio for D in the logistic regression of Y on D, and let $\hat{\xi}^*$ be the estimator of the log odds ratio for $\hat{D}$ in the logistic regression of Y on $\hat{D}$(i.e., the two-stage 2SPS estimator). As the sample size gets larger, $\hat{D} \longrightarrow D$ and $|\hat{\xi}^* - \hat{\xi}| \xrightarrow{p} 0$ (74; 75), i.e., $\hat{\xi}^*$ converges in probability to ξ under the true model conditional on D, which is $P(Y = 1|D) = expit(\eta + \xi D)$. We now find an expression for ξ as a function of the log odds ratio for treatment received among compliers under the principal stratification framework.

When R=0, only always-takers will receive the treatment; when R=1, both always-takers and compliers will get the treatment. It follows that:

$$d_0 = E(Z|R = 0) = \rho_A \tag{2.3.1}$$

and

$$d_1 = E(Z|R=1) = \rho_A + \rho_C. \tag{2.3.2}$$

Then for the second stage logistic regression we have:

$$\begin{aligned} & logit \Pr(Y=1|R=0) \\ & = logit \Pr(Y=1|D=d_0) \\ & = \eta + \xi d_0, \end{aligned}$$

$$\begin{aligned} & logit \Pr(Y=1|R=1) \\ & = logit \Pr(Y=1|D=d_1) \\ & = \eta + \xi d_1. \end{aligned}$$

Solving the above two equations for ξ, we have:

$$\xi = \frac{logit \Pr(Y|R=1) - logit \Pr(Y|R=0)}{d_1 - d_0}.$$

Under the five assumptions stated in Section 2 and the above parameter settings, the probability of observed Y given R can be expressed as the conditional probability of potential outcome $Y^{(0)}$ and $Y^{(1)}$. We can then calculate $\Pr(Y|R=1)$ and $\Pr(Y|R=0)$ as follows:

$$logit \Pr(Y|R=1) = logit\left(\rho_A\omega_A^1 + \rho_C\omega_C^1 + \omega_N^0 - \rho_A\omega_N^0 - \rho_C\omega_N^0\right),$$

$$logit \Pr(Y|R=0) = logit\left(\rho_A\omega_A^1 + \rho_C\omega_C^0 + \omega_N^0 - \rho_A\omega_N^0 - \rho_C\omega_N^0\right).$$

The full proof of these equations is in Appendix A1. From the above equation, we can calculate ξ as follows:

$$\xi = \frac{logit\Pr(Y|R=1) - logit\Pr(Y|R=0)}{d_1 - d_0} \tag{2.3.3}$$

$$= \frac{\begin{array}{c} logit\left(\rho_A\omega_A^1 + \rho_C\omega_C^1 + \omega_N^0 - \rho_A\omega_N^0 - \rho_C\omega_N^0\right) - \\ logit\left(\rho_A\omega_A^1 + \rho_C\omega_C^0 + \omega_N^0 - \rho_A\omega_N^0 - \rho_C\omega_N^0\right) \end{array}}{\rho_C}.$$

Since $\hat{\xi}$ converges in probability to ξ, equation (2.3.3) is a closed form expression for the probability limit of the two-stage logistic regression estimator of $\hat{\xi}$.

2.3.2 Bias analysis

Having derived the closed form expression of ξ, we can calculate the difference between ψ and ξ, the asymptotic bias of the two-stage logistic regression.

$$\begin{aligned} B_{2SPS} &= \xi - \psi \qquad (2.3.4)\\ &= \frac{1}{\rho_C}\begin{pmatrix} logit\left(\rho_A\omega_A^0 + \rho_C\omega_C^1 + \omega_N^0 - \rho_A\omega_N^0 - \rho_C\omega_N^0\right) \\ -logit\left(\rho_A\omega_A^0 + \rho_C\omega_C^0 + \omega_N^0 - \rho_A\omega_N^0 - \rho_C\omega_N^0\right) \end{pmatrix} \\ &- \left(logit\left(\omega_C^1\right) - logit\left(\omega_C^0\right)\right) \\ &= \frac{1}{\rho_C}\begin{pmatrix} logit(\rho_A\omega_A^0 + \rho_C\omega_C^1 + expit\left(logit\left(\omega_C^0\right) + \delta\right)\rho_N) \\ -logit(\rho_A\omega_A^0 + \rho_C\omega_C^0 + expit\left(logit\left(\omega_C^0\right) + \delta\right)\rho_N) \end{pmatrix} \\ &- \left(logit\left(\omega_C^1\right) - logit\left(\omega_C^0\right)\right). \end{aligned}$$

In the above equation, we re-parameterize the ω_N^0 and introduce a new parameter δ as follow,

$$logit\left(\omega_N^0\right) = logit\left(\omega_C^0\right) + \delta,$$

then

$$\omega_N^0 = expit\left(logit\left(\omega_C^0\right) + \delta\right) = \omega_C^0 \frac{e^\delta}{\omega_C^0 e^\delta - \omega_C^0 + 1}.$$

The parameter δ is the difference between ω_N^0 and ω_C^0 on the logit scale, so it is the log odds ratio of never-takers over compliers regarding the outcome. Given differences between principal strata are due to unmeasured confounders related to outcome, δ in equation (2.3.4) can be interpreted as the magnitude of confounding, where $\delta = 0$ implies no confounding because ω_N^0=ω_C^0.

From the equation (2.3.4), we can easily see:

a) When $\rho_C = 1$ (every one is a complier), $B_{2SPS} = 0$. This is because when $\rho_C = 1$, both ρ_A and ρ_N are 0. In equation (2.3.4), if we replace ρ_C by 1 and both ρ_A and ρ_N by 0, we have $B_{2SPS} = 0$.

b) When $\omega_C^1 = \omega_C^0$ (there is no causal effect), $B_{2SPS} = 0$. If we replace ω_C^1 by ω_C^0 in equation (2.3.4), all terms are canceled out and we have $B_{2SPS} = 0$.

c) The bias function does not include R, thus bias is not related to $\Pr(R = 1)$.

d) Bias can exist even when there is no confounding, that is, when $\rho_A = 0$ and $\omega_C^0 = \omega_N^0$. Replacing ρ_A by 0 in equation (2.3.4), we have

$$\begin{aligned} B_{2SPS} &= \frac{\begin{array}{c} logit\left(\rho_A\omega_A^1 + \rho_C\omega_C^1 + \omega_N^0 - \rho_A\omega_N^0 - \rho_C\omega_N^0\right) \\ -logit\left(\rho_A\omega_A^1 + \rho_C\omega_C^0 + \omega_N^0 - \rho_A\omega_N^0 - \rho_C\omega_N^0\right) \end{array}}{\rho_C} - logit\left(\omega_C^1\right) + logit\left(\omega_C^0\right) \\ &= \frac{logit\left(\rho_C\omega_C^1 + \omega_N^0 - \rho_C\omega_N^0\right) - logit\left(\omega_N^0\right)}{\rho_C} - logit\left(\omega_C^1\right) + logit\left(\omega_C^0\right). \end{aligned}$$

In this equation, B_{2SPS} is generally not 0, because ρ_C in the denominator can not be canceled out with the ρ_C in the logit function of the numerator.

With the closed form expression (2.3.4), we can analyze the magnitude of bias under different parameter settings according to specific scenarios. To simplify the analysis and show the relationship between bias and confounding, we create four such scenarios when there are no always-takers. We plot bias against δ while fixing all other parameters in Fig. 2.1 to 2.4.

All four plots show that the bias is not 0 when there is no confounding ($\delta = 0$). When the compliance rate decreases from 0.8 to 0.5, the bias on the logit scale is about 5 time larger (compare plot 2.1 and plot 2.2). Comparing plot 2.2 and plot 2.3, we can see that when the event rate is lower, the bias range is larger, but when the event rate is decreased from 0.03 to 0.003 (Fig. 2.4), the absolute bias does not increase further.

2.4 Bias of Two-Stage Residual Inclusion (2SRI)

In this section, we extend to the 2SRI estimator, the derivation in Section 3 of bias of the 2SPS under the principal stratification framework. In the first stage regression of treatment received on the treatment assignment R as an IV, the residual is $E = Z - E(Z|R)$, and the second stage regression model is

$$Pr(Y = 1) = expit\left(\lambda_0 + \lambda_1 Z + \lambda_2 E\right). \tag{2.4.1}$$

The estimator of λ_1 is an estimate of the causal log odds ratio for receiving treatment among compliers. We derive a closed form expression for the probability limit of the estimator of λ_1. This enables us to derive a closed form expression for the asymptotic

difference between the probability limit of the estimator of λ_1 and the causal log odds ratio among compliers.

2.4.1 Closed form expression for the probability limit of the estimator

For the 2SRI approach, in general, equation (2.4.1) is not the true model for $Pr(Y = 1|Z, E)$, as the true model includes the interaction term between Z and E; this makes it much more difficult to develop a closed form expression for the probability limit of the estimator. However, if we assume that there are no always-takers, so that $\Pr(Z = 1, R = 0) = 0$, then the true model does not have the interaction term and the 2SRI model in equation (2.4.1) is the true model (see the details in Appendix A2). In this section, we develop a closed form expression for the probability limit of the estimator of λ_1 only under the no always-taker assumption. The no always-taker assumption is true in clinical trials when patients in the placebo group cannot access the study drug. In contrast, the bias results for the 2SPS estimator depend on a true model conditional on just Z (treatment-received) that does not require the absence of always-takers.

The residual $E = Z - E(Z|R)$ is estimated from the first stage regression, and is included as a covariate in the second stage regression. Letting $\hat{E} = Z - \hat{E}(Z|R)$, we consider the second stage regression $Pr(Y = 1|Z, \hat{E}) = expit(\lambda_0 + \lambda_1 Z + \lambda_2 \hat{E})$. The 2SRI approach estimates the causal log odds ratio with the estimated coefficient for Z in the logistic regression of Y on Z and $\hat{E}$. Let $\hat{\lambda}_1$ denote the estimated coefficient

for Z in the logistic regression of Y on Z and E, and let $\hat{\lambda}_1^*$ denote the estimated coefficient for Z in the logistic regression of Y on Z and $\hat{E}$. As the sample size gets larger, $\hat{E} \longrightarrow E$ and $|\hat{\lambda}_1^* - \hat{\lambda}_1| \xrightarrow{p} 0$ (74; 75). The estimator $\hat{\lambda}_1^*$ converges in probability to λ_1 under the model $Pr(Y = 1|Z, E) = expit(\lambda_0 + \lambda_1 Z + \lambda_2 E)$ when there are no always-takers. When there are always-takers, the 2SRI model is misspecified. In this situation, $\hat{\lambda}_1^*$ estimated from the second stage logistic regression converges to the point that minimizes the Kullback-Leibler distance between the family of probability distributions being maximized over the true probability distribution (76).

Under the no always-taker assumption, we can find an expression for λ_1 as follows. From the equations (2.3.1) and (2.3.2), we have

$$E(Z|R) = \rho_A + \rho_C R,$$

so

$$E = Z - E(Z|R) = Z - \rho_A - \rho_C R.$$

Note that Z, E and Z, R contain the same information; i.e., knowing Z, E tells us Z, R and vice versa, so that $Pr(Y = 1|Z, E) = Pr(Y = 1|Z, R)$. For the second stage regression, we have

$$\begin{aligned} & logit \Pr(Y = 1|Z, E) \qquad (2.4.2) \\ &= \lambda_0 + \lambda_1 Z + \lambda_2 E \\ &= \lambda_0 + \lambda_1 Z + \lambda_2 (Z - \rho_A - \rho_C R) \\ &= \lambda_0 - \lambda_2 \rho_A + (\lambda_1 + \lambda_2) Z - \lambda_2 \rho_C R \\ &= logit \Pr(Y = 1|Z, R). \end{aligned}$$

Then we have three equations based on the possible values of Z and R ((Z=1,R=0) is not possible because there are no always-takers):

$$
\begin{aligned}
&logit \Pr\left(Y=1|Z=1,R=1\right) \qquad (2.4.3)\\
&= logit \Pr\left(Y^{(1)}=1|Z=1,R=1\right)\\
&= logit\left(\frac{\rho_A}{\rho_A+\rho_C}\omega_A^1+\frac{\rho_C}{\rho_A+\rho_C}\omega_C^1\right)\\
&= \lambda_0-\lambda_2\rho_A+(\lambda_1+\lambda_2)-\lambda_2\rho_C,
\end{aligned}
$$

$$
\begin{aligned}
&logit \Pr\left(Y=1|Z=0,R=1\right) \qquad (2.4.4)\\
&= logit \Pr\left(Y^{(0)}=1|Z=0,R=1\right)\\
&= logit \Pr(Y^{(0)}=1|NT)\\
&= logit(\omega_N^0)\\
&= \lambda_0-\lambda_2\rho_A-\lambda_2\rho_C,
\end{aligned}
$$

$$
\begin{aligned}
&logit \Pr\left(Y=1|Z=0,R=0\right) \qquad (2.4.5)\\
&= logit \Pr\left(Y^{(0)}=1|Z=0,R=0\right)\\
&= logit\left(\frac{1-\rho_A-\rho_C}{1-\rho_A}\omega_N^0+\frac{\rho_C}{1-\rho_A}\omega_C^0\right)\\
&= \lambda_0-\lambda_2\rho_A.
\end{aligned}
$$

Solving equations (2.4.3), (2.4.4) and (2.4.5) for λ_1 yields the closed form expression for λ_1 as:

$$
\begin{aligned}
\lambda_1 = &logit\left(\frac{\rho_A}{\rho_A+\rho_C}\omega_A^0+\frac{\rho_C}{\rho_A+\rho_C}\omega_C^1\right)-logit(\omega_N^0) \qquad (2.4.6)\\
&-\frac{1}{\rho_C}logit\left(\frac{1-\rho_A-\rho_C}{1-\rho_A}\omega_N^0+\frac{\rho_C}{1-\rho_A}\omega_C^0\right)+\frac{1}{\rho_C}logit(\omega_N^0).
\end{aligned}
$$

2.4.2 Bias analysis

With the closed form expression for the probability limit of $\hat{\lambda}_1$, we can calculate B_{2SRI}, the bias defined as the difference between the log odds ratio for treatment-received among compliers and the estimated log odds ratio with the 2SRI approach.

$$
\begin{aligned}
B_{2SRI} &= \lambda_1 - \psi \qquad (2.4.7)\\
&= logit\left(\frac{\rho_A}{\rho_A+\rho_C}\omega_A^1 + \frac{\rho_C}{\rho_A+\rho_C}\omega_C^1\right) - logit(\omega_N^0)\\
&- \frac{1}{\rho_C} logit\left(\frac{1-\rho_A-\rho_C}{1-\rho_A}\omega_N^0 + \frac{\rho_C}{1-\rho_A}\omega_C^0\right) + \frac{1}{\rho_C} logit(\omega_N^0)\\
&- logit\left(\omega_C^1\right) + logit\left(\omega_C^0\right)\\
&= logit\left(\frac{\rho_A}{\rho_A+\rho_C}\omega_A^1 + \frac{\rho_C}{\rho_A+\rho_C}\omega_C^1\right) - logit\left(expit\left(logit\left(\omega_C^0\right)+\delta\right)\right)\\
&- \frac{1}{\rho_C} logit\left(\frac{1-\rho_A-\rho_C}{1-\rho_A}\left(expit\left(logit\left(\omega_C^0\right)+\delta\right)\right) + \frac{\rho_C}{1-\rho_A}\omega_C^0\right)\\
&+ \frac{1}{\rho_C} logit\left(expit\left(logit\left(\omega_C^0\right)+\delta\right)\right) - logit\left(\omega_C^1\right) + logit\left(\omega_C^0\right).
\end{aligned}
$$

δ is the same parameter as in equation (2.3.4). The following conclusions follow from equation (2.4.7):

a) When $\rho_C = 1$ (every one is a complier), $B_{2SRI} = 0$. If $\rho_C = 1$, both ρ_A and ρ_N equal to 0. Plug in these values of ρ_C,ρ_A and ρ_N to the equation (2.4.7), $B_{2SRI} = 0$.

b) When $\omega_C^0 = \omega_N^0$, and $\omega_A^1 = \omega_C^1$ (there is no confounding), we replace ω_N^0 with ω_C^0, and ω_A^1 with ω_C^1 in equation (2.4.7), yielding $B_{2SRI} = 0$. That is, when there is no confounding, the 2SRI approach is unbiased.

As in section 3 with the 2SPS estimator, we use equation (2.4.7) to analyze the magnitude of bias of the 2SRI estimator under different scenarios as in Fig. 2.4-2.8.

All four plots (Fig 2.5 to 2.8) show that when there is no confounding ($\delta = 0$),

the bias of the 2SRI estimator is zero. The first scenario (Fig. 2.5) shows that when the compliance rate is high (0.8), the bias is small for a wide range of confounding. The second scenario (Fig. 2.6) shows that if the outcome is not rare, the bias is very small unless δ is smaller than -1 or greater than 2, which means that the odds ratio comparing compliers to never-takers with respect to the potential outcomes is smaller than 0.37 or greater than 7.4. These scenarios correspond to very strong confounding. Fig. 2.7 shows the scenario when the outcome is rare, with ω_C^1 and ω_C^0 one tenth of those in scenario 1, The bias for this scenario is larger than that of scenario 1, but the bias is still moderate if the confounding is not very severe. In scenario 4 (Fig. 2.8), we make the outcome even rarer. The magnitude of bias does not change much compared to the bias under scenario 3. Therefore, we can conclude that for the 2SRI model, there is bias when there is confounding, but the bias is small to moderate if the confounding is not severe.

2.5 Simulation

2.5.1 Simulation algorithm

We simulated the data sets according to the following algorithm:

Step 1: Generate a data set with total number of N subjects. Among these subjects, always-takers (ATs), compliers (Cs), and never-takers (NTs) are generated from a multinomial distribution with probability of ρ_A for ATs, probability of ρ_C for Cs and probability of ρ_N for NTs. With the statistical programming package R, this

step can be implemented by W=t(rmultinom(n, 1, c(ρ_A,ρ_C, ρ_N))).

Step 2: With the probability of $Pr(R = 1) = r$, randomly assign about rN of the subjects to R=1 and the rest of $(1 - r)N$ subject to $R = 0$. This step can be implemented by R=t(rmultinom(n, 1, c(r,1-r))) in the package R.

Step 3: Simulate $Y^{(0)}$ and $Y^{(1)}$ based on the value of AT, C or NT, and the parameter ω_A^1, ω_C^1, ω_N^1, ω_A^0, ω_C^0, and ω_N^0. For instance, if an subject is AT, then $Pr(Y^{(0)} = 1) = \omega_A^0$, and $Pr(Y^{(1)} = 1) = \omega_A^1$. With these probabilities, we can create $Y^{(1)}$ and $Y^{(0)}$ with the binomial distribution. We implemented this step in the package R with the following program:

prY0=W[,1]*ω_A^0+W[,2]*ω_C^0+W[,3]*ω_N^0

dim(prY0)=c(n,1)

prY1=W[,1]*ω_A^1+W[,2]*ω_C^1+W[,3]*ω_N^1

dim(prY1)=c(n,1)

Y0=apply(prY0, 1, function (x) rbinom(1,1,x))

Y1=apply(prY1, 1, function (x) rbinom(1,1,x))

Step 4: Based on AT, C or NT, and R, determine Z. For instance, if an observation is in either the AT or C group, and the treatment assignment R=1, then Z=1.

Step 5: Based on Z , $Y^{(0)}$ and $Y^{(1)}$, determine Y

$Y = Y^{(1)}Z + Y^{(0)}(1 - Z)$.

2.5.2 Simulation results

For each setting, we ran the simulation 2000 times, with the sample size of n=10,000. For both 2SPS and 2SRI approaches, we simulated data with different selection of parameters. As examples, Table 2.1 shows the results with the parameter settings without always-takers: $\rho_A = 0$; $\rho_C = 0.5$ (thus $\rho_N = 0.5$); $\omega_C^0 = 0.3$ or $\omega_C^0 = 0.03$; $\omega_C^1 = 0.6$ or $\omega_C^1 = 0.06$; δ varies among 2, 1.5, 1, 0.5, 0, -0.5, -1, -1.5 or -2. For these simulations, the bias is calculated as the difference between the mean of estimated log odds ratio ($\hat{\xi}$ for 2SPS and $\hat{\lambda}_1$ for 2SRI) and the log odds ratio among compliers ψ. The mean square of error (MSE) is calculated as the mean square of the difference between the estimated log odds ratio and the log odds ratio among compliers.

Under all parameter settings without always-takers, the bias resulting from simulations is consistent with the analytic results, and when there is no confounding, the bias is not zero for 2SPS but is zero for 2SRI (Table 2.1). The simulation results of MSE follow the same pattern as the results for absolute bias with these large sample simulations. We are currently doing further research on the MSE properties of the different estimators.

We also performed simulations including always-takers with the parameter settings: $\rho_A = 0.2$; $\rho_C = 0.5$ (thus $\rho_N = 0.3$); $\omega_C^0 = 0.3$ or $\omega_C^0 = 0.03$; $\omega_C^1 = 0.6$ or $\omega_C^1 = 0.06$; δ varies among 2, 1.5, 1, 0.5, 0, -0.5, -1, -1.5 or -2. Under these parameter settings, the analytic results are available for the 2SPS procedure, but are not possible for the 2SRI approach as discussed in Section 4. As shown in table 2.2, the bias from

simulated data is consistent with the analytic results for the 2SPS approach when there are always-takers. For 2SRI, the results show that the bias is smaller than for 2SPS, and is close to 0 when δ is 0, but for some parameter settings with strong confounding, the bias is larger than for 2SPS.

2.6 Discussion

The instrumental variable approach has been applied to logistic regression to control for unmeasured confounding in estimating treatment effects under non-adherence in randomized trials and under actual medical care in observational studies. However, there has been little if no evaluation of the bias of this use of instrumental variables in the context of estimating the effect of treatment among those who are compliers or take the treatment. Accordingly, we have developed closed form expressions for the asymptotic bias of the 2SRI and 2SPS approaches to two-stage logistic regression, and we have shown that these analytic results are consistent with the simulation results under different parameter settings. Terza et al. (39) showed that the 2SRI approach is unbiased when the true model is conditional on the unmeasured confounder. For the treatment effect conditional on compliance or receiving treatment, Nagelkerke et al. (70) and Ten Have et al. (67) presented simulations showing that the bias of 2SRI approach increases as the magnitude of confounding increases. Our analytical and simulation results confirm such bias for the 2SRI as well as for the 2SPS approach. We further show that unlike the 2SRI approach, the 2SPS procedure is biased even when there is no unmeasured confounding.

An important contribution of this research is the expression of the conditional distribution of observed outcomes Y given treatment assignment R as a function of the probability of compliance and the conditional distribution of potential outcomes $Y^{(0)}$ and $Y^{(0)}$, given compliance status. With this contribution, we can analytically present probability limits and therefore the bias of the estimators of the causal effects of treatment given compliance and treatment status. Further, we provide analytic estimates of bias for a variety of situations. These analytic estimates of bias can help researchers evaluate if the bias is small under specific conditions (e.g. high compliance, and moderate confounding). Hence, our results can be used as a guide for deciding if the 2SRI or 2SPS strategy is appropriate. This method can be potentially applied to the bias analysis of causal inference with other non-linear two-stage regressions, such as regressions of probit models and log linear models.

When the 2SRI or 2SPS is appropriately used, these approaches have the advantage that they are very easy to implement with any software package that can do logistic regression (e.g., SAS, R, or STATA). Logistic regression is used for both the first and second stages of either the 2SRI or 2SPS procedures. The predicted or residual values from the first stage logistic regression of treatment on the IV are used as covariates in the second stage logistic regression: the predicted value of treatment replaces observed treatment for 2SPS, whereas the residual from the first stage regression is added as a covariate along with observed treatment for 2SRI.

The bias for both the 2SPS and 2SRI approaches occurs when all of the IV assumptions are met. Additional research is needed in resolving such bias, and also in assessing departures from the IV assumptions under the logistic IV model. To resolve

the bias of the 2SRI and 2SPS approaches, the logistic structural nested mean model of Vansteelandt and Goetghebeur (77) in the randomized trial context when controls do not have access to the treatment can be extended to the observational data context when all subjects have access to treatment. Additionally, such a modeling approach may be modified to assess departures from the exclusion restriction using a similar weighted estimating equations approach as in Ten Have et al. (2007) (78).

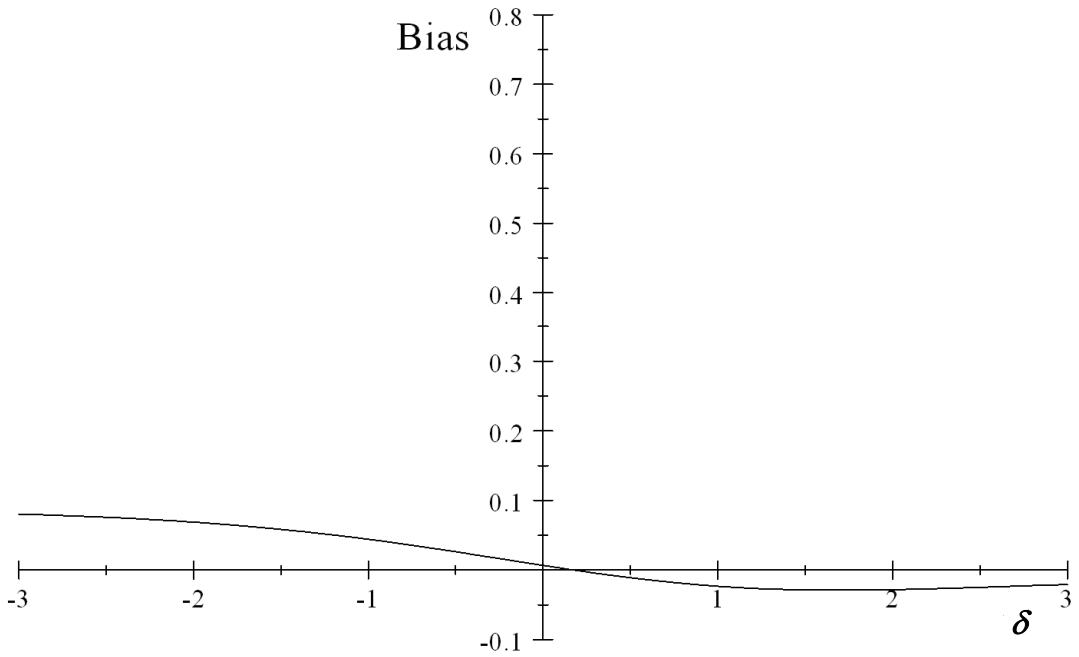

Fig 2.1. Plot of bias on magnitude of confounding δ with 2SPS approach: ρ_A=0, ρ_C =0.8, $\omega^1{}_c$=0.6, $\omega^0{}_c$ =0.3.

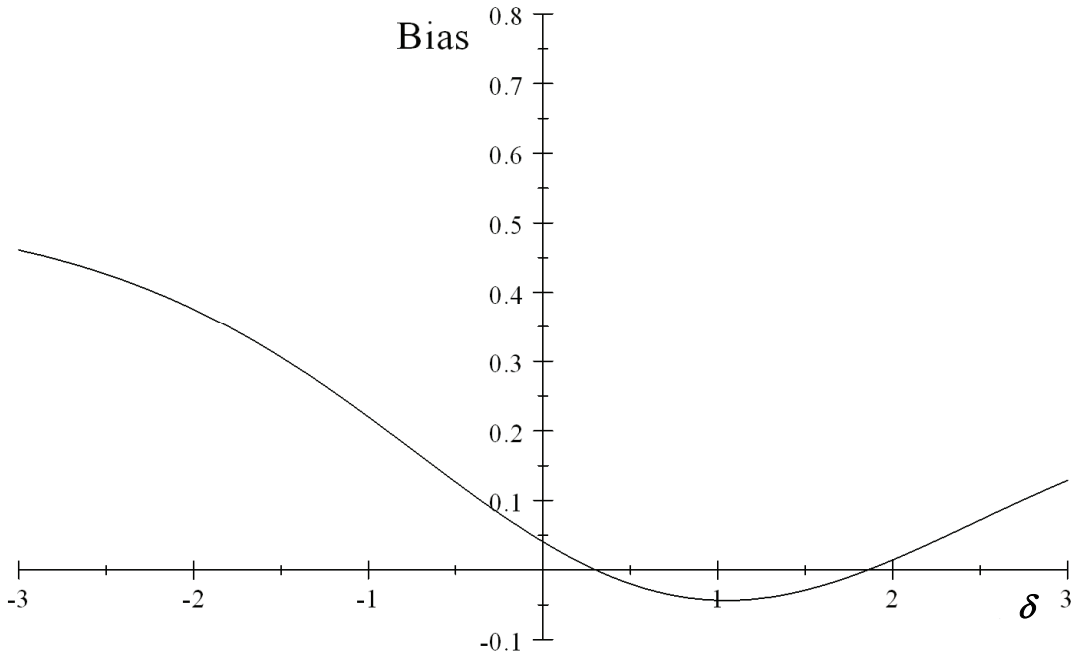

Fig 2.2. Plot of bias on magnitude of confounding δ with 2SPS approach: ρ_A=0, ρ_C =0.5, $\omega^1{}_c$ =0.6, $\omega^0{}_c$ =0.3.

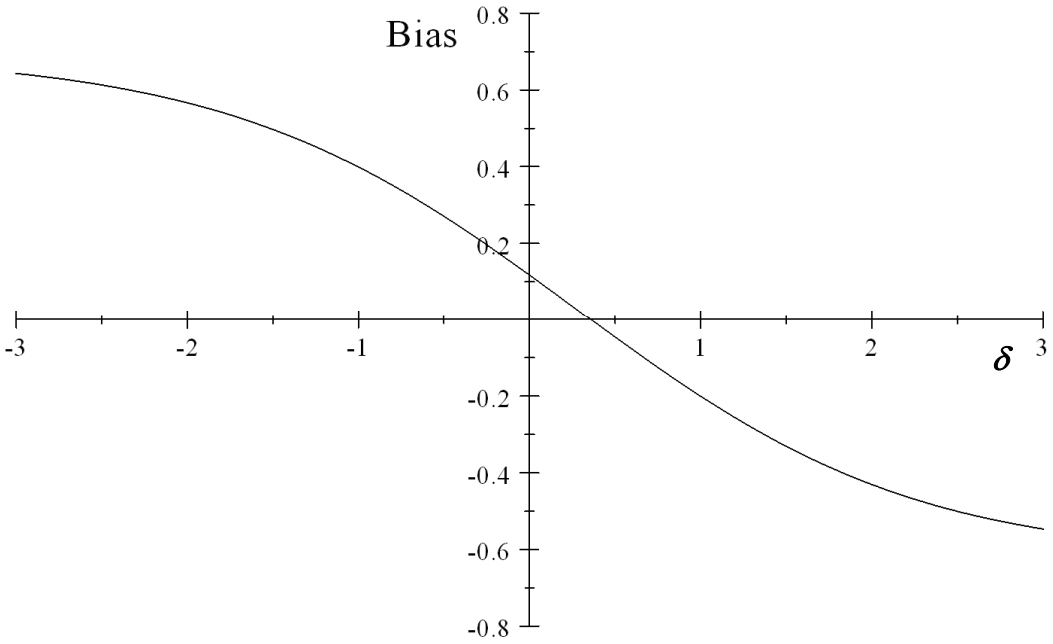

Fig 2.3. Plot of bias on magnitude of confounding δ with 2SPS approach: ω_1=0, ω_2=0.5, ω_{12}=0.06, ω_{02}=0.03.

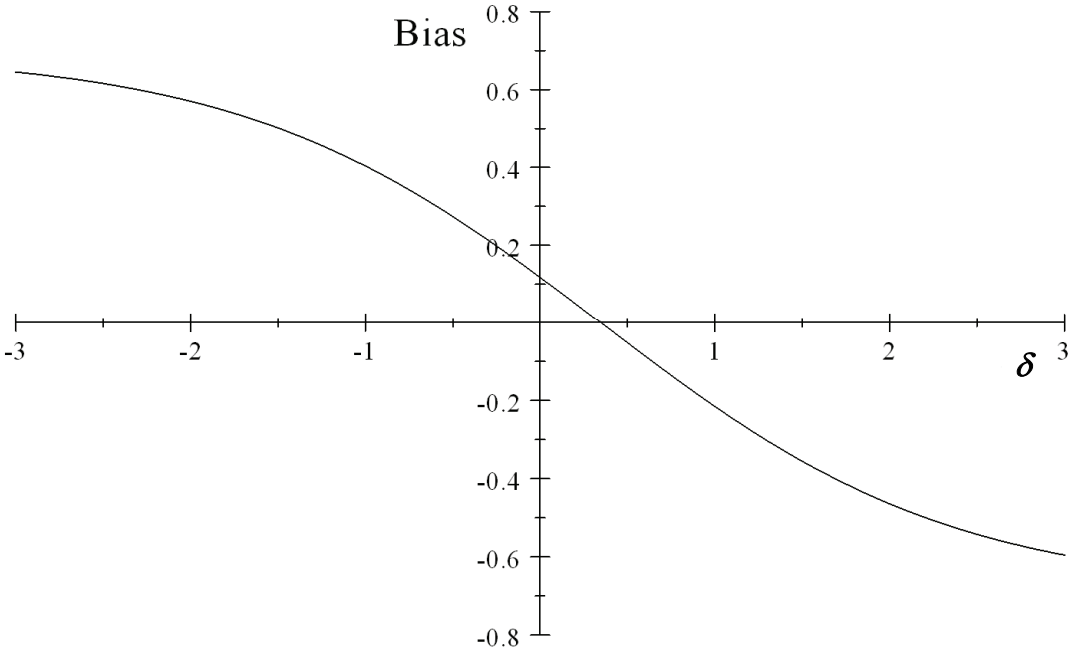

Fig 2.4. Plot of bias on magnitude of confounding δ with 2SPS approach: ρ_A=0, ρ_C=0.5, ω^1_c=0.006, ω^0_c=0.003.

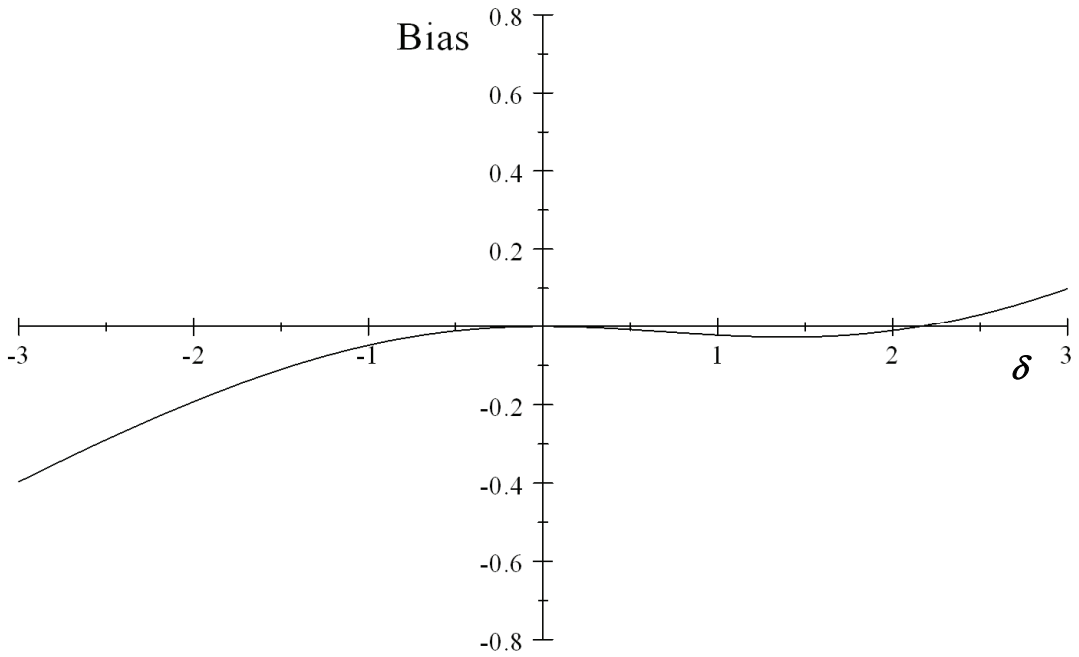

Fig 2.5. Plot of bias on magnitude of confounding δ with 2SRI approach: ρ_A=0, ρ_C =0.8, $\omega^1{}_c$=0.6, $\omega^0{}_c$ =0.3.

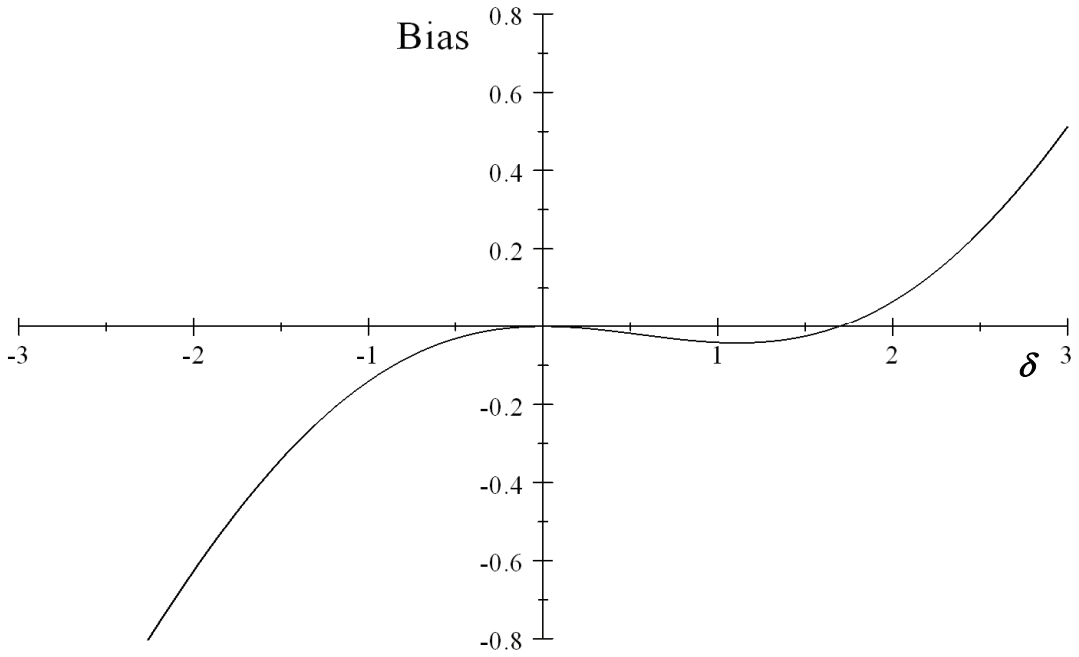

Fig 2.6. Plot of bias on magnitude of confounding δ with 2SRI approach: ρ_A=0, ρ_C =0.5, $\omega^1{}_c$=0.6, $\omega^0{}_c$ =0.3.

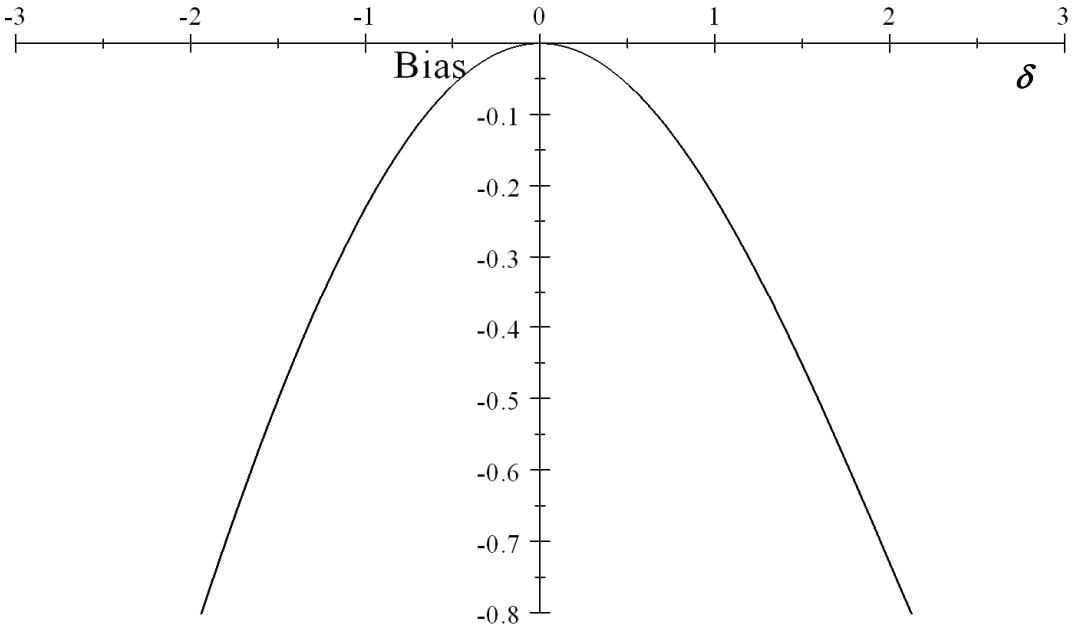

Fig 2.7. Plot of bias on magnitude of confounding δ with 2SRI approach: ρ_A=0, ρ_C =0.5, $\omega^1{}_c$=0.06, $\omega^0{}_c$ =0.03.

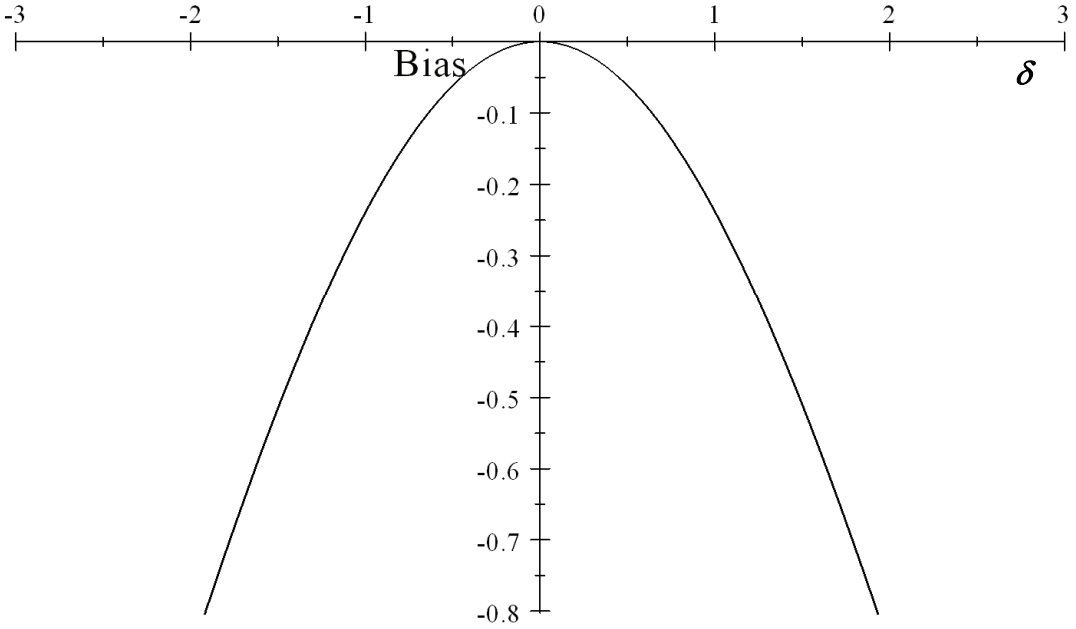

Fig 2.8. Plot of bias on magnitude of confounding δ with 2SRI approach: ρ_A=0, ρ_C =0.5, $\omega^1{}_c$=0.006, $\omega^0{}_c$ =0.003.

Table 2.1. Comparison of simulation results and analytic results when there are no always-takers.

				2SPS				2SRI			
ω^0_C	ω^1_C	True LogOR	δ	LogOR by Regression	Observed Bias	Analytic Result of Bias	MSE	LogOR by Regression	Observed Bias	Analytic Result of Bias	MSE
0.3	0.60	1.2528	-2.0	1.6295	0.3768	0.3754	0.1500	0.6256	-0.6272	-0.6266	0.4095
			-1.5	1.5601	0.3073	0.3061	0.1024	0.9112	-0.3416	-0.3415	0.1295
			-1.0	1.4740	0.2213	0.2200	0.0567	1.1127	-0.1400	-0.1410	0.0301
			-0.5	1.3813	0.1286	0.1263	0.0238	1.2244	-0.0284	-0.0309	0.0095
			0.0	1.2961	0.0433	0.0405	0.0088	1.2559	0.0031	0.0000	0.0075
			0.5	1.2362	-0.0166	-0.0200	0.0069	1.2383	-0.0145	-0.0179	0.0071
			1.0	1.2079	-0.0449	-0.0435	0.0090	1.2103	-0.0425	-0.0413	0.0088
			1.5	1.2228	-0.0300	-0.0289	0.0081	1.2268	-0.0259	-0.0250	0.0079
			2.0	1.2666	0.0138	0.0145	0.0080	1.3172	0.0644	0.0651	0.0123
0.03	0.0600	0.7246	-2.0	1.2894	0.5648	0.5666	0.3901	-0.1732	-0.8978	-0.8474	0.9745
			-1.5	1.2215	0.4969	0.4973	0.3131	0.2011	-0.5235	-0.5015	0.3865
			-1.0	1.1225	0.3980	0.3994	0.2181	0.4788	-0.2458	-0.2314	0.1432
			-0.5	0.9900	0.2654	0.2709	0.1232	0.6522	-0.0724	-0.0589	0.0666
			0.0	0.8374	0.1128	0.1175	0.0585	0.7161	-0.0084	0.0000	0.0485
			0.5	0.6770	-0.0475	-0.0459	0.0387	0.6630	-0.0616	-0.0571	0.0406
			1.0	0.5198	-0.2048	-0.2005	0.0705	0.5002	-0.2243	-0.2169	0.0790
			1.5	0.3911	-0.3334	-0.3310	0.1335	0.2658	-0.4587	-0.4525	0.2339
			2.0	0.2932	-0.4314	-0.4306	0.2026	-0.0107	-0.7352	-0.7297	0.5593

Note: The probability of always-takers ρ_A=0, the probability of compliers ρ_C=0.5 and the probability of never-takers ρ_N=0.5.

Table 2.2. Comparison of simulation results and analytic results when there are always-takers.

				2SPS				2SRI			
ω^0_C	ω^1_C	True LogOR	δ	LogOR by Regression	Observed Bias	Analytic Result of Bias	MSE	LogOR by Regression	Observed Bias	Analytic Result of Bias	MSE
0.3	0.60	1.2528	-2.0	1.3159	0.0631	0.0615	0.0098	1.2554	0.0026	NA	0.0090
			-1.5	1.3007	0.0480	0.0461	0.0081	1.2624	0.0096	NA	0.0085
			-1.0	1.2809	0.0281	0.0257	0.0065	1.2677	0.0149	NA	0.0079
			-0.5	1.2574	0.0046	0.0016	0.0057	1.2668	0.0140	NA	0.0074
			0.0	1.2338	-0.0190	-0.0220	0.0061	1.2559	0.0031	NA	0.0066
			0.5	1.2167	-0.0361	-0.0389	0.0073	1.2380	-0.0148	NA	0.0067
			1.0	1.2112	-0.0416	-0.0434	0.0083	1.2221	-0.0306	NA	0.0077
			1.5	1.2201	-0.0327	-0.0346	0.0077	1.2216	-0.0311	NA	0.0076
			2.0	1.2393	-0.0135	-0.0162	0.0071	1.2410	-0.0118	NA	0.0071
0.03	0.0600	0.7246	-2.0	0.8826	0.1580	0.1583	0.0753	0.9577	0.2331	NA	0.1092
			-1.5	0.8623	0.1378	0.1390	0.0677	0.9177	0.1931	NA	0.0895
			-1.0	0.8312	0.1067	0.1093	0.0578	0.8633	0.1387	NA	0.0677
			-0.5	0.7880	0.0634	0.0652	0.0483	0.7983	0.0737	NA	0.0507
			0.0	0.7276	0.0030	0.0034	0.0410	0.7250	0.0005	NA	0.0413
			0.5	0.6471	-0.0774	-0.0766	0.0421	0.6443	-0.0803	NA	0.0427
			1.0	0.5549	-0.1696	-0.1704	0.0598	0.5541	-0.1705	NA	0.0600
			1.5	0.4575	-0.2671	-0.2683	0.0971	0.4389	-0.2857	NA	0.1073
			2.0	0.3686	-0.3560	-0.3586	0.1472	0.2962	-0.4284	NA	0.2042

Note: The probability of always-takers ρ_A=0.2, the probability of compliers ρ_C=0.5 and the probability of never-takers ρ_N=0.3.

2.7 Appendix

A1. Prove that the probability of observed Y given R can be expressed by the following equations.

$$\Pr\left(Y|R=1\right)=\rho_A\omega_A^0+\rho_C\omega_C^1+\omega_N^0-\rho_A\omega_N^0-\rho_C\omega_N^0,$$

and

$$\Pr\left(Y|R=0\right)=\rho_A\omega_A^0+\rho_C\omega_C^0+\omega_N^0-\rho_A\omega_N^0-\rho_C\omega_N^0.$$

In these equations, AT means always-taker, C means complier, and NT means never-taker, and

$$\omega_A^1=\Pr\left(Y^{(1)}=1|AT\right),$$

$$\omega_C^1=\Pr\left(Y^{(1)}=1|C\right),$$

$$\omega_N^1=\Pr\left(Y^{(1)}=1|NT\right),$$

$$\omega_A^0=\Pr\left(Y^{(0)}=1|AT\right),$$

$$\omega_C^0=\Pr\left(Y^{(0)}=1|C\right),$$

$$\omega_N^0=\Pr\left(Y^{(0)}=1|NT\right),$$

$$r=\Pr\left(R=1\right),$$

$$\rho_A=\Pr(AT),$$

$$\rho_C=\Pr(C),$$

$$\rho_N=\Pr(NT).$$

Proof:

$$
\begin{aligned}
&\Pr\left(Y^{(1)}=1|Z=1,R=1\right)\\
&=\Pr\left(Y^{(1)}=1,Z=1,R=1\right)/\Pr(R=1,Z=1)\\
&=\frac{\Pr(Y^{(1)}=1,AT,R=1)+\Pr(Y^{(1)}=1,C,R=1)}{\Pr(R=1,AT)+\Pr(R=1,C)}\\
&=\frac{\Pr(Y^{(1)}=1,AT)\Pr(R=1)+\Pr(Y^{(1)}=1,C)\Pr(R=1)}{\Pr(R=1)\Pr(AT)+\Pr(R=1)\Pr(C)}\\
&=\frac{\Pr(Y^{(1)}=1|AT)\Pr(AT)+\Pr(Y^{(1)}=1|C)\Pr(C)}{\Pr(R=1)\Pr(AT)+\Pr(R=1)\Pr(C)}\\
&=\frac{\Pr(AT)}{\Pr(AT)+\Pr(C)}\Pr(Y^{(1)}=1|AT)+\frac{\Pr(C)}{\Pr(AT)+\Pr(C)}\Pr(Y^{(1)}=1|C)\\
&=\frac{\rho_A}{\rho_A+\rho_C}\omega_A^1+\frac{\rho_C}{\rho_A+\rho_C}\omega_C^1.
\end{aligned}
$$

Note: According to the assumptions of the IV, R is independent of $Y^{(1)}$ and the principal stratum, thus in the above equation, $Pr(Y^{(1)}=1,AT,R=1)=Pr(Y^{(1)}=1,AT)Pr(R=1)$ and $Pr(Y^{(1)}=1,C,R=1)=Pr(Y^{(1)}=1,C)Pr(R=1)$.

$$
\begin{aligned}
&\Pr\left(Y^{(0)}=1|Z=0,R=0\right)\\
&=\frac{\Pr(NT)}{\Pr(NT)+\Pr(C)}\Pr(Y^{(0)}=1|NT)+\frac{\Pr(C)}{\Pr(NT)+\Pr(C)}\Pr(Y^{(0)}=1|C)\\
&=\frac{1-\rho_A-\rho_C}{1-\rho_A}\omega_N^0+\frac{\rho_C}{1-\rho_A}\omega_C^0,
\end{aligned}
$$

$$
\begin{aligned}
&\Pr\left(Y=1|R=1\right)\\
&=\Pr\left(Y^{(1)}=1,Z=1|R=1\right)+\Pr\left(Y^{(0)}=1,Z=0|R=1\right)\\
&=\Pr(Y^{(1)}=1|Z=1,R=1)\Pr(Z=1|R=1)+\\
&\Pr(Y^{(0)}=1|Z=0,R=1)\Pr(Z=0|R=1)\\
&=\left(\frac{\rho_A}{\rho_A+\rho_C}\omega_A^0+\frac{\rho_C}{\rho_A+\rho_C}\omega_C^1\right)(\rho_A+\rho_C)+\omega_N^0\left(1-\rho_A-\rho_C\right)\\
&=\rho_A\omega_A^0+\rho_C\omega_C^1+\omega_N^0-\rho_A\omega_N^0-\rho_C\omega_N^0,
\end{aligned}
$$

$$
\begin{aligned}
&\Pr\left(Y=1|R=0\right)\\
&=\Pr\left(Y^{(1)}=1,Z=1|R=0\right)+\Pr\left(Y^{(0)}=1,Z=0|R=0\right)\\
&=\Pr(Y^{(1)}=1|Z=1,R=0)\Pr(Z=1|R=0)+\\
&\Pr(Y^{(0)}=1|Z=0,R=0)\Pr(Z=0|R=0)\\
&=\omega_A^0\rho_A+\left(\frac{1-\rho_A-\rho_C}{1-\rho_A}\omega_N^0+\frac{\rho_C}{1-\rho_A}\omega_C^0\right)(1-\rho_A)\\
&=\rho_A\omega_A^0+\rho_C\omega_C^0+\omega_N^0-\rho_A\omega_N^0-\rho_C\omega_N^0.
\end{aligned}
$$

A2. Prove: $\Pr(Y=1|Z,E)=expit(\lambda_0+\lambda_1 Z+\lambda_2 E)$ is not the true model and the true model should include the interaction between Z and E, or the interaction between Z and R. When there are no always-takers, the true model does not include the interaction.

Proof: The true model is

$$\begin{aligned}
\Pr(Y=1|Z,E) &= \Pr(Y=1|Z,R) \\
&= E(Y|Z,R) \\
&= I_{(Z=0,R=0)}E(Y|Z=0,R=0)+I_{(Z=1,R=0)}E(Y|Z=1,R=0) \\
&\quad +I_{(Z=0,R=1)}E(Y|Z=0,R=1)+I_{(Z=1,R=1)}E(Y|Z=1,R=1) \\
&= E(Y|Z=0,R=0) \\
&\quad +Z\left[E(Y|Z=1,R=0)-E(Y|Z=0,R=0)\right] \\
&\quad +R\left[E(Y|Z=0,R=1)-E(Y|Z=0,R=0)\right] \\
&\quad +ZR\left[\begin{array}{c} E(Y|Z=1,R=1)-E(Y|Z=1,R=0) \\ -E(Y|Z=0,R=1)+E(Y|Z=0,R=0) \end{array}\right] \\
&= \lambda_0+\lambda_1 Z+\lambda_2 R+\lambda_3 ZR.
\end{aligned}$$

In the above equations,

$$\begin{aligned}
\lambda_0 &= E(Y|Z=0,R=0), \\
\lambda_1 &= \left[E(Y|Z=1,R=0)-E(Y|Z=0,R=0)\right], \\
\lambda_2 &= \left[E(Y|Z=0,R=1)-E(Y|Z=0,R=0)\right], \\
\lambda_3 &= \begin{array}{c} E(Y|Z=1,R=1)-E(Y|Z=1,R=0) \\ -E(Y|Z=0,R=1)+E(Y|Z=0,R=0) \end{array} \\
&= E(Y|Z=1,R=1)-(\lambda_0+\lambda_1+\lambda_2).
\end{aligned}$$

So the true model includes the interaction between Z and R.

When there are no always-takers, we have $I_{(Z=1,R=0)} \equiv 0$, then the true model

becomes

$$\begin{aligned}
\Pr(Y=1|Z,E) &= \Pr(Y=1|Z,R) \\
&= E(Y|Z,R) \\
&= I_{(Z=0,R=0)}E(Y|Z=0,R=0) \\
&+ I_{(Z=0,R=1)}E(Y|Z=0,R=1) + I_{(Z=1,R=1)}E(Y|Z=1,R=1) \\
&= E(Y|Z=0,R=0) \\
&+ R\left[E(Y|Z=0,R=1) - E(Y|Z=0,R=0)\right] \\
&+ Z\left[E(Y|Z=1,R=1) - E(Y|Z=0,R=1)\right] \\
&= \lambda_0 + \lambda_1 R + \lambda_2 Z.
\end{aligned}$$

In the above equations,

$$\begin{aligned}
\lambda_0 &= E(Y|Z=0,R=0), \\
\lambda_1 &= \left[E(Y|Z=0,R=1) - E(Y|Z=0,R=0)\right], \\
\lambda_2 &= \left[E(Y|Z=1,R=1) - E(Y|Z=0,R=1)\right].
\end{aligned}$$

The true model does not include the interaction term.

A3. Some details about the bias analysis.

a)When there is no confounding, the treatment effect estimated with 2SPS can be biased.

The bias of 2SPS estimator is:

$$B_{2SPS} = \frac{\begin{array}{c} logit\left(\rho_A\omega_A^1 + \rho_C\omega_C^1 + \omega_N^0 - \rho_A\omega_N^0 - \rho_C\omega_N^0\right) \\ -logit\left(\rho_A\omega_A^1 + \rho_C\omega_C^0 + \omega_N^0 - \rho_A\omega_N^0 - \rho_C\omega_N^0\right) \end{array}}{\rho_C} - logit\left(\omega_C^1\right) + logit\left(\omega_C^0\right).$$

One no-confounding scenario is that there are no always-takers, and compliers and never-takers have the same probability of potential outcome, e.g., $\rho_A = 0$ and $\omega_C^0 = \omega_N^0$. Plugging in these values to the above equation, we have,

$$\begin{aligned} B_{2SPS} &= \frac{\begin{array}{c} logit\left(0\omega_A^1 + \rho_C\omega_C^1 + \omega_C^0 - 0\omega_C^0 - \rho_C\omega_C^0\right) \\ -logit\left(0\omega_A^1 + \rho_C\omega_C^0 + \omega_C^0 - 0\omega_C^0 - \rho_C\omega_C^0\right) \end{array}}{\rho_C} - logit\left(\omega_C^1\right) + logit\left(\omega_C^0\right) \\ &= \frac{logit\left(\rho_C\omega_C^1 + \omega_C^0 - \rho_C\omega_C^0\right) - logit\left(\omega_C^0\right)}{\rho_C} - logit\left(\omega_C^1\right) + logit\left(\omega_C^0\right). \end{aligned}$$

This equation generally not 0. We can easily see that it is 0 if on linear scale instead of logit scale.

b)When there is no confounding, the treatment effect estimated with 2SRI is unbiased.

The bias of the 2SRI estimator with no always-takers is:

$$\begin{aligned} B_{2SRI} &= \lambda_1 - \psi \\ &= logit\left(\frac{\rho_A}{\rho_A + \rho_C}\omega_A^1 + \frac{\rho_C}{\rho_A + \rho_C}\omega_C^1\right) - logit(\omega_N^0) \\ &- \frac{1}{\rho_C} logit\left(\frac{1 - \rho_A - \rho_C}{1 - \rho_A}\omega_N^0 + \frac{\rho_C}{1 - \rho_A}\omega_C^0\right) + \frac{1}{\rho_C} logit(\omega_N^0) \\ &- logit\left(\omega_C^1\right) + logit\left(\omega_C^0\right). \end{aligned}$$

. Plug in $\rho_A = 0$ and $\omega_C^0 = \omega_N^0$ to this equation, we have:

$$
\begin{aligned}
B_{2SRI} &= \lambda_1 - \psi \\
&= logit\left(\frac{0}{0+\rho_C}\omega_A^1 + \frac{\rho_C}{0+\rho_C}\omega_C^1\right) - logit(\omega_C^0) \\
&- \frac{1}{\rho_C} logit\left(\frac{1-0-\rho_C}{1-0}\omega_C^0 + \frac{\rho_C}{1-0}\omega_C^0\right) + \frac{1}{\rho_C} logit(\omega_C^0) \\
&- logit\left(\omega_C^1\right) + logit\left(\omega_C^0\right) \\
&= logit\left(\omega_C^1\right) - logit(\omega_C^0) - \frac{1}{\rho_C} logit(\omega_C^0) + \frac{1}{\rho_C} logit(\omega_C^0) \\
&- logit\left(\omega_C^1\right) + logit\left(\omega_C^0\right) \\
&= 0.
\end{aligned}
$$

.

Chapter 3

Variance Estimate of Causal Odds Ratio with Instrumental Variable Two-Stage Logistic Regression

3.1 Introduction

Instrumental variable (IV) methods are used to estimate effects of receiving treatment or exposure to risk factor on an outcome when there is unmeasured confounding in medical research, such as in clinical trials under non-adherence to treatment (45; 43; 40; 80; 81; 82; 47), or in non-randomized studies (41; 42; 83). While there has some research on IV estimates of causal odds ratios for binary responses (84; 69; 67), little has been written on variance estimation beyond cross-validation estimation. In this paper, under a logistic regression model for the confounded effect of treatment or exposure on a binary outcome, we propose sandwich variance estimators for two

different two-stage instrumental variable estimators. The two stage logistic regression approaches we consider are the two-stage predictor substitution (2SPS) and two-stage residual inclusion (2SRI) methods (70; 85; 39; 79). We evaluate the bias of these variance estimators relative to the true variability of the IV point estimates by simulations and evaluate the coverage of confidence interval calculated with the variance estimator we proposed. In this paper, we use "treatment" to represent either treatment received or exposure to a risk factor. An IV has the following properties: a) it is associated with treatment; b) it has no direct causal effect on the outcome; and c) it is independent of all unmeasured confounders of the treatment-outcome relationship (45; 40; 41; 86). Under these conditions, the IV analysis of the treatment-outcome relationship controls for measured and unmeasured confounding (43; 47; 48; 49). For randomized trials, the IV is randomized treatment assignment, but for observational studies it needs to be a carefully selected under the above assumptions The 2SPS and 2SRI IV approaches generally involve, as a first stage, the modeling of treatment as a function of the IV and any baseline covariates and then the second stage modeling of outcome as some function of predicted treatment and the covariates from the first stage regression. Under the 2SPS approach, predicted treatment from the first stage model replaces observed treatment as the principal covariate in the second stage model relating outcome to treatment (88; 55). Under the 2SRI method (70; 85), predicted and observed treatment are used to compute a residual that is included as a covariate in the second stage model where the principal covariate is observed treatment.

Angrist, Imbens and Rubin provided a good interpretation for the causal effect of

the general instrumental variable strategy (43). Under the potential outcome framework, they set up principal stratification framework under the assumptions that the proposed IV is associated with treatment, is independent of unmeasured confounders given the measured confounders and that the IV only affects outcome through treatment received and there are no defiers. With the principal stratification framework, patients are classified by the compliance status of treatment assignment as always-takers, compliers and never-takers. For the linear model, they analytically proved that under the above assumptions, the treatment effect estimated by the 2SPS IV method can be interpreted as average causal effect of compliers, which is called local average treatment effect (LATE), or compliers average causal effect (CACE). Since the 2SPS and 2SRI approaches give the same estimates with linear regression (70), the linear 2SRI also has the interpretation of LATE or CACE (43; 85).

For the linear two stage methods, the naive variance estimate obtained with ordinary least squares under the second stage regression model of treatment on outcome is not correct, since it does not adjust for the variability of the predicted treatment as a covariate in this model. The variance estimator for the 2SPS IV estimator is based on a heteroskedasticity-robust or sandwich estimator of variance involving cross products of the predicted treatment vector and a scalar dispersion factor based on the observed treatment factor (75). For the 2SRI approach, we have not found any published research on the variance estimator, but the estimate needs to be adjusted in a similar way, as the 2SRI and 2SPS approaches yield the same estimate of treatment effect for the linear case (70).

For the two-stage logistic regression, we have derived the sandwich variance es-

timator using a similar approach by Murphy and Topel (87) for the parameters in the second stage when the second stage model include parameters estimates obtained from the first stage. This approach is concisely described by Wooldridge (75) and it was also used by Zeger and Liang to get asymptotically unbiased variance estimators of the generalized estimating equations model (GEE) for longitudinal data analysis (89; 90). Other variance estimation approaches such as those based on the full information maximum likelihood (FIML) require the specification of the joint likelihood, which is generally not done with two stage procedures (91). When the joint likelihood is specified, this procedure uses the joint likelihood functions with respect to coefficients in both steps to yield efficient estimators and asymptotically correct estimates of variances (91). However, when there are many parameters to be estimated, this approach is computationally impractical. For this reason and because the joint likelihood is not specified under the two stage approaches, we implement the method by Murphy and Topel in estimating the standard errors of the two stage estimators. This paper is organized as follows. In Section 2, we introduce notation on observed and potential variables based on the principal stratification framework and we specify the two-stage logistic regression models. We use the Wooldridge's approach for two-step M estimation to derive the variance estimator of the 2SPS and 2SRI estimators in section 3. In Section 4, we did simulations to compare the naive variance estimate and our adjusted estimates with the observed variance of estimates of the treatment effect. We also compare variance, mean standard error (SME) and 95% coverage of 2SPS and 2SRI by simulation. Finally, we discuss the results and conclusions.

3.2 Notation and parameter setting

We define R as the IV, which in randomized trial setting is randomized treatment assignment, and Z as the treatment received. R=1 means that a patient is assigned to the study treatment, and R=0 means a patient is assigned to the other treatment (or no treatment). Similarly, Z=1 means that a patient receives the study treatment, and Z=0 means that a patient receives the other treatment (or non-treatment). Y is the observed binary outcome. With this definition, R is the instrumental variable. Under the principal stratification framework which provides the causal estimand for the two stage procedures, we also define as the probability of a subject being in the always-taker (AT) class, and as the corresponding probability for the complier class. Under the principal stratification framework with the no-defier assumption, only ATs can get the study treatment when they are assigned to other treatment or no treatment. Consequently, the first stage logistic regression for both the 2SPS and 2SRI approaches is parameterized as:

$$E\left(Z|\mathbf{r}\right) = expit\left(\mathbf{r}^T\boldsymbol{\rho}\right) = expit\left(\rho_A + \rho_C r\right). \tag{3.2.1}$$

In the above equation, $\mathbf{r}^T = (1, r)$ and $\boldsymbol{\rho}^T = (\rho_A, \rho_C)$.

For the 2SPS approach, the second stage logistic regression is the outcome on the predicted treatment-received (i.e., the expected value of Z conditional on R) from the first stage regression, which is,

$$E\left(Y|\hat{z}\right) = expit\left(\hat{\mathbf{z}}^T\boldsymbol{\lambda}\right) = expit\left(\lambda_1 + \lambda_2\widehat{z}\right). \tag{3.2.2}$$

In the above equation $\hat{\mathbf{z}}^T = (1, \hat{z})$ and $\boldsymbol{\lambda}^T = (\lambda_1, \lambda_2)$ With this regression, the

treatment effect is λ_2.

For the 2SRI, the second stage logistic regression is the model of outcome on treatment received AND the residual from the first stage regression:

$$\begin{aligned} EY &= expit\left(\lambda_1 + \lambda_2 z_i + \lambda_2 \hat{e}\right) \\ &= expit\left(\lambda_1 + \lambda_2 z_i + \lambda_2 \left[z_i - expit\left(\rho_A + \rho_C r_i\right)\right]\right). \end{aligned} \tag{3.2.3}$$

With this regression, the treatment effect is also λ_2. Our goal is to derive and evaluate variance estimators for the 2SPS and 2SRI estimators of λ_2.

3.3 Variance estimate of 2 stage logistic regression

We use the Wooldridge's approach for two-step M estimation to derive the variance estimator of the 2SPS and 2SRI estimators of the CACE log odds ratio for receiving treatment. Accordingly, we derive separate objective functions for the first and second stage models from which we obtain separate score and Hessian equations.

3.3.1 Variance estimate of 2SPS

For the 2SPS approach, we derive score and Hessian functions for the first and second stage models. For the first stage model in (3.2.1), the objective function for the parameters is defined as the log of the binomial mass function for an individual response:

$$q_1\left(z, \mathbf{r}; \boldsymbol{\rho}\right) = z_i\left(\rho_A + \rho_C r_i\right) - \ln\left[1 + \exp\left(\rho_A + \rho_C r_i\right)\right]. \tag{3.3.1}$$

The estimators of ρ_A and ρ_C maximize $\sum q_1(z, \mathbf{r}; \rho)$, and solves the first-order condition:

$$\sum \mathbf{s}_i(z, \mathbf{r}; \boldsymbol{\rho}) = 0, \tag{3.3.2}$$

where $\mathbf{s}_i(z, \mathbf{r}; \rho)$ is the two-dimensonal vector score for the objective function $q_1(z, \mathbf{r}; \boldsymbol{\rho})$ for an individual subject, derived by taking the first order partial derivatives of $q_1(z, \mathbf{r}; \boldsymbol{\rho})$ with respect to the dimensional parameter vector $\boldsymbol{\rho}$.

Similarly, let denote the Hession matrix of $q_1(z, \mathbf{r}; \boldsymbol{\rho})$ with respect to $\boldsymbol{\rho}$ for an individual subject as $\mathbf{H}_1(z, \mathbf{r}; \rho) \equiv \partial^2 q_1(z, \mathbf{r}; \rho) / \partial\rho\partial\rho'$, then by a Taylor series expansion with $\hat{\boldsymbol{\rho}}$ converging in probability to $\boldsymbol{\rho}^*$, we have,

$$\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}^* \approx \left(\sum \mathbf{H}_1(z, \mathbf{r}; \hat{\boldsymbol{\rho}})\right)^{-1} \left(-\sum \mathbf{s}_1(z, \mathbf{r}; \hat{\boldsymbol{\rho}})\right), \tag{3.3.3}$$

where $\hat{\boldsymbol{\rho}}^T = (\hat{\rho}_A, \hat{\rho}_C)$ from the first stage regression.

For the second stage of the 2SPS approach, with $\boldsymbol{\rho}$ replaced with $\hat{\rho}$ from the first stage, the objective function based on equation (3.2.2) is,

$$\begin{aligned} q_2(y, \mathbf{r}; \hat{\boldsymbol{\rho}}, \boldsymbol{\lambda}) &= y\hat{\mathbf{z}}^T\boldsymbol{\lambda} - \ln\left(1 + \exp\left(\hat{\mathbf{z}}^T\boldsymbol{\lambda}\right)\right) \qquad (3.3.4) \\ &= y(\lambda_1 + \lambda_2\hat{z}) - \ln\left[1 + \exp(\lambda_1 + \lambda_2\hat{z})\right] \\ &= y(\lambda_1 + \lambda_2 expit(\hat{\rho}_A + \hat{\rho}_C r)) - \ln\left[1 + \exp(\lambda_1 + \lambda_2 expit(\hat{\rho}_A + \hat{\rho}_C r))\right]. \end{aligned}$$

The estimators of λ_1 and λ_2 maximize $q_2(y, \mathbf{r}; \hat{\boldsymbol{\rho}}, \boldsymbol{\lambda})$, i.e., solves the first-order condition,

$$\sum \mathbf{s}_2(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \boldsymbol{\lambda}) = 0, \tag{3.3.5}$$

where $\mathbf{s}_2(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \boldsymbol{\lambda})$ is the score of the objective function $q_2(y, \mathbf{r}; \hat{\boldsymbol{\rho}}, \boldsymbol{\lambda})$ for individual i with respect to λ. Similarly, let $H_2(y, \mathbf{r}; \hat{\boldsymbol{\rho}}, \boldsymbol{\lambda})$ denote the Hession matrix of

the objective function $q_2(y, \mathbf{r}; \hat{\rho}, \lambda)$, with respect to $\boldsymbol{\lambda}$.

We then take the following Taylor expansion with $\hat{\boldsymbol{\lambda}}$ converging in probability to $\boldsymbol{\lambda}^*$, we have,

$$\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^* \approx \left(\sum \mathbf{H}_2\left(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right)\right)^{-1} \left(-\sum \mathbf{s}_2\left(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right)\right), \tag{3.3.6}$$

where $\mathbf{s}_2\left(y_i, \mathbf{r}_i; \hat{\rho}, \hat{\lambda}\right)$ and $\mathbf{H}_2\left(y_i, \mathbf{r}_i; \hat{\rho}, \hat{\lambda}\right)$ are obtained from $\mathbf{s}_2(y_i, \mathbf{r}_i; \hat{\rho}, \lambda)$ and $\mathbf{H}_2(y_i, \mathbf{r}_i; \hat{\rho}, \lambda)$ respectively by replacing λ with $\hat{\lambda}$. Given the series expansion in (3.3.6), one variance estimate of $\widehat{\lambda}$ is,

$$\begin{aligned}\widehat{\mathbf{V}}\left(\hat{\boldsymbol{\lambda}}\right)_{naive} &= \left(\sum \mathbf{H}_2\left(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right)\right)^{-1} \sum \mathbf{s}_2\left(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right) \mathbf{s}_2\left(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right)^T \\ &\quad \left(\sum \mathbf{H}_2\left(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \hat{\lambda}\right)\right)^{-1}. \end{aligned} \tag{3.3.7}$$

However, this variance estimate does not take into account that the variability of $\hat{\mathbf{z}}$ through $\hat{\boldsymbol{\rho}}$. That is, it does not take into account the series expansion in (6) for $\hat{\boldsymbol{\rho}}$. Consequently, we achieve such an adjustment by incorporating the Taylor series expansion in (3.3.3) with the following Taylor series expansion of $\sum \mathbf{s}_2(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \boldsymbol{\lambda})$ around $\boldsymbol{\rho}^*$, the probability limit of $\hat{\boldsymbol{\rho}}$, as follows:

$$\sum \mathbf{s}_2(y, \mathbf{r}; \hat{\boldsymbol{\rho}}, \boldsymbol{\lambda}) \approx \sum \mathbf{s}_2(y, \mathbf{r}; \boldsymbol{\rho}^*, \boldsymbol{\lambda}) + \sum \frac{\partial \mathbf{s}_2(y, \mathbf{r}; \boldsymbol{\rho}^*, \boldsymbol{\lambda})}{\partial \rho} (\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}^*), \tag{3.3.8}$$

where $\frac{\partial \mathbf{s}_2(y, \mathbf{r}; \boldsymbol{\rho}^*, \boldsymbol{\lambda})}{\partial \boldsymbol{\rho}}$ is the partial derivative of $\mathbf{s}_2(y, \mathbf{r}; \boldsymbol{\rho}^*, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\rho}$ evaluated at $\boldsymbol{\rho}^*$. From the approximation in (3.3.3) for $(\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}^*)$, the expansion in (3.3.3) is asymptotically equivalent to:

$$\begin{aligned}\sum \mathbf{s}_2(y, \mathbf{r}; \hat{\boldsymbol{\rho}}, \boldsymbol{\lambda}) &\approx \sum \mathbf{s}_2(y, \mathbf{r}; \boldsymbol{\rho}^*, \boldsymbol{\lambda}) + \sum \frac{\partial \mathbf{s}_2(y, \mathbf{r}; \boldsymbol{\rho}^*, \boldsymbol{\lambda})}{\partial \boldsymbol{\rho}} \left(\sum \mathbf{H}_1(z, \mathbf{r}; \hat{\boldsymbol{\rho}})\right)^{-1} \\ &\quad \left(-\sum \mathbf{s}_1(z, \mathbf{r}; \hat{\boldsymbol{\rho}})\right). \end{aligned} \tag{3.3.9}$$

Based on (3.3.9), we follow the general case of Wooldridge (75) by adjusting the score evaluated at $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\lambda}}$, $\mathbf{s}_2\left(y, \mathbf{r}; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right)$, as $\mathbf{g}\left(y, \mathbf{r}; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right)$,

$$\mathbf{g}\left(y, \mathbf{r}; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right) = \mathbf{s}_2\left(y, \mathbf{r}; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right) + \sum \frac{\partial \mathbf{s}_2\left(y, \mathbf{r}; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right)}{\partial \boldsymbol{\rho}} \left(\sum \mathbf{H}_1\left(z_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}\right)\right)^{-1} \left(-\mathbf{s}_1\left(z_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}\right)\right). \tag{3.3.10}$$

Replacing the $\mathbf{s}_2\left(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right)$ by $\mathbf{g}\left(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right)$, the variance estimator for the 2SPS estimator of $\boldsymbol{\lambda}$, adjusted for the first stage regression estimate of $\boldsymbol{\rho}$ is,

$$\begin{aligned} \widehat{\mathbf{V}}\left(\hat{\boldsymbol{\lambda}}\right)_{adjust} &= \left(\sum \mathbf{H}_2\left(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right)\right)^{-1} \sum \mathbf{g}\left(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right) \mathbf{g}\left(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right)^T \\ &\quad \left(\sum \mathbf{H}_2\left(y_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}}\right)\right)^{-1}. \end{aligned} \tag{3.3.11}$$

3.3.2 Variance estimator for the 2SRI approach

For the 2SRI approach, we take the same strategy as above and adjust the second stage objective function for $\boldsymbol{\lambda}$ with the corresponding approximation in (3.3.3) to adjust for $hat\boldsymbol{\rho}$ under the first stage regression. Accordingly, the second stage objective function for $\boldsymbol{\lambda}$ under the 2SRI approach is:

$$q_2\left(y, z; \hat{e}, \boldsymbol{\lambda}\right) = y\left(\lambda_1 + \lambda_2 z + \lambda_3 \hat{e}\right) - ln\left(1 + exp\left(\lambda_1 + \lambda_2 z + \lambda_3 \hat{e}\right)\right) \tag{3.3.12}$$

The corresponding adjusted score is,

$$\mathbf{g}\left(y, z; \hat{e}, \hat{\boldsymbol{\lambda}}\right) = \mathbf{s}_2\left(y, z; \hat{e}, \hat{\boldsymbol{\lambda}}\right) + \sum \frac{\partial \mathbf{s}_2\left(y, z; \hat{e}, \hat{\boldsymbol{\lambda}}\right)}{\partial \boldsymbol{\rho}} \left(\sum \mathbf{H}_1\left(z_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}\right)\right)^{-1} \left(-\mathbf{s}_1\left(z_i, \mathbf{r}_i; \hat{\boldsymbol{\rho}}\right)\right) \tag{3.3.13}$$

which is then plugged into the adjusted variance estimator for in (3.3.11) to get the adjusted variance estimator for 2SRI.

For two stage linear regression, we analytically prove that the above 2SPS and 2SRI variance estimators, which are equivalent, equal the heteroskedasticity-robust variance estimate of the simple two stage linear regression (See the detail in attachment A).

3.4 Simulations

Since our research is focused on the causal inference of two stage logistic regression, we simulated data with the principal stratification settings of the Angrist-Rubin model of causal inference. The data sets were generated with the following algorithm.

Step 1: Generate a data set with total number of N subjects. Among these subjects, always-takers (ATs), compliers (Cs), and never-takers (NTs) are generated from multinomial distributions with probability ρ_A for ATs, ρ_C for Cs, and ρ_N for NTs.

Step 2: With the probability of r, randomly assign rN subjects to R=1 and the rest of $(1-r)N$ subjects to R=0.

Step 3: Create the potential outcome $Y^{(0)}$ and $Y^{(1)}$ based on the compliance status of each subject, and the probability of potential outcome for each compliance status.

Step 4: Determine treatment received Z of each subject based on the treatment assignment R and compliance status.

Step 5: With the following equation, determine observed outcome of each subject based on the potential outcome and treatment received,

$$Y = Y^{(1)}Z + Y^{(0)}(1-Z)$$

We simulated data sets with different compliance rates, different confounding factors, and different sample sizes with information from our example data analysis. For each setting of the parameters, we simulated 2000 data sets and performed the 2SPS and 2SRI methods for logistic regression. For each of these estimation approaches, we calculated the simulation-based variance of the log odds ratio estimates, in addition to the averages of the naive variance, adjusted variance, and bootstrap variance estimated at each iteration for the corresponding estimator of the log odds ratio for treatment on outcome. We also used these different variance estimates to calculate 95% coverage estimates for the log odds ratio. Finally, we calculated the mean square error for the 2SPS and 2SRI estimates of the log odds ratio for treatment on outcome.

3.5 Result

Tables 3.1, 3.2, 3.3 and 3.4 present the observed simulated variance of the two stage log odds ratio estimates and percentage difference between the observed sample variance and the mean of the different variance estimates listed above. Tables 3.1 and 3.2 present results for different sample sizes for the 2SPS approach, and Tables 3.3 and 3.4 do the same for the 2SRI approach. First in Tables 3.1 and 3.2, we examine the impact of the sample size on the simulated variance of the 2SPS log odds ratio estimates. We can see that this variance decreases by about 50% when the sample size increases from 500 to 1000. In both tables, the simulated variance increases slightly when the confounding becomes severe, but it increases dramatically when the compliance rate decreases. For example, in Table 3.2, the variance increases about 6 times

from 0.03 to 0.18 when the compliance rate decreases from 0.7 to 0.3. In Tables 3.1 and 3.2, we next compare the average of the naive and adjusted variance estimates with the simulated variance of the 2SPS log odds ratio estimates. The percentage difference between the average adjusted variance and the simulated variance is occasionally greater than 5%, but usually very small. The maximum difference from all simulations is 7.34%. In contrast, the percentage difference between the average naive variance estimate and the simulated variance is often large with a maximum of 32.85%. This difference is larger when the compliance is lower, and increases as the confounding becomes more severe. In Tables 3.1 and 3.2, we also compare the average of the adjusted and bootstrap variance estimates. The adjusted variance is usually closer to the simulated variance than is the bootstrap variance, especially when the sample size is smaller. The difference between the average bootstrap variance and the simulated variance is as high as 18.42% when the compliance rate is 0.3 and the confounding is very severe ($\delta = 3$) for sample size 500. As with the average naive variance estimate, the difference between the average bootstrap variance and the simulated variance increases as the compliance rate becomes lower and the confounding becomes more severe. Nevertheless, overall, the average bootstrap variance estimate is closer to the simulated variance than is the average naive variance estimate. Tables 3.3 and 3.4 provide similar comparisons for the 2SRI approach as do Tables 3.1 and 3.2 for the 2SPS approach. For the 2SRI approach, the simulated variance of the log odds ratio estimates also decreases by about 50% when the sample size increase from 500 to 1000, and increases as the compliance rate is lower and as the confounding gets more severe. Again, the change of the simulated variance is much more sensitive

to compliance rate than to severity of confounding. For the 2SRI approach, Tables 3.3 and 3.4 show similar results as displayed in Tables 3.1 and 3.2 for the 2SPS approach. For the 2SRI method, the average adjusted variance estimates are closer to the simulated variance than are the average naive and bootstrap variance estimates. The average naive variance estimate differs substantially from the simulated variance (as high as 27.27%) when the confounding is severe. Similarly to 2SPS approach, the average bootstrap estimated variance of the 2SRI log odds ratio estimator is substantially different from the simulated variance when the compliance rate is low and confounding severe. With the sample size of 500, this difference can be as high as 19%. Tables 3.5 and 3.6 present the width and coverage of 95% confidence intervals for the 2SPS and 2SRI approaches, respectively. For the 2SPS approach, the 95% confidence interval coverage based on the adjusted variance estimate is low (minimum of 92.35%) when the compliance rate is 0.3 and δ is -3 to -2 (indicating very severe confounding). Other than these settings, the 95% confidence intervals based on the adjusted variance estimate for both the 2SPS and 2SRI approaches have coverage close to 95%. For all of the different settings and different sample sizes, the 2SPS approach has narrower adjusted variance-based 95% confidence intervals than the 2SRI approach. Fig. 3.1 and 3.2 present the bias, variance and MSE of the 2SPS and 2SRI log odds ratio estimates, respectively. The results show that for most of the settings, the 2SRI logs odds ratio estimator has larger variance than the 2SPS log odds ratio estimate, even though the bias of the 2SRI estimator is often smaller than that of the 2SPS approach when confounding is not severe. For many settings, the 2SRI variance is twice that of the 2SPS variance, which leads to a higher MSE for

the 2SRI approach compared to the 2SPS approach. When the sample size is 5000 (Table 3.2), the variance of both log odds ratio estimators is small, so the MSE is mainly determined by the bias. As shown in Cai et al. (79), both the 2SPS and 2SRI approaches are similarly biased for the log odds ratio of receiving treatment among compliers. In this situation, the MSE of 2SRI approach is close to the 2SPS approach.

3.6 Discussion

In this paper, we applied the theory of two-step M estimation to obtain the adjusted variance estimators for the 2SPS and 2SRI IV estimators of the log odds ratio of receiving treatment among compliers. Our simulation results show that the adjusted variance estimators provide good estimates of the simulated variance of the causal log odds ratio. In addition, these adjusted variance estimators perform better than the corresponding bootstrap estimators. We found that the average bootstrap variance is not accurate when the compliance rate is low, which is consistent with other studies that have shown that the validity of bootstrap is questionable when the IV is weakly correlated with the endogenous explanatory variable (92).

Furthermore, our simulation results indicate that the naive variance estimate without the adjustment for two stage regression can be severely biased when the compliance rate is low and confounding is severe. This is true even for the 2SRI approach, when the causal log odds ratio is the coefficient for the variable of treatment received, instead of the expected value of treatment estimated from the first stage, which is the case for the 2SPS approach. Our simulation results also demonstrate that the

adjusted variance estimators for the 2SPS and 2SRI methods are very sensitive to the compliance rate, but not very sensitive to the severity of confounding. This result informs the estimation of power and sample size when planning the IV analysis of two stage logistic regression. In our simulation results, the 95% confidence interval coverage based on the adjusted variance is good for both the 2SPS and 2SRI approaches, even when the bias of the corresponding IV log odds ratio estimators is severe, which is encouraging for the application of these two IV approaches. Our previous research focused on the bias of the 2SPS and 2SRI log odds ratio estimators of receiving treatment among compliers (79). In this paper, when comparing the 2SPS and 2SRI approaches with respect to the bias, variance and MSE of these estimators, we found that the variance of the 2SRI estimates is usually larger than that of 2SPS approach, so the MSE of 2SRI approach is usually greater than that of 2SPS approach even though 2SRI is less biased for some settings as shown by Cai et al. (79). With this result, we conclude that 2SRI approach does not have an advantage over the 2SPS approach.

Table 3.1. Comparison of adjusted variance estimates with naïve estimates and the variance estimated by bootstrap for the percentage difference from the sample variance of simulation: 2SPS approach with small sample size.

Compliance Rate ρ_C	Delta	Sample Variance	Naïve Estimate (% Difference)	Adjusted Variance (% Difference)	Bootstrap Estimate (% Difference)
0.3	-3	0.3821	32.85	4.02	12.76
	-2.5	0.3839	29.96	3.06	11.37
	-2	0.3939	23.52	-0.19	10.78
	-1.5	0.3814	23.06	1.95	9.61
	-1	0.3717	20.45	3.41	7.22
	-0.5	0.3563	19.09	6.65	8.20
	0	0.3663	10.11	3.92	9.63
	0.5	0.3865	0.98	0.92	8.29
	1	0.4108	-5.31	-0.18	7.00
	1.5	0.4596	-12.93	-2.48	12.57
	2	0.5114	-17.67	-2.37	14.97
	2.5	0.5830	-23.37	-4.16	17.15
	3	0.6357	-25.73	-3.27	18.43
0.5	-3	0.1206	23.01	2.77	5.79
	-2.5	0.1205	22.37	2.84	5.03
	-2	0.1223	19.52	1.42	4.72
	-1.5	0.1227	17.52	1.20	6.97
	-1	0.1240	14.30	0.47	5.15
	-0.5	0.1267	9.45	-1.02	5.09
	0	0.1254	8.33	1.41	5.50
	0.5	0.1250	7.06	4.31	7.04
	1	0.1339	-0.49	1.18	5.50
	1.5	0.1396	-3.73	1.90	4.83
	2	0.1471	-7.15	1.79	8.61
	2.5	0.1529	-8.92	2.67	10.06
	3	0.1599	-11.40	2.01	8.95
0.7	-3	0.0629	11.81	2.32	2.96
	-2.5	0.0628	11.88	2.56	3.26
	-2	0.0630	11.38	2.44	3.14
	-1.5	0.0637	10.02	1.68	4.32
	-1	0.0645	8.34	0.86	3.29
	-0.5	0.0657	5.96	-0.34	3.44
	0	0.0652	6.33	1.36	2.46
	0.5	0.0650	6.24	2.95	2.52
	1	0.0688	0.29	-1.07	0.69
	1.5	0.0695	-0.83	-0.53	2.13
	2	0.0703	-1.85	-0.13	3.48
	2.5	0.0708	-2.48	0.31	3.17
	3	0.0717	-3.65	-0.13	2.35

Note: Number of iteration of bootstrap p=500; Sample size N=500, Simulation time M=2000.

Outcome rate for treatment group $\omega_{11}= \omega_{12}= \omega_{13}=0.6$;

Outcome rate for comparison group for always-takers and compliers $\omega_{01}= \omega_{02}=0.3$.

Rate of always-taker $\rho_A=0.2$ and never-taker $\rho_A =1- \rho_C- \rho_N$; Delta=logit ω_{03}-logit ω_{02}.

Table 3.2. Comparison adjusted variance estimates with naïve estimates and the variance estimated by bootstrap for the percentage difference from the sample variance of simulation: 2SPS approach with large sample size.

Compliance Rate ρ_C	δ	Sample Variance	Naïve Estimate (% Difference)	Adjusted Variance (% Difference)	Bootstrap Estimate (% Difference)
0.3	-3	0.1859	30.15	0.64	3.69
	-2.5	0.1799	32.28	3.50	3.74
	-2	0.1774	30.80	4.24	4.43
	-1.5	0.1790	25.12	2.45	6.44
	-1	0.1781	20.07	1.87	3.33
	-0.5	0.1782	13.81	0.88	4.03
	0	0.1798	7.34	0.03	4.44
	0.5	0.1838	1.62	-0.09	5.11
	1	0.1843	0.98	4.67	7.14
	1.5	0.2084	-8.18	0.45	5.57
	2	0.2342	-14.09	-1.15	4.10
	2.5	0.2614	-18.42	-1.85	5.67
	3	0.2892	-22.14	-2.79	5.87
0.5	-3	0.0636	15.97	-3.34	2.66
	-2.5	0.0640	14.49	-3.99	1.64
	-2	0.0633	14.70	-2.90	0.25
	-1.5	0.0628	14.19	-1.94	-0.13
	-1	0.0648	8.78	-4.63	0.23
	-0.5	0.0653	5.68	-4.71	0.64
	0	0.0670	0.96	-5.74	0.81
	0.5	0.0675	-1.40	-4.22	1.33
	1	0.0687	-3.44	-2.19	0.86
	1.5	0.0744	-10.16	-5.27	-1.22
	2	0.0800	-15.11	-7.34	-1.94
	2.5	0.0833	-16.87	-6.65	0.18
	3	0.0860	-18.12	-6.09	0.61
0.7	-3	0.0332	5.86	-3.27	-0.48
	-2.5	0.0334	5.13	-3.74	-0.39
	-2	0.0331	5.87	-2.76	-0.22
	-1.5	0.0332	5.25	-2.85	0.01
	-1	0.0336	3.87	-3.43	0.78
	-0.5	0.0337	3.11	-3.14	1.73
	0	0.0345	0.27	-4.53	2.99
	0.5	0.0346	-0.16	-3.38	2.91
	1	0.0357	-3.54	-4.98	-1.38
	1.5	0.0365	-5.70	-5.54	-1.18
	2	0.0370	-6.85	-5.35	-1.43
	2.5	0.0374	-7.89	-5.34	-1.39
	3	0.0378	-8.54	-5.27	-0.09

Note: Number of iteration of bootstrap p=700. Sample size N=1000, Simulation time M=2000.

Outcome rate for treatment group $\omega_{11}= \omega_{12}= \omega_{13}=0.6$;

Outcome rate for comparison group for always-takers and compliers $\omega_{01}= \omega_{02}=0.3$.

Rate of always-taker $\omega_1= 0.2$ and never-taker $\omega_3=1- \omega_1- \omega_2$; Delta=logit ω_{03}-logit ω_{02}.

Table 3.3. Comparison of adjusted variance estimates with naïve estimates and the variance estimated by bootstrap for the percentage difference from the sample variance of simulation: 2SRI approach with small sample size.

Compliance Rate ρ_C	Delta	Sample Variance	Naïve Estimate (% Difference)	Adjusted Variance (% Difference)	Bootstrap Estimate (% Difference)
0.3	-3	0.8518	-9.06	3.86	16.43
	-2.5	0.8134	-9.46	1.95	14.82
	-2	0.7689	-10.73	-1.25	14.10
	-1.5	0.6629	-5.92	1.57	11.43
	-1	0.5693	-2.42	2.88	9.33
	-0.5	0.4767	2.97	6.24	9.67
	0	0.4347	0.97	3.52	10.61
	0.5	0.4168	-2.74	0.90	8.86
	1	0.4205	-6.52	-0.40	7.55
	1.5	0.4620	-13.19	-2.56	13.12
	2	0.5200	-18.12	-2.18	15.61
	2.5	0.6098	-24.32	-3.57	18.29
	3	0.6894	-27.27	-2.39	19.45
0.5	-3	0.2077	-6.03	1.57	6.83
	-2.5	0.2007	-4.83	2.01	5.30
	-2	0.1948	-5.13	0.54	5.33
	-1.5	0.1845	-4.22	0.11	7.69
	-1	0.1720	-3.00	-0.21	5.73
	-0.5	0.1606	-2.80	-1.34	5.54
	0	0.1459	0.41	1.24	5.74
	0.5	0.1351	2.99	4.31	7.45
	1	0.1382	-2.11	1.02	6.06
	1.5	0.1406	-4.10	1.89	5.33
	2	0.1479	-7.35	1.81	9.06
	2.5	0.1554	-9.38	2.75	10.44
	3	0.1649	-12.15	2.20	9.50
0.7	-3	0.0795	-0.97	1.60	4.48
	-2.5	0.0788	-0.55	1.83	4.69
	-2	0.0781	-0.36	1.73	4.41
	-1.5	0.0776	-0.72	0.98	5.48
	-1	0.0767	-0.93	0.29	4.26
	-0.5	0.0758	-1.52	-0.78	4.09
	0	0.0727	0.59	1.00	3.08
	0.5	0.0701	2.29	2.70	3.04
	1	0.0723	-2.28	-1.47	1.22
	1.5	0.0715	-2.33	-0.80	2.55
	2	0.0714	-2.68	-0.28	3.88
	2.5	0.0715	-2.96	0.24	3.58
	3	0.0722	-4.00	-0.19	2.79

Note: Number of iteration of bootstrap p=500; Sample size N=500, Simulation time M=2000.

Outcome rate for treatment group $\omega_{11}= \omega_{12}= \omega_{13}=0.6$;

Outcome rate for comparison group for always-takers and compliers $\omega_{01}= \omega_{02}=0.3$.

Rate of always-taker $\rho_A=0.2$ and never-taker $\rho_A =1- \rho_C- \rho_N$; Delta=logit ω_{03}-logit ω_{02}.

Table 3.4. Comparison adjusted variance estimates with naïve estimates and the variance estimated by bootstrap for the percentage difference from the sample variance of simulation: 2SRI approach with large sample size.

Compliance Rate ρ_C	δ	Sample Variance	Naïve Estimate (% Difference)	Adjusted Variance (% Difference)	Bootstrap Estimate (% Difference)
0.3	-3	0.4105	-10.76	0.21	6.77
	-2.5	0.3755	-7.16	2.71	5.56
	-2	0.3401	-4.43	3.79	6.07
	-1.5	0.3086	-4.10	1.91	7.53
	-1	0.2701	-2.22	1.54	4.17
	-0.5	0.2365	-1.09	0.79	4.65
	0	0.2118	-1.09	-0.02	4.94
	0.5	0.1977	-2.00	-0.08	5.42
	1	0.1881	-0.13	4.49	7.35
	1.5	0.2090	-8.31	0.41	5.81
	2	0.2380	-14.60	-1.10	4.25
	2.5	0.2740	-19.71	-1.64	5.70
	3	0.3152	-24.26	-2.57	5.74
0.5	-3	0.1072	-10.07	-3.40	5.24
	-2.5	0.1049	-9.98	-4.06	3.94
	-2	0.0993	-7.86	-2.86	2.06
	-1.5	0.0928	-5.64	-1.88	1.22
	-1	0.0885	-6.55	-4.32	1.44
	-0.5	0.0822	-5.68	-4.70	1.27
	0	0.0778	-6.51	-6.12	1.17
	0.5	0.0730	-5.24	-4.39	1.56
	1	0.0706	-4.76	-2.17	1.20
	1.5	0.0747	-10.35	-5.21	-1.04
	2	0.0802	-15.15	-7.27	-1.76
	2.5	0.0843	-17.09	-6.48	0.35
	3	0.0882	-18.46	-5.65	0.79
0.7	-3	0.0416	-5.66	-3.43	0.70
	-2.5	0.0415	-5.95	-3.91	0.73
	-2	0.0407	-4.63	-2.85	0.82
	-1.5	0.0402	-4.36	-2.95	0.93
	-1	0.0397	-4.48	-3.52	1.67
	-0.5	0.0387	-3.82	-3.33	2.38
	0	0.0384	-4.92	-4.73	3.49
	0.5	0.0372	-3.77	-3.58	3.27
	1	0.0373	-5.59	-5.00	-1.13
	1.5	0.0374	-6.85	-5.58	-0.92
	2	0.0375	-7.46	-5.38	-1.20
	2.5	0.0377	-8.23	-5.38	-1.18
	3	0.0379	-8.73	-5.27	0.11

Note: Number of iteration of bootstrap p=700. Sample size N=1000, Simulation time M=2000.

Outcome rate for treatment group $\omega_{11}= \omega_{12}= \omega_{13}=0.6$;

Outcome rate for comparison group for always-takers and compliers $\omega_{01}= \omega_{02}=0.3$.

Rate of always-taker $\omega_1= 0.2$ and never-taker $\omega_3=1- \omega_1- \omega_2$; Delta=logit ω_{03}-logit ω_{02}.

Table 3.5. Comparison of width and coverage of 95% confidence intervals for the true log odds ratio between 2SPS and 2SRI approaches with small sample size.

Compliance Rate ρ_C	δ	2SPS				2SRI			
		Adjusted Estimate		Bootstrap Estimate		Adjusted Estimate		Bootstrap Estimate	
		Width	%Coverage	Width	%Coverage	Width	%Coverage	Width	%Coverage
0.3	-3	2.4383	92.35	2.5509	91.50	3.6301	96.15	3.8499	97.20
	-2.5	2.4335	92.95	2.5436	92.15	3.5154	96.00	3.7203	97.25
	-2	2.4264	93.10	2.5314	92.50	3.3647	95.70	3.5453	97.00
	-1.5	2.4144	94.45	2.5167	94.35	3.1715	96.05	3.3366	96.70
	-1	2.4011	95.70	2.4951	94.55	2.9601	96.60	3.0994	96.40
	-0.5	2.3882	96.30	2.4778	95.90	2.7551	96.45	2.8747	96.50
	0	2.3895	96.20	2.4752	96.25	2.5969	96.25	2.6992	96.35
	0.5	2.4152	96.00	2.5021	95.70	2.5077	96.20	2.6041	95.80
	1	2.4775	95.40	2.5705	95.75	2.5034	95.45	2.6039	96.00
	1.5	2.5843	94.70	2.6817	96.10	2.5900	94.75	2.6935	96.25
	2	2.7194	95.45	2.8362	95.90	2.7450	95.65	2.8699	96.15
	2.5	2.8669	95.95	3.0059	96.60	2.9421	95.95	3.0937	97.05
	3	3.0017	96.20	3.1629	97.10	3.1415	95.85	3.3204	97.10
0.5	-3	1.3755	94.45	1.3993	94.75	1.7917	95.55	1.8477	96.20
	-2.5	1.3756	94.80	1.3992	94.95	1.7653	95.55	1.8166	96.15
	-2	1.3761	94.65	1.3990	95.00	1.7269	95.60	1.7748	96.25
	-1.5	1.3771	95.05	1.3995	95.70	1.6775	95.65	1.7205	96.30
	-1	1.3793	95.40	1.4011	96.00	1.6175	95.25	1.6558	96.05
	-0.5	1.3842	95.20	1.4064	95.50	1.5549	95.10	1.5903	95.85
	0	1.3936	95.85	1.4167	95.95	1.5014	95.75	1.5331	96.05
	0.5	1.4107	95.35	1.4351	95.45	1.4667	95.90	1.4972	95.85
	1	1.4379	94.95	1.4636	95.00	1.4593	94.80	1.4894	95.30
	1.5	1.4724	95.55	1.5013	95.00	1.4776	95.60	1.5102	94.95
	2	1.5100	95.40	1.5418	95.80	1.5143	95.45	1.5498	95.95
	2.5	1.5452	95.70	1.5815	95.95	1.5585	95.60	1.5987	96.10
	3	1.5746	95.80	1.6136	96.00	1.6002	95.65	1.6438	96.15
0.7	-3	0.9936	95.25	1.0022	95.95	1.1125	95.15	1.1314	95.50
	-2.5	0.9942	95.15	1.0027	96.10	1.1090	95.15	1.1274	95.60
	-2	0.9951	95.40	1.0036	95.95	1.1036	95.30	1.1212	95.70
	-1.5	0.9965	95.30	1.0049	96.10	1.0962	95.30	1.1128	96.00
	-1	0.9989	94.85	1.0075	95.70	1.0860	94.95	1.1017	95.80
	-0.5	1.0022	94.80	1.0113	96.05	1.0739	95.20	1.0888	95.90
	0	1.0069	95.40	1.0165	95.95	1.0613	95.40	1.0755	95.80
	0.5	1.0133	95.70	1.0233	95.40	1.0508	95.75	1.0645	95.50
	1	1.0215	94.35	1.0317	94.90	1.0448	94.75	1.0581	94.75
	1.5	1.0295	94.45	1.0406	95.45	1.0430	94.50	1.0568	95.55
	2	1.0371	94.45	1.0485	95.20	1.0447	94.40	1.0586	95.40
	2.5	1.0432	94.65	1.0554	95.15	1.0478	94.65	1.0623	95.30
	3	1.0478	94.45	1.0609	94.95	1.0510	94.55	1.0663	95.05

Note: Number of iteration of bootstrap p=500; Sample size N=500; Simulation time M=2000.

Outcome rate for treatment group $\omega_{11}= \omega_{12}= \omega_{13}=0.6$;

Outcome rate for comparison group for always-takers and compliers $\omega_{01}= \omega_{02}=0.3$.

Rate of always-taker $\rho_A=0.2$ and never-taker $\rho_A =1- \rho_C- \rho_N$; Delta=logit ω_{03}-logit ω_{02}.

Table 3.6. Comparison of width and coverage of 95% confidence intervals for the true log odds ratio between 2SPS and 2SRI approaches with large sample size.

Compliance Rate ρ_C	δ	2SPS				2SRI			
		Adjusted Estimate		Bootstrap Estimate		Adjusted Estimate		Bootstrap Estimate	
		Width	%Coverage	Width	%Coverage	Width	%Coverage	Width	%Coverage
0.3	-3	1.6852	89.10	1.7148	89.40	2.4967	95.25	2.5585	95.70
	-2.5	1.6814	89.65	1.7109	90.40	2.4178	96.15	2.4773	95.80
	-2	1.6762	91.40	1.7053	90.70	2.3140	95.85	2.3679	96.10
	-1.5	1.6694	92.00	1.6971	92.60	2.1846	95.60	2.2329	95.95
	-1	1.6609	93.05	1.6869	93.95	2.0412	95.10	2.0819	95.75
	-0.5	1.6538	95.10	1.6786	95.90	1.9038	95.60	1.9385	96.30
	0	1.6543	95.40	1.6769	95.75	1.7949	95.20	1.8227	95.80
	0.5	1.6715	95.00	1.6941	95.50	1.7334	95.30	1.7593	95.55
	1	1.7121	95.35	1.7377	95.10	1.7282	95.75	1.7560	95.15
	1.5	1.7825	95.40	1.8114	95.55	1.7845	95.45	1.8153	95.65
	2	1.8726	95.65	1.9060	95.70	1.8880	95.55	1.9237	95.70
	2.5	1.9694	95.30	2.0083	96.30	2.0183	94.85	2.0609	96.45
	3	2.0597	94.90	2.1021	96.30	2.1523	92.15	2.1995	94.25
0.5	-3	0.9702	93.90	0.9755	94.35	1.2587	94.85	1.2732	96.05
	-2.5	0.9704	94.10	0.9754	94.40	1.2407	94.70	1.2544	95.70
	-2	0.9706	94.25	0.9755	94.40	1.2146	95.00	1.2267	95.65
	-1.5	0.9713	94.65	0.9761	94.65	1.1801	94.95	1.1909	95.95
	-1	0.9729	94.35	0.9778	95.25	1.1384	94.55	1.1485	95.75
	-0.5	0.9764	94.75	0.9812	95.25	1.0949	94.55	1.1036	95.40
	0	0.9832	94.55	0.9881	94.85	1.0578	94.95	1.0653	94.90
	0.5	0.9953	94.65	1.0006	95.00	1.0339	94.55	1.0410	95.20
	1	1.0141	94.10	1.0200	94.75	1.0283	94.35	1.0358	95.05
	1.5	1.0385	93.85	1.0454	94.15	1.0412	93.95	1.0492	94.15
	2	1.0649	93.30	1.0728	94.55	1.0668	93.30	1.0758	94.50
	2.5	1.0902	93.85	1.0991	95.15	1.0980	94.00	1.1081	95.20
	3	1.1111	94.20	1.1210	94.95	1.1274	94.20	1.1387	94.65
0.7	-3	0.7020	94.55	0.7029	94.95	0.7850	94.45	0.7883	94.85
	-2.5	0.7024	94.00	0.7031	94.90	0.7826	94.45	0.7856	94.70
	-2	0.7030	94.30	0.7037	94.85	0.7787	94.55	0.7817	94.55
	-1.5	0.7040	94.15	0.7047	94.75	0.7734	94.70	0.7762	95.25
	-1	0.7056	94.40	0.7064	94.55	0.7663	94.60	0.7691	95.50
	-0.5	0.7081	94.40	0.7091	94.85	0.7577	94.55	0.7605	95.35
	0	0.7115	94.00	0.7126	95.00	0.7491	94.70	0.7517	95.50
	0.5	0.7161	94.50	0.7175	95.15	0.7418	94.45	0.7445	95.35
	1	0.7217	94.10	0.7229	94.60	0.7376	94.15	0.7399	94.80
	1.5	0.7275	94.30	0.7290	94.05	0.7364	94.35	0.7388	94.30
	2	0.7328	93.80	0.7346	94.00	0.7375	93.80	0.7401	94.20
	2.5	0.7374	93.90	0.7394	94.25	0.7399	94.15	0.7427	94.25
	3	0.7408	93.80	0.7430	94.45	0.7422	93.75	0.7453	94.55

Note: Number of iteration of bootstrap p=700; Sample size N=1000; Simulation time M=2000.

Outcome rate for treatment group $\omega_{11}= \omega_{12}= \omega_{13}=0.6$;

Outcome rate for comparison group for always-takers and compliers $\omega_{01}= \omega_{02}=0.3$.

Rate of always-taker $\rho_A=0.2$ and never-taker $\rho_A =1- \rho_C- \rho_N$; Delta=logit ω_{03}-logit ω_{02}.

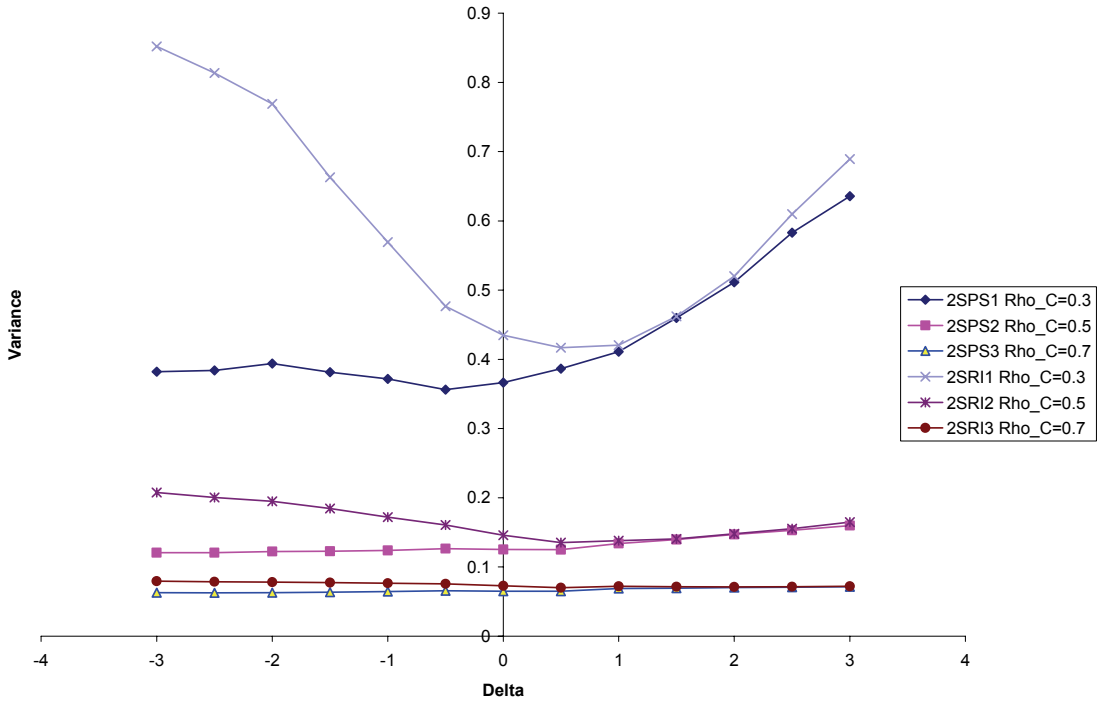

Figure 3.1. Comparison of Variance between 2SPS and 2SRI with sample size N=500.

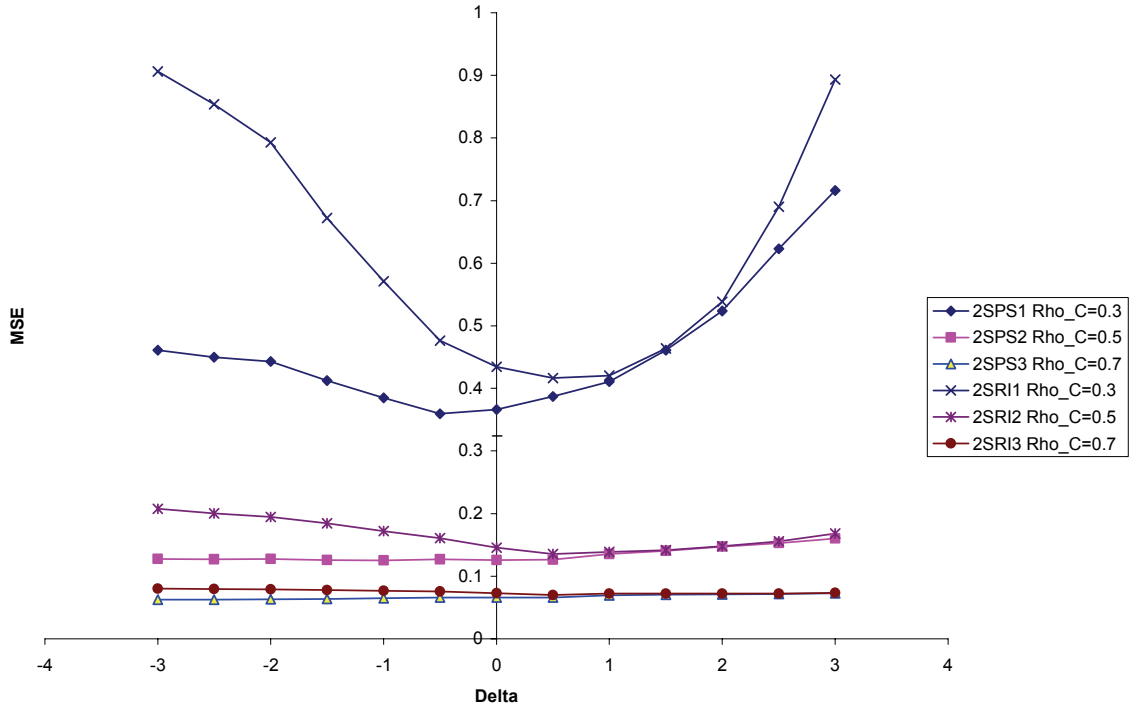

Figure 3.2. Comparison of MSE between 2SPS and 2SRI with sample size N=500.

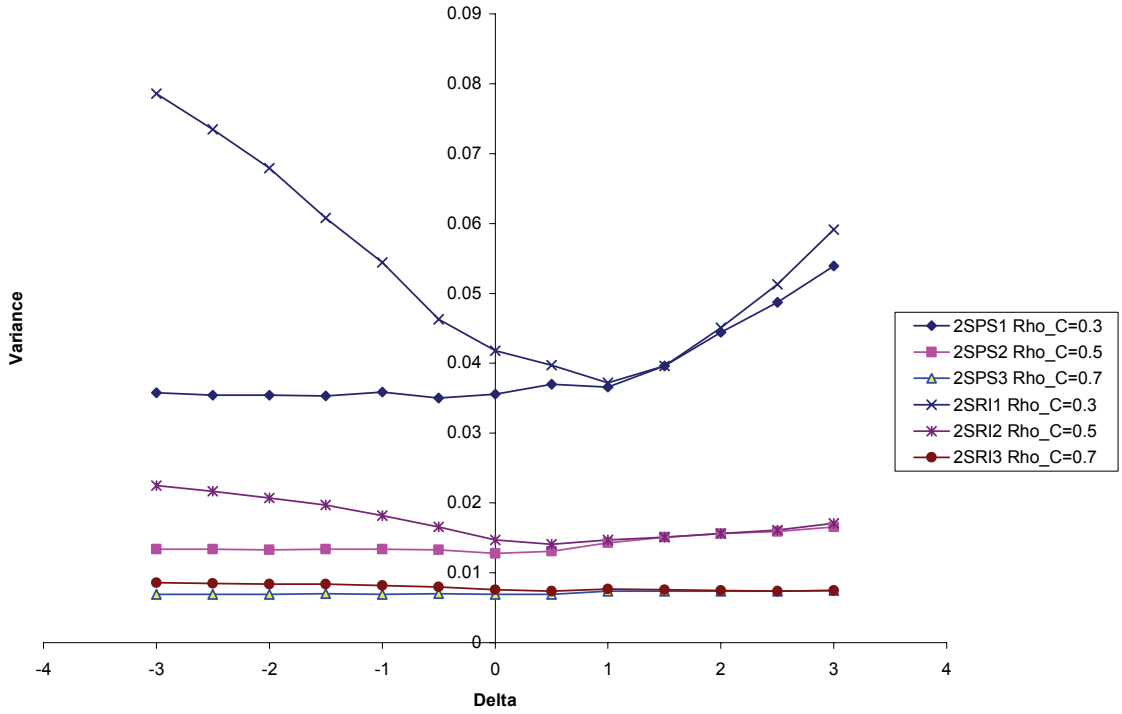

Figure 3.3. Comparison of Variance between 2SPS and 2SRI with sample size N=5000.

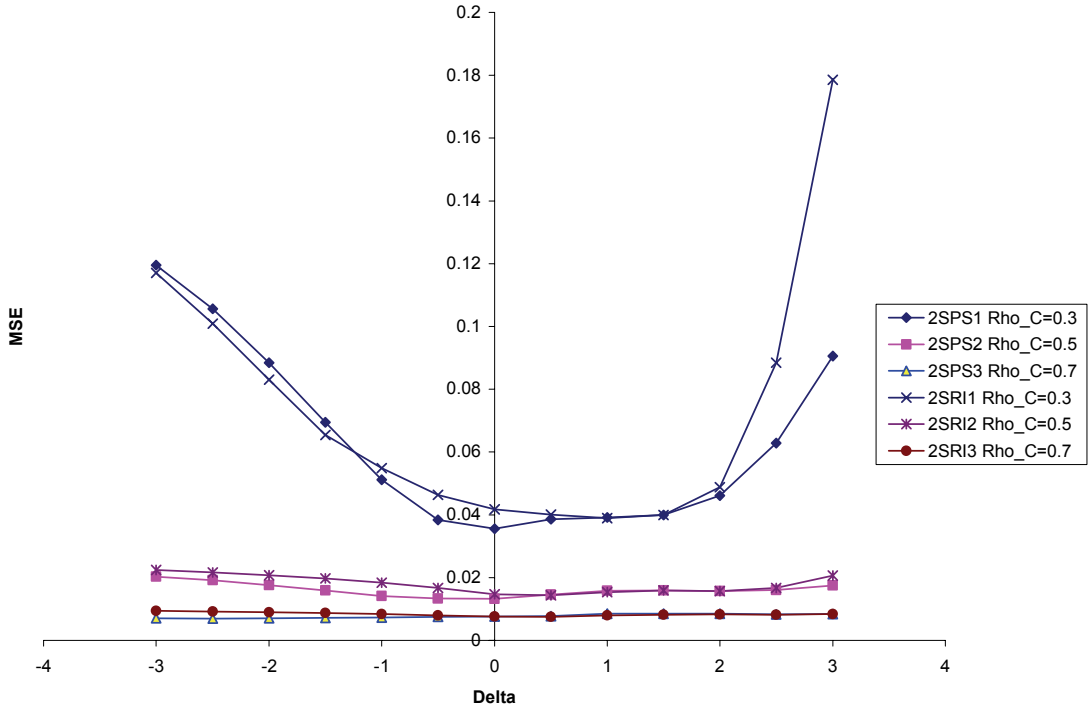

Figure 3.4. Comparison of MSE between 2SPS and 2SRI with sample size N=5000.

3.7 Appendix: The adjusted variance estimate is equal to the heteroskedasticity robust variance estimate for the simple linear case.

Prove:

The adjusted variance estimate is equal to the heteroskedasticity robust variance estimate for the simple linear case.

Proof:

For the simple two stage linear regression, we assume that the two stage linear model is as followed.

First stage:

$$z_i = \rho r_i + u_i.$$

Second stage:

$$y_i = \hat{z}_i \lambda + v_i.$$

Then we have the objective function for the first stage:

$$q_1(\rho) = (z - \rho r)^2,$$

and the score is,

$$s_1(\rho) = \frac{\partial q_1(\rho)}{\partial \rho} = (z - \rho r) r,$$

and the Hession is,

$$H_1(\rho) = \frac{\partial q_1^2(\rho)}{\partial \rho^2} = -r^2.$$

The objective function for the second stage is,

$$q_2(\lambda) = (y - \hat{z}\lambda)^2.$$

The score is,

$$s_2(\lambda) = \frac{\partial q_2(\lambda)}{\partial \lambda} = (z - \lambda\hat{z})\,\hat{z} = (z - \lambda\hat{\rho}r)\,\hat{\rho}r,$$

and the Hession is,

$$H_2(\lambda) = \frac{\partial q_2^2(\lambda)}{\partial \lambda^2} = -\hat{z}^2.$$

The adjusted score function is,

$$\begin{aligned}
\mathbf{g}\left(y, r; \hat{\rho}, \hat{\lambda}\right) &= \mathbf{s}_2\left(y, r; \hat{\rho}, \hat{\lambda}\right) + \sum \frac{\partial \mathbf{s}_2\left(y, r; \hat{\rho}, \hat{\lambda}\right)}{\partial \rho} \left(\sum \mathbf{H}_1\left(z_i, r_i; \hat{\rho}\right)\right)^{-1} \\
&\quad \left(-\mathbf{s}_1\left(z_i, r_i; \hat{\rho}\right)\right) \\
&= \left(y - \hat{\lambda}\hat{\rho}r\right)\hat{\rho}r + \sum \left(y_i r_i - 2\hat{\lambda}\hat{\rho}r_i^2\right)\left(\sum -r_i^2\right)^{-1}\left[-\left(z_i - \hat{\rho}r_i\right)r_i\right].
\end{aligned}$$

So the estimated variance of two stage linear regression by the Wooldrige process is,

$$\begin{aligned}
\widehat{\mathbf{V}}\left(\hat{\lambda}\right)_{adjust} &= \left(\sum \mathbf{H}_2\left(y_i, r_i; \hat{\rho}, \hat{\lambda}\right)\right)^{-1} \sum \mathbf{g}\left(y_i, r_i; \hat{\rho}, \hat{\lambda}\right)\mathbf{g}\left(y_i, r_i; \hat{\rho}, \hat{\lambda}\right)^T \\
&\quad \left(\sum \mathbf{H}_2\left(y_i, r_i; \hat{\rho}, \hat{\lambda}\right)\right)^{-1} \\
&= \left[\sum \left(\hat{z}_i\right)^2\right]^{-2} \sum \left[\begin{array}{c} \left(y_i - \hat{\lambda}\hat{\rho}r_i\right)\hat{\rho}r_i + \\ \sum \left(y_i r_i - 2\hat{\lambda}\hat{\rho}r_i^2\right)\left(\sum -r_i^2\right)^{-1}\left[-\left(z_i - \hat{\rho}r_i\right)r_i\right] \end{array}\right]^2
\end{aligned}$$

Comparing with the heteroskedasticity-robust estimate of variance:

$$\begin{aligned}
\hat{Var}\left(\hat{\lambda}\right) &= \left(\hat{\mathbf{Z}}'\hat{\mathbf{Z}}\right)^{-1}\left(\sum \hat{u}_i^2 \hat{\mathbf{z}}_i' \hat{\mathbf{z}}_i\right)\left(\hat{\mathbf{Z}}'\hat{\mathbf{Z}}\right)^{-1} \\
&= \left(\sum \hat{z}_i^2\right)^{-2} \sum \left[\left(y_i - z_i\hat{\lambda}\right)\left(\hat{z}_i\right)\right]^2,
\end{aligned}$$

we need to show,

$$\sum \left[\left(y_i - z_i \hat{\lambda} \right) (\hat{z}_i) \right]^2 = \sum \begin{bmatrix} \left(y_i - \hat{\lambda} \hat{\rho} r_i \right) \hat{\rho} r_i + \sum \left(y_i r_i - 2 \hat{\lambda} \hat{\rho} r_i^2 \right) \\ \left(\sum -r_i^2 \right)^{-1} \left[- \left(z_i - \hat{\rho} r_i \right) r_i \right] \end{bmatrix}.$$

Since,

$$\begin{aligned} RHS &= \sum \left[\left(y_i - \hat{\lambda} \hat{z}_i \right) \hat{z}_i + \sum \left(y_i r_i - 2 \hat{\lambda} \hat{\rho} r_i^2 \right) \left(\sum -r_i^2 \right)^{-1} \left[- \left(z_i - \hat{\rho} r_i \right) r_i \right] \right]^2 \\ &= \sum \begin{bmatrix} \left(y_i - z_i \hat{\lambda} \right) (\hat{z}_i) + (z_i - \hat{z}_i) \hat{\lambda} \hat{z}_i + \sum \left(y_i r_i - 2 \hat{\lambda} \hat{\rho} r_i^2 \right) \\ \left(\sum -r_i^2 \right)^{-1} \left[- \left(z_i - \hat{\rho} r_i \right) r_i \right] \end{bmatrix}^2. \end{aligned}$$

we need to prove,

$$(z_i - \hat{z}_i) \hat{\lambda} \widehat{z_i} + \sum \left(y_i r_i - 2 \hat{\lambda} \hat{\rho} r_i^2 \right) \left(\sum -r_i^2 \right)^{-1} \left[- \left(z_i - \hat{\rho} r_i \right) r_i \right] = 0.$$

Proof:

$$\begin{aligned} & (z_i - \hat{z}_i) \hat{\lambda} \hat{z}_i + \sum \left(y_i r_i - 2 \hat{\lambda} \hat{\rho} r_i^2 \right) \left(\sum -r_i^2 \right)^{-1} \left[- \left(z_i - \hat{\rho} r_i \right) r_i \right] \\ = \; & (z_i - \hat{z}_i) \hat{\lambda} \hat{z}_i + \sum \left(y_i r_i - 2 \hat{\lambda} \hat{\rho} r_i^2 \right) \left(\sum r_i^2 \right)^{-1} (z_i - \hat{z}_i) r_i \\ = \; & (z_i - \hat{z}_i) \left[\hat{\lambda} \hat{z}_i + \sum \left(y_i r_i - 2 \hat{\lambda} \hat{\rho} r_i^2 \right) \left(\sum r_i^2 \right)^{-1} r_i \right] \\ = \; & (z_i - \hat{z}_i) \left(\sum r_i^2 \right)^{-1} \left[\hat{\lambda} \hat{z}_i \left(\sum r_i^2 \right)^{-1} + \sum \left(y_i r_i - 2 \hat{\lambda} \hat{\rho} r_i^2 \right) r_i \right] \\ = \; & (z_i - \hat{z}_i) \left(\sum r_i^2 \right)^{-1} \left[\frac{\sum r_i z_i}{\sum r_i^2} r_i \hat{\lambda}_i \left(\sum r_i^2 \right)^{-1} + r_i \sum \left(y_i r_i - 2 \hat{\lambda} \hat{\rho} r_i^2 \right) \right] \\ = \; & (z_i - \hat{z}_i) \left(\sum r_i^2 \right)^{-1} r_i \left[\sum r_i z_i \hat{\lambda}_i + \sum \left(y_i r_i - 2 \hat{\lambda} \hat{\rho} r_i^2 \right) \right] \\ = \; & (z_i - \hat{z}_i) \left(\sum r_i^2 \right)^{-1} r_i \sum \left(r_i z_i \hat{\lambda}_i + y_i r_i - 2 \hat{\lambda} \hat{\rho} r_i^2 \right) \\ = \; & (z_i - \hat{z}_i) \left(\sum r_i^2 \right)^{-1} r_i \sum r_i \left(z_i \hat{\lambda}_i - \hat{\lambda} \hat{\rho} r_i + y_i - \hat{\lambda} \hat{\rho} r_i \right) \\ = \; & (z_i - \hat{z}_i) \left(\sum r_i^2 \right)^{-1} r_i \sum r_i \left[\hat{\lambda}_i (z_i - \hat{z}_i) + \left(y_i - \hat{\lambda} \hat{z}_i \right) \right] \\ = \; & (z_i - \hat{z}_i) \left(\sum r_i^2 \right)^{-1} r_i \sum r_i \left[\hat{\lambda}_i u_i + v_i \right] \\ = \; & 0 \end{aligned}$$

So,

$$\begin{aligned} RHS &= \sum \left[\left(y_i - \hat{\lambda} \hat{z}_i \right) \hat{z}_i + 0 \right]^2 \\ &= LHS \end{aligned}$$

Chapter 4

Different Approaches of Instrumental Variable Analysis of Antidiabetic Effect of Bezafibrate

4.1 Introduction

Two-stage logistic regression and generalize structural mean model (GSMM) are the two different approaches to analyze causal inference of binary outcomes. GSMM was proposed by Vansteelandt and Geotghebeur (66) as an extension of the structural mean model (SMM) (96; 97; 104; 68) to the logistic regression for binary outcomes under the randomized clinical trial (RCT) setting when patients assigned to the placebo group can not access the study treatment. Having analyzed bias, variance, mean standard error, and 95% confidence coverage of two-logistic logistic regression approaches and found that the two stage logistic regression is biased, we were motivated by the

epidemiologic study of casual antidiabetic effect of bezafibrate to apply the GSMM together with the two-stage logistic regressions to the analysis of non-randomized observational study. Vansteelandt and Geotghebeur mentioned that the GSMM can be extended to observational study (66), but there was no detailed discussion how it is implemented, and what is the property of GSMM when it is extended to non-randomized study. In this research, we extended the GSMM to the setting when patients assigned to the placebo group can access the study treatment, and we did simulations to evaluate bias, variance, mean standard error of the GSMM and compare it with the two-stage logistic regression approaches. The two-stage regression approaches include two-stage predictor substitution (2SPS) and two-stage residual inclusion (2SRI). Under the 2SPS approach, the first stage model yields the predicted value of treatment as a function of an instrument and covariates, and in the second stage model for the outcome, this predicted treatment replaces the observed treatment as covariate. Under the 2SRI approach, the first stage is the same, but the residual term of the first stage regression is included in the second stage regression, retaining the observed treatment as a covariate (70; 85). Angrist, Imbens and Rubin (43) proposed the causal model under the potential outcome framework originating from Neyman and Fisher in the early 20th century (20). Under specific assumptions we will described in the second section, they classified patients based on the potential compliance status, and analyzed treatment effect based on this classification, which is called principal stratification. Patients are called always-takers if they would take the study treatment no matter what their treatment assignment is; Patients are called never-takers if they would not take the study treatment no matter what their treatment

assignment is; Patients are called compliers if they follow the treatment assignment; Patients are called defiers if they would take treatment opposite to their treatment assignment. With this principle stratification, they proved that linear 2SPS estimator can be interpreted as the local average treatment effect (LATE) or the complier average causal effect (CACE) (43). Nagelkerke and Terza proved 2SRI estimates the same treatment effect as 2SPS (70; 85) for the linear model, thus it can also be interpreted as LATE or CACE. However, when the two-stage-linear regression is extended to the two-stage logistic regression for the binary outcome, both 2SPS and 2SRI are biased in general (79). In this research, we want to answer the question if the treatment effect estimated from the GSMM is unbiased estimate of CACE. More importantly, we want to evaluate the performance of the GSMM estimator when it is extended to be applied to the non-randomized observational study. To do this, we first wrote an R program to implement the GSMM extended to the setting when patients in both arms can access to the study treatment and used this program to perform simulations to evaluated performance of GSMM method under the principal stratification framework. Then we applied GSMM together with 2SPS and 2SRI to the analysis of antidiabetic effect of bezafibrate using the GPRD database. Bezafibrate is a fibrate that is widely prescribed in the U. K. for the treatment of dyslipidemia. Recently, it was found by conducting post-hoc analysis of the randomized clinical trial data that this fibrate may prevent diabetes in patients with cardiovascular disease or obesity (106; 107; 108; 109), while other fibrates [like ciprofibrate, clofibrate, fenofibrate, or gemfibrozil] were not found to have such effect. This may due to the fact that bezafibrate is a pan-PPAR agonist that can activate PPAR-α as well as PPAR-γ,

and thus have the same effect as anti-diabetes drug in the class PPAR-γ agonist like thiazolidinedione (TZD) (94), while other fibrates are much more selective for PPAR-α. Based on this observation, Flory et al performed a retrospective cohort study using General Practice Research Database (GPRD) to compare the instance of diabetes between bezafibrate users and other fibrate users. Their results suggested a protective effect of bezafibrate against diabetes (95). In this study, we will apply different IV approaches to the analysis of the same data to explore these methods and further assess whether there is causal effect of bezafibrate against diabetes. This paper is organized as follows. In the second section on method, we will first introduce notations and assumptions of causal inference frame work of principle stratification. Then we will describe GSMM model followed by two approaches of two-stage logistic regression models. About the GSMM model, we will describe how it was extended to the setting when patients assigned to the placebo arm can access the study treatment. Also in the method section, we will describe how datasets were generated by simulation with the principle stratification framework, how we did analysis with the simulated data. Then we will describe GPRD bezafibrate data and the strategy of analyzing this data. In the third and fourth section, we will present our result and discussion.

4.2 Method

4.3 Assumptions and notations

An IV is a factor that has the following properties: a) it is associated with treatment; b) it has no direct causal effect on the outcome; and c) it is independent of all unmeasured confounders of the treatment-outcome (45; 43; 46; 41). An IV analysis can provide causal inference of treatment effect that controls for all confounding including unmeasured confounding (42). According to the Angrist-Imbens-Rubin model, we have the following five assumption about IV analysis under the potential outcome framework (43): 1) Stable unit treatment value assumption (SUTVA) (71; 105), which means that potential outcomes for each person is unrelated to the treatment status of other individuals; this assumption also implies the consistency assumption, which means the potential outcome of a certain treatment will be the same regardless of the treatment assignment mechanism (73); 2) Random assignment assumption, which means that the IV is unrelated, as the randomized assignment, to all confounders in the randomized clinical trials, or it is unrelated to the unmeasured confounders (conditional on the measured confounders) of the treatment-outcome relationship in observational studies; 3) Exclusion restriction, which means that any effect of treatment assignment on outcomes must be via an effect of treatment assignment on treatment received; 4) Nonzero average causal effect of treatment assignment on treatment received, which means that the treatment assignment should be associated with treatment received; and 5) Monotonicity, which means that there is no one

who does the opposite of his/her treatment assignment, regardless of the actual assignment. Under these assumptions, we define Z as the observed treatment received; $Z = 1$ means a patient takes the study treatment and $Z = 0$ means a patient takes comparison treatment or non-treatment. As an IV, R is defined as treatment assignment in the randomized clinical trial setting; $R = 1$ means a patient is assigned to the study treatment and $R = 0$ means a patients is assigned to the comparison treatment or non-treatment. In non-randomized studies, R is defined as an IV that is associated with treatment received and meets all above IV criteria. For instance, if the physicians' preferences are used as IV, $R = 1$ means a physician's preference is the study treatment, and $R = 0$ means a physician's preference is the comparison treatment or non-treatment. For both contexts of clinical trial and observational study, we define Y as the observed outcome, $Y^{(1)}$ as the potential outcome if a patient takes study treatment, $Y^{(0)}$ as the potential outcome if a patients takes comparison treatment or non-treatment. We denote ω_{1C} as the potential outcome among compliers when exposed to the study treatment, and ω_{0C} as the potential outcome among compliers when untreated. Lastly, we define ρ_A, ρ_C , and ρ_N as the probability of always-takers, compliers, and never-takers respectively. Under the above monotonicity assumption, there are no never-takers. In the clinical trial when patients in the placebo arm can not access the study treatment, there are no always-takers, which means $\rho_A = 0$, but with observational studies, ρ_A can be any value from 0 to 1.

4.3.1 Generalized Structural Mean Models

The general class of structural mean models (SMMs) is defined under the potential outcome framework and estimated by G-estimation (96; 97; 104; 68; 110; 6). The SMM expresses the contrast of the means of observed outcomes and the potential treatment-free outcomes as a function of exposure to treatment. Under the randomization assumption, the average treatment effect can be estimated with a G-estimation equation based on the assumption that the means of potential outcome in the two randomization arms are equal. This process applies to both the linear model for estimating difference in means and the log linear model for estimating the log-ratio of means. However, for a logistic SMM with a binary outcome, no unbiased estimating equation exists for the causal parameter of the logistic SMM (68), because integration of the mean of a binary outcome under a logistic model with the law of iterated expectation does not produce a marginal mean for which the logistic model holds. To solve this problem, Vansteelandt and Goetghebeur (66) proposed the GSMM that augments the logistic SMM with a logistic regression model for the association between the binary outcome and observed exposure in each randomization arm or at level of the instrumental variable.

In the randomized trial context when controls cannot access treatment, the first component of the GSMM is the structural model for the causal log odds ratio of exposure on outcome conditional on the observed levels of the IV, exposure factor,

and any covariates:

$$\log it\left\{E\left(Y_i|Z_i,\mathbf{x}_{i,}R_i=1\right)\right\}-\log it\left\{E\left(Y_i^{(0)}|Z_i,\mathbf{x}_{i,}R_i=1\right)\right\} \quad (4.3.1)$$

$$= \eta_s'\left(Z_i,\mathbf{x}_i\right)\psi.$$

where ψ is the treatment effect at the logit scale, and $\eta_s'\left(Z_i,\mathbf{x}_i\right)$ is a function of Z_i and $\mathbf{x}_i$. For intendance, if $\eta_s'\left(Z_i,\mathbf{x}_i\right)=Z_i$, then ψ is the causal log odd ratio. If $\eta_s'\left(Z_i,\mathbf{x}_i\right)$ contains x, it means the causal effect is different for patients with different baseline characteristics $\mathbf{x}$. This structural model is the same regardless of whether it's a randomized trial with controls not having access to treatment and the IV context where everyone has access to treatment.

The second component of the GSMM is the association model for the log odds ratio of exposure on outcome conditional on observed levels of the exposure and covariates factors among those in the randomized to treatment group:

$$\log it\left\{E\left(Y_i|Z_i,\mathbf{x}_{i,}R_i=1\right)\right\}=\eta_{a1}'\left(Z_i,\mathbf{x}_i\right)\beta_1, \quad (4.3.2)$$

where subscript 1 in η_{a1} and β_1 means the association model is from treatment arm. $\eta_{a1}'\left(Z_i,\mathbf{x}_i\right)$ can be any function of Z_i and $\mathbf{x}_i$. β_1 is the parameter for the association of observed outcome and observed treatment and covariate $\mathbf{x}$. For instance, the usual form of $\eta_{a1}'\left(Z_i,\mathbf{x}_i\right)\beta_1$ is $Z\beta_{11}+x\beta_{12}$. For the clinical trial with controls not having access to treatment, equation (4.3.2) is the only association model because in the control arm, $Z=0$ for all patients, but in the observational studies, Z can be either 0 or 1 in the control arm, so we have the second association model in the control arm as follows,

$$\log it\left\{E\left(Y_i|Z_i,\mathbf{x}_{i,}R_i=0\right)\right\}=\eta_{a0}'\left(Z_i,\mathbf{x}_i\right)\beta_0. \quad (4.3.3)$$

Subscript 0 in η_{a0} and β_0 means the association model is from control arm.

Estimation:

In clinical trials when patients in the placebo arm can not access the study treatment, the exposure-free outcome for each subject i can be expressed as,

$$H_i(\psi,\beta) = \exp it\left\{\eta'_{a1}(Z_i,\mathbf{x}_i)\beta_1 - \eta'_s(Z_i,\mathbf{x}_i)\psi\right\}R_i + Y_i(1-R). \tag{4.3.4}$$

By the randomization assumption, the treatment arm and the placebo arm have the same conditional mean of H_i given $\mathbf{x}_i$, which leads to the equation,

$$\begin{aligned} & d(\mathbf{x}_i)\,E\left\{H_i(\psi,\beta)\,|\mathbf{x}_i, R_i=1\right\} - q(\mathbf{x}_i) \qquad (4.3.5)\\ = \;& d(\mathbf{x}_i)\left[E\left\{H_i(\psi,\beta)\,|\mathbf{x}_i, R_i=0\right\} - q(\mathbf{x}_i)\right] \end{aligned}$$

and the estimating equation for ψ as,

$$\begin{aligned} & \sum \frac{d(\mathbf{x}_i)R_i}{pr(R_i=1|\mathbf{x}_i)}\left\{H_i(\psi,\beta) - q(\mathbf{x}_i)\right\} \qquad (4.3.6)\\ = \;& \frac{d(\mathbf{x}_i)(1-R_i)}{pr(R_i=0|\mathbf{x}_i)}\left\{H_i(\psi,\beta) - q(\mathbf{x}_i)\right\}, \end{aligned}$$

where $d(\mathbf{x})$ and $q(\mathbf{x})$ can be any functions of covariates. We should select these functions to maximize the efficiency of the estimations (see Vansteelandt and Goetghebeur (66) regarding how to select $d(\mathbf{x})$ and $q(\mathbf{x})$.

The estimating equations for β_0 and β_1 is:(7)

$$\sum d_b(Z_i,\mathbf{x}_i)R_i\left[Y_i - \exp it\left\{\eta'_{ab}(Z_i,\mathbf{x}_i)\beta_b\right\}\right], \tag{4.3.7}$$

for b=0,1 and where $d(z,\mathbf{x})$ is an arbitary vector function of Z_i and $\mathbf{x}_i$, for instance, the maximum likelihood score equation.

In a randomized clinical trial when the control group can not access the study treatment, we can only obtain the estimate of β_1 from the estimation equation (4.3.7), and then obtain $H_i(\psi, \beta)$ from (4.3.4), so that can be used in equation (4.3.6) to estimate ψ by iterative form of G-estimation.

In the IV constext when every subject has access to treatment, the exposure-free outcome for each subject expressed as,

$$\begin{aligned} H_i(\psi, \beta) &= \exp it\left\{\eta'_{a1}(Z_i, \mathbf{x}_i)\beta_1 - \eta'_s(Z_i, \mathbf{x}_i)\psi\right\} R_i \\ &\quad + \exp it\left\{\eta'_{a0}(Z_i, \mathbf{x}_i)\beta_0 - \eta'_s(Z_i, \mathbf{x}_i)\psi\right\}(1-R) \end{aligned} \tag{4.3.8}$$

We need to estimate both β_0 and β_1 from equation (4.3.7), and then obtain $H_i(\psi, \beta)$ from (4.3.8), so that can be used in equation (4.3.6) to estimate ψ by iterative form of G-estimation.

In that above models, we need to select covariate x that is associated with both exposure and outcome in treatment arm and placebo arm.

4.3.2 Two stage logistic regression

The first stage logistic regression is the treatment received Z_i on the treatment assignment R_i and other covariate x_i.

$$E(Z) = \exp it(\rho_1 + \rho_2 R + \rho x)$$

For the 2SPS, the second stage logistic regression is the outcome on the expected value of Z_i from the first stage regression, which is,

$$E(Y) = \exp it(\lambda_1 + \lambda_2 \widehat{z} + \lambda_3 x)$$

With this regression, the causal treatment effect is λ_2, the coefficient of the expected value of Z.

For the 2SRI, the second stage logistic regression is the outcome on the treatment received AND the residual from the first stage regression, which is

$$E(Y) = \exp it(\lambda_1 + \lambda_2 z + \lambda_3 \widehat{e} + \lambda_4 x)$$

With this regression, the causal treatment effect is λ_2, the coefficient of the observed treatment received(6, 24, 34).

4.3.3 Simulations

In our previous simulations (79), we set the probability of compliance status as constant. In this study, we make our simulation more generalized by introducing a baseline covariate x and letting the probability of compliance status vary over x. The detailed algorithm of simulation is as follows:

Step 1: Generate a data set with total number of N subjects. Each subject as a continuous baseline variable x that follows the normal distribution with mean of 0 and variance of 1.

Step 2: For each subject, determine the probability of always-takers (ATs), Compliers (Cs), and never-takers (NTs) as follows:

$\rho_A = P(C) = expit(b_1 + b_2 x)$

$\rho_C = P(AT) = a(1 - P(C))$

$\rho_N = P(NT) = 1 - P(C) - P(AT)$

a, b_1, and b_2 are predetermined parameters that can determine probability of Cs,

ATs, and NTs. For instance, $a = 0$ means there are no ATs; when $b_1 = 0$ and the mean of x is 0, the probability of Cs is 0.5.

Step 3: With multinomial distribution, determine each subject's compliance situation, in R, the following program is used:

Comp=t(apply(w,1, function(w) rmultinom(1, 1, c(w[1], w[2], w[3]))))

w[1], w[2], w[3] are the probability of AT, C, and NT respectively.

Step 4: With the probability of probability r, randomly assign rN subjects to $R = 1$ and the rest of $(1 - r)\, N$ subjects to $R = 0$.

Step 5: Create the potential outcome $Y^{(0)}$ and $Y^{(1)}$ based on the compliance status of each subject, and the probability of potential outcome for each compliance status.

Step 6: Determine treatment received Z of each subject based on the treatment assignment R and compliance status.

Step 7: Determine observed outcome of each subject based on the potential outcome and treatment received.

$Y = Y^{(1)}Z + Y^{(0)}(1 - Z)$

With this setting, patients in the control arm can access the study treatment means that there are ATs ($a \neq 0$). If($a = 0$), we have $Pr(AT) = 0$, which means that patients in the control arm can not access the study treatment. We also introduce a parameter δ as a measure of severity of confounding. The δ is denoted as the difference between the logit of outcome of never-takers and compliers,

With the same simulated data, we used GSMM, 2SPS and 2SRI methods to estimate the treatment effect, and we compare bias, variance and MSE among those three approaches.

4.3.4 Bezafibrate Data from the GPRD Database

Our data is from the static version of the General Practitioner Research Database (GPRD) that ceased to be updated in 2002. This is an electronic medical record database from 754 general practitioner practice in UK. The database includes registration, demographic information, all prescriptions written by the general practitioners, clinical diagnoses, etc. The GPRD data were originally collected for clinical purposes but the database is widely used for epidemiological research(15-17, 21, 40). Patients in the GPRD database represent the UK population because 98% of people in UK receive all forms of health care through general practitioners and each person has to register to a specific general practice.

We used the data from 1998 through 2002 to create a cohort of fibrate users and classify those patients into bezafibrate users and other fibrate users based on their first fibrate use recorded in the database. To be included in the cohort, patients needed to have at least 12 months up-to-standard (which means the data quality met criteria set by GPRD) GPRD record before the first prescription of fibrate, and needed to have continuous prescription of bezafibrate (study group) or other fibrate (comparison group) for at least 90 days. The continuous prescription means that the gap between the treatments was less than 60 days. We began follow-up from the 91st day of fibrate treatment (T0) to the 30th day after the end of fibrate treatment (T1). If a patient switched treatment group, T1 was defined as the date this patient switched treatment class. Patient with any evidence of diabetes before T0 were excluded.

4.3.5 IV analysis of the Bezafibrate Data

In our IV analysis, the treatment Z is the class of fibrate, Z=1 mean bezafibrate and Z=0 mean other fibrate. The outcome Y is the occurrence of diabetes during the follow-up (between T0 and T1). The occurrence of diabetes is defined by at least two diabetes codes, which can be of clinical diagnoses for diabetes, or of treatment of diabetes.

In this analysis, we used the prior fibrate prescription from the same practice as the IV R. Since we defined the exposure as the initial fibrate treatment, the prior prescription from the same practice was always prescribed to a different patient than the current patient. If a patients was the first one who are prescribed a fibrate, there was no IV for this patients, thus this patient was excluded from the analysis.

For the covariate x, we included calendar year, age, sex, body mass index (BMI) and smoking status were not included as covariate in the analysis because a large portion of patients had missing values of the variables), hypertension, history of myocardial infarction (MI), history of stroke, use of potentially protective drugs (ACE-inhibitors), and common potentially diabetogenic drugs (beta blockers, thiazide diuretics, corticosteroids). We did analyses with and without these covariates and we also tested if the covariates should be included in the models.

Before we did the IV analysis, we tested if the IV is associated with the exposure by calculating correlation of IV and Z, odds ratio, and p value for the null hypothesis that the odds ratio is 1.

We use both approaches of 2-stage logistic regression, 2SPS and 2SRI, as well as

the GSMM to analyze the data and compare bezafibrate users vs other fibrate users regarding rate of diabetes. We also applied traditional logistic regression without IV to the analysis of the same data set. We compared point estimate and standard error from all these 4 approaches of analysis.

4.4 Results

4.4.1 Simulation results

We did the simulations for both frequent outcomes ($\omega_{0C} = 0.3$, and $\omega_{1C} = 0.6$) and rare outcomes ($\omega_{0C} = 0.03$, and $\omega_{1C} = 0.06$) with sample size N=10000. The compliance rate is 0.5, thus the rate of NT is also 0.5. Table 4.1 is of the results of the simulations when there are no always-takers. For the frequent outcome, the bias was very small (less than 0.5%); for simulation of rare outcome, the bias was also very small. It was less than 5% when the δ was less than 1. When the δ is 1 and 1.5, the bias was greater than 5% but less than 10%. When the δ was 2, we didn't have the simulation results because the model didn't converge for some of the simulations. This result shows that the sample size 10000 is not enough for the rare outcome. All simulations have very accurate of 95% CI coverage.

Table 4.2 is of the results of the simulations when we extended the GMSS model to the situation that there are always-takers. In these simulations, the compliance rate was 0.5, and the rate of ATs and NTs were both 0.25. Similar to the non-AT situation, the bias of the treatment effect estimator was also very small (less than

0.3%) for the frequent outcome. For the rare outcome, the bias estimated from the simulation was larger than the frequent outcome setting, but it was still very small (less than 5%). In these simulations, the 95% CI coverage was also close to 95%, although it was slightly higher than 95%.

Table 4.3 compares bias, variance and SME of the three IV approaches, which are 2SPS, 2SRI logistic regression and GSMM. In the simulations with sample size N=3000, there were no ATs, and we set compliance rate as 0.3, 0.5 and 0.7, thus the rate for NTs was 0.7, 0.5, and 0.3 respectively. The δ varied from -2 to 2 in the simulations. For GSMM, the estimated bias was less than 2% in the low compliance situation and less than 1% in high compliance situations. In contrast, the 2SPS and 2SRI can be severely biased. The percentage bias was as high as 53.9% for the 2SPS approach and it was as high as 68.62% for the 2SRI approach. For these two approaches, there was a trend that the bias deceases with the increase of compliance rate (i.e., strength of the IV) and it also decreased with the decrease of confounding. When there is no confounding, the 2SRI was unbiased (the estimated bias is less than 0.5%), which was consistent with our previous report.

Comparing the variance of the three approaches, we can see that the GSMM generally had the smallest variance, and 2SRI had the highest variance. The GSMM also had the smallest MSE among the three approaches. It is difficult to tell which approach has smaller MSE when comparing 2SRI and 2SPS.

4.4.2 IV analysis of bezafibrate data

We first estimated the association of IV and treatment and present results in table 4.4. When the prior prescription from the same practice was bezafibrate, 79.4% of patients actually had bezafibrate prescription, in comparison, when the prior prescription from the same practice was other fibrates, only 60.7% of patients had bezafibrate prescription. The OR for the association of the IV and the treatment received was 2.49 with 95% confidence interval 2.31-2.69 and p value less than 0.001. The correlation between the IV and exposure was 0.1907, which means the prior prescription from the same practice was a weak IV.

Table 4.5 shows the association between exposure and outcome as well as between IV and outcome. The unadjusted odds ratio of actual bezafibrate treatment vs. other fibrate on the outcome of diabetes was 0.67 (95% CI 0.53-0.85) with p value 0.0009, indicating strong association. On the other hand, the association of IV and the outcome was very weak, with odds ratio 0.97 (%95 CI 0.76-1.26), and p value 0.8417. This association is equivalent to intent-to-treat analysis in clinical trial.

In Table 4.6, we compare treatment effect estimated by different approaches, with and without adjustment of covariates. With traditional logistic regression without IV analysis, the log odds ratio was -0.3942 without covariates and the estimated log odds ratio is almost the same when adding covariates (-0.4028). For the IV analysis of 2-stage regression approaches, 2SPS and 2SRI yielded very similar estimates of causal effect. When the covariates were not in the models, 2SPS and 2SRI provide log odds ratio of -0.1391 and -0.1067 respectively; when the covariates were added to

the models, 2SPS and 2SRI yield log odds ratio of -0.2682 and -0.2532 respectively. On the other hand, the GSMM approaches provided log odds ratio -0.1959 when the covariates were not in the model, but the estimated log odds ratio was 0.4040 when the covariates are added to the model. As shown in the same table, the log odds ratio estimated by GSMM had the smallest standard error among the three IV approaches, which was consistent with our simulation results.

Theoretically, the IV analysis can control for all measured and unmeasured confounders without adding covariates in the model, unless the IV is associated with some measured confounders, in which situation, we should add measured confounders as covariates in the models. As the GSMM approach yielded opposite estimates of log odds ratio when the covariates were added into the model, we tested the association of the IV and all covariates to see if we should add any covariates into the model. Table 4.7 shows that only gender and patients' history of ACE inhibitor/angiotensin receptor blocker use was associate with IV, but the p value for the association of gender and the outcome was 0.3008 and 0.7250 in the group with IV being bezafibrate and with IV being other fibrate respectively, and the p value for the association of history of ACE inhibitor/angiotensin receptor blocker use and the outcome was 0.9120 and 0.7305 respectively. With these results, we assumed that no covariates should be included in the GSMM model.

4.5 Discussion

In this research, we developed an R program to implement the GSMM model proposed by Vansteelandt and Geotghebeur (66) when the model is extended from randomized clinical trial that patients assigned to the placebo group can not access the study treatment to observational study that patients in the comparison group as determined by IV can access the study treatment. We then did simulations with both situations when patients in the comparison group can or can not access the study treatment. Our simulation results demonstrated that for both situations, the GSMM gives unbiased estimates of treatment effect. When Vansteelandt and Geotghebeur presented their GSMM model with clinical trial setting, they also did simulations to demonstrate how the model worked. As most investigors do, they simulated data sets according to the model itself in their simulations. In our simulations, we simulated data sets with principle stratification framework. Our unbiased results demonstrate that the GSMM yields unbiased estimates of the complier average causal effect (CACE) on the logit scale, thus our simulation study creates a linkage between the GSMM and the principal stratification framework. It will be very interesting to analytically prove that GSMM is unbiased estimator of odds ratio of CACE. Our simulation results not only show that the GSMM is unbiased estimator of odds ratio of CACE while both 2SPS and 2SRI are biased, but also demonstrate that GSMM has smaller variance and MSE. All these results proved that GSMM is better estimator than 2SPS and 2SRI. However, during our simulations, we found that when the outcome rate is low, GSMM needs much larger sample size to converge than 2SPS

and 2SRI. For instance, when the outcome rate is 0.06 for the treatment group and 0.03 for the placebo group, the GSMM can not converge when the sample size is 5000, but with the same outcome rates, 2SPS and 2SRI can converge even when the sample size is 1000. In our previous study, we have shown that for 2SPS and 2SRI approach, higher compliance rate yields less bias and smaller variance, but it may not be the case for GSMM approach. This is because when compliance rate is high, there will be small number of patients with z=0 in the treatment arm, thus estimates from the association model may have larger variance, or it may be even worse, the model may not converge. The limitation of applying the IV analysis to epidemiology research is that it is difficult to find an IV that meets all assumptions, and violation of those assumptions may lead to invalid inference. Potential variables that can be used as IV in medical research includes physician's prescribing preference (55; 57; 54; 101; 111), clinic or hospital (58), and geographic region (93; 61; 59). Among them, physician's prescribing preference is the mostly widely used IV in pharmacoepidemiology. It was reported that the last prescription of the same physician is a stronger IV (i.e., more closely associated with exposure) than an IV based on all prior prescriptions (57). When we analyzed the bezafibrate data with IV approaches, we found that the previous prescription from the same practice is associated with current prescription with p value less than 0.001, which implies it can be used as an IV. We use is variable instead of prescription from the same physician because in this GPRD database with total number of 754 practices, 217 were missing all prescriber IDs, and only half fibrate prescriptions had prescriber ID. Even when prescriber IDs are available, the prescriber ID may not represent the physician who prescribed the drug, because it

can represent any worker in the practice such as a nurse who entered the prescription into the system. We used the immediate prior prescription from the same practice instead of all prior prescriptions from the same practice as IV because there is report that immediate prior prescription is a stronger IV (57). We assumed that the prior prescription of the same practice is independent of outcome conditional on the treatment, and this variable is independent of all unmeasured confounders. In UK, patients have to register to a practice and they can only see doctors in the practice where they registered, thus they have much less freedom to select physicians than people do in other European countries. People mostly select practices to register based on the geographic area for their convenience. Our IV assumption is valid unless patients select practice based on the health outcome of the patients registered to the practice. In this bezafibrate data analysis, over specifying the GSMM model yielded opposite result, with an odds ratio greater than 1. On the contrary, for both 2SPS and 2SRI approaches, adding all covariates into the model didn't change the result very much. It looks like that GSMM is more sensitive to model selection than 2SPS and 2SRI. This may related to the fact that GSMM needs larger sample size to converge, and when the sample size is too small, the model is not stable. The important result of this analysis is that all three approaches have odd ratios less than 1, which suggests a casual protective effect of bezafibrate against diabetes. However, the estimated protective effect is not statistically significant in this analysis. This may due to the fact that the IV is weak and the rate of the outcome is low, thus the sample size is not large enough. On the other hand, our analysis of logistic regression without IV approaches showed statistically significant protective result that is similar

with result of survival analysis by Flory et al (95). This is an example that there is tradeoff between accuracy and precision with and without IV approaches. A larger sample size is needed for this study with IV analysis.

Table 4.1. Simulation results of GSMM estimator without always-takers.

Outcome Rate w_{0c}	Outcome Rate w_{1c}	δ	Estimated Log OR	Bias	Bias %	Estimated Variance	Width of 95% CI	% Coverage
0.30	0.60	-2.0	1.2544	0.0017	0.1339	0.0050	0.2782	95.00
		-1.5	1.2540	0.0012	0.0982	0.0053	0.2866	95.10
		-1.0	1.2545	0.0017	0.1375	0.0058	0.2981	95.25
		-0.5	1.2558	0.0030	0.2401	0.0064	0.3125	94.70
		0.0	1.2566	0.0038	0.3043	0.0070	0.3280	95.05
		0.5	1.2566	0.0038	0.3055	0.0076	0.3421	94.20
		1.0	1.2549	0.0021	0.1675	0.0081	0.3518	94.65
		1.5	1.2555	0.0027	0.2156	0.0083	0.3567	94.80
		2.0	1.2549	0.0021	0.1710	0.0083	0.3570	94.95
0.03	0.06	-2.0	0.7355	0.0110	1.4904	0.0261	0.6325	95.35
		-1.5	0.7377	0.0131	1.7777	0.0285	0.6601	95.80
		-1.0	0.7414	0.0168	2.2706	0.0325	0.7032	95.85
		-0.5	0.7449	0.0203	2.7279	0.0391	0.7685	95.60
		0.0	0.7484	0.0239	3.1876	0.0500	0.8644	96.05
		0.5	0.7584	0.0338	4.4572	0.0701	1.0094	96.50
		1.0	0.7659	0.0413	5.3949	0.1086	1.2087	95.80
		1.5	0.7828	0.0582	7.4356	0.1892	1.4969	95.20
		2.0 §	0.8147	0.0901	11.060	4.4463	2.2107	95.21

Note: The Sample size is 10000; The true logOR when w0c=0.3 and w1c=0.6 is 1.2528; the true logOR when w0c=0.03 and w1c=0.06 is 0.7246. The compliance rate is 0.5. The bias is defined as the difference of estimated logOR and the true logOR.

§ Based on the 1275 simulations instead of 2000 simulations because the R program stopped when the model didn't converge or the system was computationally singular.

Table 4.2. Simulation results of GSMM estimator without always-takers.

Outcome Rate w_{0c}	Outcome Rate w_{1c}	δ	Estimated Log OR	Bias	Bias %	Estimated Variance	Width of 95% CI	% Coverage
0.30	0.60	-2.0	1.2556	0.0029	0.2290	0.0065	0.3171	95.90
		-1.5	1.2555	0.0028	0.2210	0.0067	0.3199	95.75
		-1.0	1.2552	0.0024	0.1924	0.0068	0.3240	95.80
		-0.5	1.2545	0.0017	0.1362	0.0071	0.3297	95.80
		0.0	1.2543	0.0015	0.1200	0.0074	0.3372	95.75
		0.5	1.2530	0.0002	0.0184	0.0078	0.3459	96.00
		1.0	1.2528	0.0000	0.0003	0.0082	0.3551	96.75
		1.5	1.2513	-0.0014	-0.1129	0.0086	0.3636	96.50
		2.0	1.2496	-0.0032	-0.2521	0.0089	0.3706	96.85
0.03	0.06	-2.0	0.7500	0.0254	3.3876	0.0397	0.7794	97.40
		-1.5	0.7506	0.0260	3.4642	0.0411	0.7928	97.45
		-1.0	0.7512	0.0266	3.5409	0.0434	0.8136	97.05
		-0.5	0.7509	0.0263	3.5038	0.0469	0.8453	97.60
		0.0	0.7538	0.0292	3.8731	0.0529	0.8959	97.35
		0.5	0.7541	0.0295	3.9149	0.0624	0.9693	97.45
		1.0	0.7538	0.0293	3.8807	0.0779	1.0746	97.10
		1.5	0.7521	0.0276	3.6666	0.1015	1.2168	97.25
		2.0	0.7511	0.0265	3.5298	0.1415	1.4122	96.30

Note: The Sample size is 10000; The true logOR when w_{0c}=0.3 and w_{1c}=0.6 is 1.2528; the true logOR when w_{0c}=0.03 and w_{1c}=0.06 is 0.7246.

Table 4.3. Comparing Bias, Variance and MSE of 2SRI, 2SPS and GMSS.

Compliance	δ	2SPS			2SRI			GSMM		
		Bias%	Variance	MSE	Bias%	Variance	MSE	Bias%	Variance	MSE
0.3	-2	53.9183	0.0751	0.5313	-68.6214	0.1389	0.8779	1.2139	0.0348	0.0350
	-1.5	42.7860	0.0703	0.3576	-36.3221	0.1074	0.3144	1.3544	0.0393	0.0396
	-1	29.8485	0.0632	0.2030	-14.0762	0.0833	0.1144	1.4558	0.0444	0.0447
	-0.5	16.6929	0.0593	0.1030	-2.4911	0.0702	0.0711	1.4846	0.0537	0.0540
	0	5.6660	0.0548	0.0598	0.4433	0.0594	0.0594	1.4981	0.0619	0.0623
	0.5	-1.1777	0.0514	0.0516	-1.1879	0.0530	0.0532	1.8688	0.0688	0.0694
	1	-4.1440	0.0542	0.0568	-3.8740	0.0544	0.0568	1.0822	0.0758	0.0759
	1.5	-1.4274	0.0559	0.0562	-0.9504	0.0561	0.0563	1.1560	0.0738	0.0740
	2	5.3301	0.0632	0.0676	10.7841	0.0660	0.0842	1.3621	0.0731	0.0733
0.5	-2	25.1660	0.0246	0.1240	-40.8972	0.0443	0.3068	0.4856	0.0168	0.0168
	-1.5	20.6969	0.0238	0.0910	-22.9951	0.0365	0.1195	0.5133	0.0179	0.0179
	-1	15.0236	0.0228	0.0583	-9.7452	0.0301	0.0450	0.4827	0.0192	0.0193
	-0.5	8.8435	0.0224	0.0347	-1.9902	0.0266	0.0272	0.6257	0.0218	0.0218
	0	3.0573	0.0221	0.0235	0.2886	0.0240	0.0240	0.7406	0.0243	0.0244
	0.5	-1.3746	0.0217	0.0220	-1.1937	0.0224	0.0226	0.5654	0.0265	0.0265
	1	-3.1885	0.0205	0.0221	-2.9974	0.0206	0.0220	0.5185	0.0262	0.0262
	1.5	-2.4763	0.0216	0.0226	-2.1383	0.0218	0.0225	0.3426	0.0271	0.0271
	2	0.4039	0.0233	0.0233	4.0417	0.0245	0.0271	0.4064	0.0276	0.0276
0.7	-2	10.8417	0.0127	0.0311	-22.0355	0.0210	0.0972	0.4427	0.0109	0.0109
	-1.5	9.1438	0.0124	0.0255	-12.6871	0.0175	0.0427	0.4473	0.0111	0.0111
	-1	6.8768	0.0123	0.0197	-5.4469	0.0153	0.0199	0.4520	0.0116	0.0116
	-0.5	4.1589	0.0121	0.0148	-1.0837	0.0138	0.0139	0.4795	0.0122	0.0122
	0	1.4723	0.0118	0.0121	0.2984	0.0126	0.0126	0.5729	0.0127	0.0128
	0.5	-0.7934	0.0119	0.0120	-0.5530	0.0122	0.0123	0.6071	0.0137	0.0137
	1	-2.0983	0.0122	0.0129	-1.9705	0.0123	0.0129	0.5887	0.0146	0.0146
	1.5	-2.1766	0.0124	0.0131	-1.9599	0.0124	0.0130	0.6478	0.0148	0.0149
	2	-1.3957	0.0128	0.0131	0.7437	0.0133	0.0134	0.7022	0.0151	0.0152

Note: N=3000, w_{0c}=0.3, w_{1c}=0.6, without AT

Table 4.4. Correlation of the IV and the exposure

	Exposure to Bezafibrate (%)	OR (95% CI)	P value	Correlation
IV=Bezafibrate	9127 (79.40)	2.49 (2.31-2.68)	<.0001	0.1904
IV=Other fibrates	2648 (60.76)			

Table 4.5. Rate of outcome associated with exposure and IV.

	Bezafibrate	Other Fibrate	OR	P Value
Exposure	209/12043 (1.74)	108/4231 (2.55)	0.67 (0.53-0.85)	0.0009
IV	216/11495 (1.88)	84/4358 (1.93)	0.97 (0.76-1.26)	0.8417

Table 4.6. Comparison of results of causal log OR by different approaches.

Model	Covariate(s)	Treatment Effect Log OR	Standard Error	P value for the treatment effect
Naïve Logistic Regression	No Covariate	-0.3942	0.1199	0.0010
	All in the list	-0.4028	0.1213	0.0009
IV 2SPS	No Covariate	-0.1391	0.6965	0.8417
	All in the list	-0.2682	0.7238	0.7110
IV 2SRI	No Covariate	-0.1067	0.6986	0.8784
	All in the list	-0.2532	0.7241	0.7265
IV GSMM	No Covariate	-0.1959	0.2540	0.4404
	All in the list	0.4040	0.3716	0.2752

Table 4.7. Association of IV and Covariate.

	IV=Other Fibrate N (%)	IV=Bezaofibrate N(%)	P Value
Male	2399(55.05)	6095(53.02)	0.0225
Age 50-59	1427 (32.74)	3814(33.18)	0.6030
Age 50-59	1545 (35.45)	4158(36.17)	0.3989
Age >=70	495(11.36)	1200 (10.44)	0.0945
MI	29 (0.67)	68 (0.59)	0.5943
Stroke	20 (0.46)	38(0.33)	0.2321
History of calcium channel blocker use	1050(24.09)	2715 (23.62)	0.5307
History of thiazide diuretic use	395 (9.06)	1012 (8.80)	0.6074
History of loop diuretic use	220 (5.05)	581 (5.05)	0.9873
History of corticosteroid use	145 (3.33)	379 (3.30)	0.9245
History of beta-blocker use	751 (17.23)	1891 (16.45)	0.2381
History of ACE inhibitor/angiotensin receptor blocker use	243 (5.58)	552 (4.80)	0.0462

Chapter 5

Conclusion

This dissertation provides three distinct contributions to causal inference of binary outcomes with IV analysis. The first contribution is focused on the point estimate of 2SPS and 2SRI logistic regression to evaluate the bias of these approaches. The second contribution is focused on the variance estimate of 2SPS and 2SRI logistic regression, and the third is to evaluate bias of GSMM in the context of principal stratification framework and compare this approach with 2SPS, 2SRI regarding bias and MSE, then apply all three approaches to the data analysis of antidiabetic effect of bezafibrate with the GPRD database. In the first part, we developed closed form expressions for the asymptotic bias of the 2SRI and 2SPS approaches to two-stage logistic regression, and we showed that these analytic results are consistent with the simulation results under different parameter settings. An important contribution of this part is the expression of the conditional distribution of observed outcomes Y given treatment assignment R as a function of the probability of compliance and the conditional distribution of potential outcomes given compliance status. With this contribution,

we can analytically present probability limits and therefore the bias of the estimators of the causal effects of treatment given compliance and treatment status. Further, we provide analytic estimates of bias for a variety of situations. These analytic estimates of bias can help researchers evaluate if the bias is small under specific conditions (e.g. high compliance, and moderate confounding). Hence, our results can be used as a guide for deciding if the 2SRI or 2SPS strategy is appropriate. This method can be potentially applied to the bias analysis of causal inference with other non-linear two-stage regressions, such as regressions of probit models and log linear models. We could provide closed form expressions for the asymptotic bias of the 2SPS approach under the assumption that there are not always-takers as well as that there are always-takers, but with 2SRI approach, we can only provide close form expressions for the asymptotic bias when this is no always-takers, even though we can evaluate it by simulations. This is because the 2SRI is a misspecified model. The true model should include a term of interaction between the treatment receive (Z) and the residual (E) in the second stage model as we have proved in the appendix. It would be interesting to use a new 2SRI model that includes this interaction term, and to analysis bias and variance of this new 2SRI model. We should be able to derive a close form expression for the asymptotic bias of the new 2SRI model when there are always-takers. Another extension of this research is to derive close form expression of asymptotic bias of both 2SPS and 2SRI approaches to the models with covariates. In the IV definition, we only need to verify that the IV is independent with unmeasured confounding. When the IV is associated with measured confounding, we just need to include the variables for measured confounding as covariate in the regressions of both stages. For this

reason, it is very important to evaluate bias of two-stage regression when covariates are included. This is not an easy task, as the bias analysis with close form will be much more complicated when covariates are included. In the second part, we applied the theory of two-step estimation to obtain the adjust variance estimator for both 2SPS and 2SRI approach of IV two stage logistic regression. Our simulation results shows that the estimator we derived for these nonlinear IV analysis provides a good estimate of variance of the causal log odds ratio, in that the estimated variance is consistent with the observed sample variance. With these results, we provide a method to obtain the variance estimate that can apply as an alternative to the bootstrap method in the nonlinear IV analysis and the variance estimator we propose is more accurate than the bootstrap method. Our simulation results indicate that the naive variance estimate without adjustment for two stage regression can be severely biased when the compliance rate is low and the confounding is severe, thus we cannot directly use the variance estimated from the second stage regression to calculate a confidence interval or p value. This is true even for the 2SRI approach, when the causal log odds ratio is the coefficient for the variable of treatment received, instead of the expected value of treatment estimated from the first stage, which is the case for the 2SPS approach. This finding may possibly correct the improper use of 2SRI approach when the variance of log odds ratio is directly estimated from the second stage without adjustment. Based on above results, we should do further research to provide methods of sample size and power calculation, either analytically or by simulations for the IV analysis with binary outcome. For instance, we found that the variance of two-stage logistic regression is very sensitive to the compliance rate, but not very sensitive to

the severity of confounding, thus the sample size and power calculation should more depend on the compliance rate, which can be interpreted as the association of IV and exposure, or strength of IV. In the third part of my dissertation, we developed an R program to implement GSMM approach when patients in the placebo arm have access to the study treatment, which is normally the case in non-randomized observational studies. We then did simulations with the principal stratification setting to test the GSMM method and the results show that the GSMM is unbiased. Our unbiased results not only validated our R program, but also demonstrated that the GSMM yields unbiased estimates of the complier average causal effect (CACE) on the logit scale, thus our simulation study creates a linkage between the GSMM and the principal stratification framework. This result should motivate future research to analytically prove that GSMM is an unbiased estimator of odds ratio of CACE. The simulation results also show that the variance of GSMM estimates is smaller than both 2SRI and 2SPS. The unbiased estimator and smaller variance of GSMM indicate that GSMM has a big advantage over 2SRI and 2SPS. When we applied all three approaches to the data analysis of antidiabetic effect of bezafibrate, they all yield odd ratios less than 1, which indicate a casual protective effect of bezafibrate against diabetes. However, the estimated protective effect is not significant in this analysis with any of the three approaches. This may be due to the fact that the IV is weak and the rate of outcome is low, thus the sample size is not large enough. On the other hand, our analysis of logistical regression without IV approaches shows significant protective result that is similar with result of survival analysis by Flory et al (9). This is an example that there is tradeoff between accuracy and precision

with and without IV approaches. A larger sample size is required for this study with IV analysis. The above result raises an important question about IV analysis: How strong does the association of the IV and treatment or exposure needs to be? With the two-stage linear regression of IV analysis, the correlation of 0.2 between the IV and is considered strong IV, but our simulation results and empirical results of different IV logistic regression for the analysis of binary outcome show that the correlation of 0.2 yield large variances, thus the large confident interval when the rate of outcome is low. More research on the relationship between IV strength and variance, sample size and power is necessary to provide guidance for the application of IV analysis to binary outcomes.

Bibliography

[1] Miettinen OS. Individual matching with multiple contols in the case of all-or-none responses. *Biometrics* 1969; **25**:339–355.

[2] Miettinen OS. Matching and design efficiency in retrospetive studies. *American Journal of Epidemiology* 1970; **91**:111–118.

[3] Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983; **70** (1):41–55.

[4] Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*1984; **79**(387):516–524.

[5] Rosenbaum PR. Conditional Permutation Tests and the Propensity Score in Observational Studies. *Journal of the American Statistical Association*1984; **79**(387):565–574.

[6] VanderWeele TJ, Vansteelandt S, Robins JM. Marginal Structural Models for Sufficient Cause Interactions. *American Journal of Epidemiology*2010; **171**:S44.

[7] Taubman S, Robins J, Mittleman M, Hernan M. Estimating the effects of hypothetical interventions in longitudinal data: Inverse probability weighting of marginal structural models versus the parametric G-formula. *American Journal of Epidemiology* 2008; **167**(11):S47.

[8] Cole SR, Hernan MA, Margolick J, Cohen M, Robins JM, Munoz A. Marginal structural mean models to estimate the effect of highly active antiretroviral therapies on CD4 count. *American Journal of Epidemiology* 2003; **157** (11):213.

[9] Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; **11**(5):561–570.

[10] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11** (5):550–560.

[11] Maclure M, Mittleman MA. Case-crossover designs compared with dynamic follow-up designs. *Epidemiology* 2008; **19** (2):176–178.

[12] Mittleman MA, Maclure M. The study base in case-crossover and other matched designs. *American Journal of Epidemiology* 1996; **143** (11):111.

[13] Mittleman M, Robins J, Maclure M. Statistical-Methods for Analyzing Case-Crossover Studies. *American Journal of Epidemiology* 1993; **138** (8):618–619.

[14] Maclure M. The Case-Crossover Design - A Method for Studying Transient Effects of Brief Exposures on the Risk of Rare Acute Events. *American Journal of Epidemiology* 1990; **132** (4):781–782.

[15] Suissa S. The Case Time-Control Design. *Epidemiology* 1995; **6** (3):248–253.

[16] Farrington CP. Relative Incidence Estimation from Case Series for Vaccine Safety Evaluation. *Biometrics* 1995; **51** (1):228–235.

[17] Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: The self-controlled case series method. *Statistics in Medicine* 2006; **25** (10):1768–1797.

[18] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10** (1):37–48.

[19] Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology* 2001; **12** (3):313–320.

[20] Neyman J. On the application of probability theory to agricultural experiments: essay on principles, section 9. *Translated in: Statistical Science* 1990; **5** (4):465–480.

[21] Terza JV. Parametric Nonlinear Regression with Endogenous Switching. *Econometric Reviews* 2009; **28** (6):555–580.

[22] Rassen JA, Schneeweiss S, Glynn RJ, Mittleman MA, Brookhart MA. Instrumental variable analysis for estimation of absolute and relative treatment effects with binary outcomes. *Pharmacoepidemiology and Drug Safety* 2008; **17**:459.

[23] Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics* 2008; **27** (3):531–543.

[24] Stefanski LA, Buzas JS. Instrumental Variable Estimation in Binary Regression Measurement Error Models. *Journal of the American Statistical Association* 1995; **90** (430):541–550.

[25] Thoresen M, Laake P. Instrumental variable estimation in logistic measurement error models by means of factor scores. *Communications in Statistics-Theory and Methods* 1999; **28** (2):297–313.

[26] Thoresen M, Laake P. A simulation study of measurement error correction methods in logistic regression. *Biometrics* 2000; **56** (3):868–872.

[27] Palmer TM, Burton PR, Thompson JR, Tobin MD. An adjusted instrumental-variable model for Mendelian randomization. *Genetic Epidemiology* 2007; **31** (6):125.

[28] Palmer TM, Thompson JR, Tobin MD, Sheehan NA, Burton PR. Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses. *International Journal of Epidemiology* 2008; **37** (5):1161–1168.

[29] Hirano K. Semiparametric Bayesian inference in autoregressive panel data models. *Econometrica* 2002; **70** (2):781–799.

[30] Goetghebeur E, Molenberghs G, Katz J. Estimating the causal effect of compliance on binary outcome in randomized controlled trials. *Statistics in Medicine* 1998; **17** (3):341–355.

[96] Goetghebeur E, Vansteelandt S. Structural mean models for compliance analysis

in randomized clinical trials and the impact of errors on measures of exposure. *Statistical Methods in Medical Research* 2005; **14** (4):397–415.

[32] Frangakis CE, Rubin DB. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* 1999; **86** (2):365–379.

[33] Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health* 2000; **21**:121–145.

[34] Wooldridge J.M. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press: 2002.

[35] Nagelkerke N, Fidler V, Bernsen R, Borgdorff M. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine* 2000; **19** (14):1849–1864.

[97] Goldberger AS. Structural equation methods in the social sciences . *Econometrica* 1972; **40**(6):979–1001.

[104] Robins JM, Blevins D, Ritter G, Wulfsohn M. G-Estimation of the Effect of Prophylaxis Therapy for Pneumocystis-Carinii Pneumonia on the Survival of Aids Patients. *Epidemiology* 1992; **3** (4):319–336.

[95] Flory JH, Ellenberg S, Szapary PO, Strom BL, Hennessy S. Antidiabetic Action of Bezafibrate in a Large Observational Database. *Diabetes Care*2009; **32** (4):547–551.

[39] Terza J, Basu A, Rathouz P. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics* 2008; **27**(3):531–543.

[40] Bellamy S, Lin J, Have T. An introduction to causal modelling in clinical trials. *Clinical Trials* 2007; **4**(1):58–73.

[41] Greenland S. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* 2000; **29**(4):722–729.

[42] Hernan M, Robins J. Instruments for causal inference - an epidemiologist's dream? *Epidemiology* 2006; **17**(4):360–372.

[43] Angrist J, Imbens G, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996; **91**(434):444–455.

[44] Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002; **58**(1):21–29.

[45] Abadie A. Semiparametric instrumental variable estimation of treatment response models. *Journal of the American Econometrics* 2003; **113**:231–263.

[46] Didelez V, Meng S, Sheehan NA. Assumptions of IV methods for observational epidemiology. *Statistical Science* 2010; **25**(1):22–40.

[47] Sommer A, Zeger S. On estimating efficacy from clinical-trials. *Statistics in Medicine* 1991; **10**(1):45–52.

[48] Frangakis C, Rubin D. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* 1999; **86**(2):365–379.

[49] Tan Z. Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association* 2006; **101**(476):1607–1618.

[50] Small D, Rosenbaum P. War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *Statistics in Medicine* 2006; **25**(12):1981–2007.

[51] Small D, Ten Have T, Joffe M, Cheng J. Random effects logistic models for analyzing efficacy of a longitudinal randomized treatment with non-adherence. *Journal of the American Statistical Association* 2008; **103**(483):924–933.

[52] Sheiner L, Rubin D. Intention-to-treat analysis and the goal of clinical trials. *Clinical Pharmacology and Therapeutics* 1995; **56**(1):6–10.

[101] Korn E, Teeter D, Baumrind S. Using explicit clinician preferences in non-randomized study designs. *Journal of Statistical Planning and Inference* 2001; **96**(1):67–82.

[54] Korn E, Rosenbaum P, Fienberg S, Rubin D. Causal inference through potential outcomes and principal stratification: Application to studies with 'censoring' due to death - comments and rejoiners. *Statistical Science* 2006; **21**(3):310–321.

[55] Brookhart M, Wang P, Solomon D, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* 2006; **17**(3):268–275.

[111] Wang P, Schneeweiss S, Avorn J, Fischer M, Mogun H, Solomon D, Brookhart M. Risk of death in elderly users of conventional vs. atypical antipsychotic medications. *New England Journal of Medicine* 2005; **353**(22):2335–2341.

[57] Hennessy S, Leonard C, Palumbo C, Shi X, Ten Have T. Instantaneous preference was a stronger instrumental variable than 3-and 6-month prescribing preference for nsaids. *Journal of Clinical Epidemiology* 2008; **61**(12):1285–1288.

[58] Johnston S. Combining ecological and individual variables to reduce confounding by indication: Case study - subarachnoid hemorrhage treatment. *Journal of Clinical Epidemiology* 2000; **53**(12):1236–1241.

[59] Wen S, Kramer M. Uses of ecologic studies in the assessment of intended treatment effects. *Journal of Clinical Epidemiology* 1999; **52**(1):7–12.

[93] Brooks J, Chrischilles E, Scott S, Chen-Hardee S. Was breast conserving surgery underutilized for early stage breast cancer? instrumental variables evidence for stage ii patients from iowa. *Health Services Research* 2003; **38**(6):1385–1402.

[61] Stukel T, Fisher E, Wennberg D, Alter D, Gottlieb D, Vermeulen M. Analysis of observational studies in the presence of treatment selection bias - effects of invasive cardiac management on ami survival using propensity score and instru-

mental variable methods. *Jama-Journal of the American Medical Association* 2007; **297**(3):278–285.

[62] Joffe M, Brensinger C. Weighting in instrumental variables and g-estimation. *Statistics in Medicine* 2003; **22**(8):1285–1303.

[63] Hogan J, Lancaster T. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research* 2004; **13**(1):17–48.

[64] Hirano K, Imbens W, Rubin B, Zhou X. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* 2000; **1**(1):69–88.

[65] Goetghebeur E, Molenberghs G. Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *Journal of the American Statistical Association* 1996; **91**(435):928–934.

[66] Vansteelandt S, Goetghebeur E. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 2003; **65**(4):817–835.

[67] Ten Have T, Joffe M, Cary M. Causal logistic models for non-compliance under randomized treatment with univariate binary response. *Statistics in Medicine* 2003; **22**(8):1255–1283.

[68] Robins J, Rotnitzky A. Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika* 2004; **91**(4):763–783.

[69] Rassen J, Schneeweiss S, Glynn R, Mittleman M, Brookhart M. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. *American Journal of Epidemiology* 2009; **169**(3):273–284.

[70] Nagelkerke N. Estimating treatment effects in randomized clinical trials in the presence of non-compliance (vol 19, pg 1849, 2000). *Statistics in Medicine* 2001; **20**(6):982–982.

[71] Rubin, D.B. Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*1978; **6**:34–58.

[105] Rubin, D.B. Statistics and Causal Inference - Which Ifs Have Causal Answers. *Journal of the American Statistical Association* 1989;**81**(396):961–962.

[73] Lin JY, Ten Have T, Elliott MR. Longitudinal nested compliance class model in the presence of time-varying noncompliance.*Journal of the American Statistical Association* 2008; **103**:462–473.

[74] Newey W, Mcfadden D. *Large Sample Estimation and Hypothesis Tesing*, chap. 36. Elsevier B.V: Amsterdam, Norm Holland, 1994; 2111–2245.

[75] Wooldridge J. *M-Estimation*, chap. 12. The MIT Press: Cambridge, Massachusetts, London, England, 2002; 341–384.

[76] Nishii R. Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivarite Analysis* 1988; **27**: 392–403.

[77] Vansteelandt S, Goetghebeur E. Using potential outcomes as predictors of treatment activity via strong structural mean models. *Statistica Sinica* 2004; **14**(3): 907–925.

[78] Ten Have T, Joffe M, Lynch K, Brown G, Maisto S, Beck A. Causal Mediation Analyses with Rank Preserving Models. *Biometrics* 2007; **63**: 926–924.

[79] Cai B, Small DS, Ten Have T. Bias of Causal Inference for the Odds Ratio Using Two Stage Instrumental Variable Methods. *Submitted to Statistics in Medicine Biometrics* 2010.

[80] Bond SJ, White IR, Walker AS. Instrumental variables and interactions in the causal analysis of a complex clinical trial. *Statistics in Medicine* 2007; **26**(7): 1473–1496.

[81] Bloom HS. Accounting for No-Shows in Experimental Evaluation Designs. *Evaluation Review* 1984; **8**(2): 225–246.

[82] Permutt T, Hebel JR. Simultaneous-Equation Estimation in A Clinical-Trial of the Effect of Smoking on Birth-Weight. *Biometrics 1989*; **45**(2): 619–622.

[83] Basu A, Heckman JJ, Navarro-Lozano S, Urzua S. Use of instrumental variables in the presence of heterogeneity and self-selection: An application to treatments of breast cancer patients. *Health Economics* 2007; **16**(11): 1133–1157.

[84] Johnston KM, Gustafson P, Levy AR, Grootendorst P. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured

confounding with applications to epidemiological research. *Statistics in Medicine* 2008; **27**(9): 1539–1556.

[85] Terza JV, Bradford WD, Dismuke CE. The use of linear instrumental variables methods in health services research and health economics: A cautionary note. *Health Services Research* 2008; **43**(3): 1102–1120.

[86] Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* 2007; **16**(4): 309–330.

[87] Murphy KM, Topel RH. Estimation and inference in two-step econometric models (Reprinted). *Journal of Business & Economic Statistics* 2002; **20**(1): 88–97.

[88] Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* 2006; **17**(3): 268–275.

[89] Liang KY, Zeger SL. Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika* 1986; **73**(1): 13-22.

[90] Zeger SL, Liang KY, Albert PS. Models for Longitudinal Data - A Generalized Estimating Equation Approach. *Biometrics* 1988; **44**(4): 1049-1060.

[91] Leiderman L. Macroeconometric Testing of the Rational-Expectations and Structural Neutrality Hypotheses for the United-States. *Journal of Monetary Economics* 1980; **6**(1): 69-82.

[92] Moreira MJ, Porter JR, Suarez GA. Bootstrap validity for the score test when instruments may be weak. *Journal of Econometrics* 2009; **149**(1): 52–64.

[93] Brooks JM, Chrischilles EA, Scott SDet al. Was breast conserving surgery underutilized for early stage breast cancer? Instrumental variables evidence for stage II patients from Iowa. *Health Services Research* 2003; **38**(6): 1385–402

[94] Fernandes-Santos C, Carneiro RE, Mendonca LDet al. Pan-PPAR agonist beneficial effects in overweight mice fed a high-fat high-sucrose diet. *Nutrition* 2009; **25**(7-8): 818–27

[95] Flory JH, Ellenberg S, Szapary POet al. Antidiabetic Action of Bezafibrate in a Large Observational Database. *Diabetes Care* 2009; **32**(4): 547–51

[96] Goetghebeur E, Vansteelandt S. Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Statistical Methods in Medical Research* 2005; **14**(4): 397–415

[97] Goldberger AS. Structural equation methods in the social sciences . *Econometrica* 1972; **40**(6): 979–1001

[98] Herrett E, Thomas SL, Schoonen WMet al. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *British Journal of Clinical Pharmacology* 2010; **69**(1): 4–14

[99] Jick SS, Kaye JA, Vasilakis-Scaramozza Cet al. Validity of the General Practice Research Database. *Pharmacotherapy* 2003; **23**(5): 686–9

[100] Johansson S, Wallander MA, de Abajo FJet al. Prospective Drug Safety Monitoring Using the UK Primary-Care General Practice Research Database Theoretical Framework, Feasibility Analysis and Extrapolation to Future Scenarios. *Drug Safety* 2010; **33**(3): 223–32

[101] Korn EL, Teeter DM, Baumrind S. Using explicit clinician preferences in nonrandomized study designs. *Journal of Statistical Planning and Inference* 2001; **96**(1): 67–82

[102] Lewis JD, Brensinger C, Bilker WBet al. Validity and completeness of the General Practice Research Database for studies of inflammatory bowel disease. *Pharmacoepidemiology and Drug Safety* 2002; **11**(3): 211–8

[103] Nagelkerke NJD, Fidler V, Buwalda M. Instrumental Variables in the Evaluation of Diagnostic-Test Procedures When the True Disease State Is Unknown. *Statistics in Medicine* 1988; **7**(7): 739–44

[104] Robins JM, Blevins D, Ritter Get al. G-Estimation of the Effect of Prophylaxis Therapy for Pneumocystis-Carinii Pneumonia on the Survival of Aids Patients. *Epidemiology* 1992; **3**(4): 319–36

[105] Rubin DB. Statistics and Causal Inference - Which Ifs Have Causal Answers. *Journal of the American Statistical Association* 1986; **81**(396): 961–2

[106] Tenenbaum A, Fisman EZ, Boyko Vet al. Attenuation of progression of insulin resistance in patients with coronary artery disease by bezafibrate. *Archives of Internal Medicine* 2006; **166**(7): 737–41

[107] Tenenbaum A, Fisman EZ, Boyko Vet al. Bezafibrate averts progression of insulin resistance in diabetic patients with coronary artery disease. *European Heart Journal* 2006; **27**: 532

[108] Tenenbaum A, Motro M, Fisman EZet al. Effect of bezafibrate on incidence of type 2 diabetes mellitus in obese patients. *European Heart Journal* 2005; **26**(19): 2032–8

[109] Tenenbaum A, Motro M, Fisman EZet al. Peroxisome proliferator-activated receptor ligand bezafibrate for prevention of type 2 diabetes mellitus in patients with coronary artery disease. *Circulation* 2004; **109**(18): 2197–202

[110] VanderWeele TJ. Marginal Structural Models for the Estimation of Direct and Indirect Effects. *Epidemiology* 2009; **20**(1): 18–26

[111] Wang PS, Schneeweiss S, Avorn Jet al. Risk of death in elderly users of conventional vs. atypical antipsychotic medications. *New England Journal of Medicine* 2005; **353**(22): 2335–41

[112] Wood L, Martinez C. The general, practice research database - Role in pharmacovigilance. *Drug Safety* 2004; **27**(12): 871–81