CAUSAL EFFECT ESTIMATION IN RANDOMIZED CONTROLLED TRIALS WITH IMPERFECT COMPLIANCE

by

Lingyun Lyu

BA, MS China Pharmaceutical University, 2008, 2011

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Lingyun Lyu

It was defended on

April 18, 2018
and approved by

Thesis Advisor:
Jonathan G. Yabes, PhD
Assistant Professor
Departments of Medicine and Biostatistics
School of Medicine and Graduate School of Public Health
University of Pittsburgh

Committee Member:
Bruce L. Jacobs, MD, MPH
Assistant Professor
Department of Urology
School of Medicine
University of Pittsburgh

Committee Member:
Ada O. Youk, PhD
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Jonathan G. Yabes, PhD

CAUSAL EFFECT ESTIMATION IN RANDOMIZED CONTROLLED TRIALS WITH IMPERFECT COMPLIANCE

Lingyun Lyu, MS

University of Pittsburgh, 2018

ABSTRACT

Randomized controlled trials (RCT) are widely considered as the gold standard in generating evidence about the efficacy and safety of an experimental treatment. In practice, however, RCTs often suffer from non-compliance to the assigned treatment threatening the validity of the study results. Intent-to-treat (ITT) has been widely adopted as the standard analyses for such trials. However, under imperfect compliance, ITT validly estimates the treatment effectiveness instead of the treatment effect as-received (efficacy). Under the potential outcomes framework and certain assumptions, the treatment effect as-received may be represented by the Complier Average Causal Effect (CACE), the average treatment effects in the subgroups of compliers. Common methods used to estimate the CACE are As-Treated and Per-Protocol, both of which may introduce confounded comparisons between treatment arms due to the inherent differences between compliers and non-compliers. To provide valid estimates of CACE, causal inference methods such as propensity score (PS) and instrumental variables (IV)-based approaches have been proposed in the literature. As long as an instrument exists, IV-based methods could provide inferences that are less model dependent. They do not necessarily require adjusting for covariates and avoids model selection and specification issues that PS-based methods face for the propensity-to-comply model. Due to random allocation, the randomization assignment often meets the assumptions imposed by an instrument and is widely accepted as a valid instrument in many situations. The most common

IV-based estimation method is 2-Stage-Least-Squares (2SLS). For binary outcome and binary treatment groups, estimating risk ratios or odds ratios have been the subject of many studies in the literature. When interest lies in estimating the risk difference (RD) as the CACE, a linear probability model in the second stage is commonly used. However, there is lack of consensus about what is the most suitable in the first stage where the observed treatment received is regressed to the treatment assignment (instrument).

The goal of this study is to empirically investigate the different IV-based approaches to estimate the risk difference as CACE in RCTs with binary outcome and binary treatment group. We compared the performance of these methods with respect to bias, efficiency, and power and compare these to PP as the standard approach to estimate CACE. We also examined how their performance is affected by varying levels of compliance, effect size, sample size. In addition, we evaluated their statistical properties when measured confounders exist.

We found that all the IV-based methods generally provide valid and very similar estimates, efficiency and power in the setting where there are no measured confounders, while the PP shows large bias in the presence of unmeasured confounders. However, when we can account for measured confounders, a 3-stage approach may provide more efficient estimates and yield higher power. As the compliance probability goes to 1 or as the sample size increases, the differences between the different IV-based methods become negligible.

Public health significance: Results of RCTs are commonly used to implement policies or recommend guidelines to improve public health and patient care. Non-compliance however is common in RCTs and threatens the validity of its results. This study compares different strategies in providing correct estimates of treatment effect under imperfect compliance. This is critical in assessing the utility of an experimental treatment for adoption in clinical practice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Notation	Meaning
$Y_i(1)$	Potential outcome for individual *i* if exposed to the treatment
$Y_i(0)$	Potential outcome for individual *i* if exposed to the control
$E\big(Y_i(t)\big)$	Average potential outcome among individuals exposed to t
T_i	Randomization assignment indicator for individual i $(0 = Control, 1 = Treatment)$
$R_i(T_i)$	Treatment received indicator for individual *i* when randomized to treatment T_i $(0 = Control, 1 = Treatment)$
ACE	Average causal effect
S	Compliance strata
C	Compliers
NT	Never-takers
AT	Always-takers
D	Defiers
NC	Non-compliers
U	Unmeasured confounder
π_T	The probability to be assigned to the treatment group
γ_1	Risk difference
2SLS	Two-stage least squares
LG-LM	Two stage logistic/linear approach
3SLS	3-stage model approach

PREFACE

Acknowledgement

To my thesis advisor Dr. Jonathan G. Yabes, for your excellent guidance throughout this project. Your continuous patience, encouragement, and confidence in me are amazing. You have set up a role model for my future career with your insights in research, attitude for work and kindness to others.

To Dr. Ada O. Youk, for being my thesis committee member and your always gracious support and help to my academic and professional development throughout my study at Pitt.

To Dr. Bruce L. Jacobs, for being my thesis committee member, and giving me invaluable comments and suggestions to the thesis.

To my parents and brother, for your unconditional love and support. I know you are always there, so I have the courage to choose the life I desired!

To my husband, Jian, for your many years continued and unfailing love and support. I appreciate I can be myself with you!

1.0 INTRODUCTION

A randomize d clinical trial (RCT) is widely considered as the “gold standard” to evaluate the efficacy and safety of a treatment. The randomization ensures that treatment groups are comparable on both measured and unmeasured baseline variables which leads to unconfounded comparison between treatment arms. Moreover, a variety of rigorous guidelines ensure the validity of RCTs. For example, International Commission on Harmonization (ICH) guidelines provide a series of complete guidance for designing, conducting, recording, and reporting RCTs. These guidelines ensure that safe, effective, and high-quality medicines are developed and registered.

In practice, however, RCTs often suffer from non-compliance in which subjects do not receive the assigned treatment as intended. Restricting the analyses to compliers may seem logical in estimating treatment efficacy as-received. However, this potentially introduces non-random selection bias which consequently undermine the validity of results. Thus, the occurrence of non-compliance makes the causal inference hard to draw. For instance, an RCT conducted by Perkin MR, et al.[1] aiming to investigate the introduction of allergenic foods into the diet of breast-fed children was severely contaminated by noncompliance (only 32% compliance in the intervention group). When all subjects were analyzed as randomized, the results showed no difference between groups. However, when analysis was restricted to subjects who adhered to their assigned treatment, results showed a significant lower frequency of food allergy in the intervention arm versus the standard arm. The divergent conclusions from these two analyses made the interpretation of the study results difficult. Therefore, obtaining an unbiased estimate of treatment efficacy from RCTs with noncompliance is a critical problem.

A number of potential approaches have been proposed. They can be categorized into two groups: intention-to-treat (ITT) and non-ITT approaches.

The principle of ITT analysis is that all participants should be analyzed in the group to which they had been assigned, irrespective of the treatment actually received. The resulting inference provides an estimate that reflects the effectiveness of the treatment, i.e., treatment effect in the real world setting. This can be important information to policy makers and health planners, but patients and clinicians generally want to know the treatment efficacy, i.e., treatment effect as-received [2]. Under perfect compliance, treatment effectiveness and efficacy are equivalent.

Several Non-ITT approaches have been used to estimate the treatment efficacy through adjustments for non-adherence. Two common Non-ITT methods are the As-Treated and the Per-Protocol (PP). Both approaches aim to estimate the effect of treatment when actually received (as-received treatment effect). The As-Treated analysis compares observed outcomes according to actual treatment received, while the PP analysis compares outcomes only in those subjects who were to compliant. However, because of inherent differences between compliers and non-compliers, they tend to give biased estimates of treatment efficacy.

Since none of the ITT, PP, or As-Treated can provide unbiased estimate in the presence of noncompliance, the possibility of estimating treatment effects only for compliers has been explored [3]. Angrist et al. [4] demonstrated that it is possible to estimate average causal treatment effect in compliers (CACE), given certain assumptions.

A number of approaches have been developed to estimate CACE, such as instrumental variable (IV) based methods and propensity score (PS) based methods. Instrumental variable is an analytical technique that uses a variable associated with the factor under study but not directly associated with the outcome variable or any potential confounders [5]. IV analysis evaluate how

the instrument predicts the exposure and the outcome, then uses that information to understand how the exposure predicts the outcome [5]. Propensity score approach is an alternative, which is based on Frangakis and Rubin's [6] principal stratification. The rationale is to fit a propensity score model based on one treatment condition in which the principal stratum membership can be identified, then apply this propensity score model to predict the principal stratum membership for individuals in the other treatment condition. After identifying individuals' principal stratum membership in two arms, the treatment effects could be estimated in each stratum[7].

The IV techniques are appealing partly because they are easy to implement. Unlike the PS based method, the IV methods do not necessarily need to adjust for covariates, thus there is no model selection issue. A commonly acknowledged difficulty however is that valid instruments are uncommon and hard to find. In well-designed RCTs with noncompliance, this is often not an issue since the (randomized) treatment assignment often meets the assumptions of an instrument. With this well-established instrument, IV based methods have been widely applied in medical and epidemiological research[8-10].

There has been extensive research[8-11] regarding the application of instrumental variable in clinical and epidemiology. However, most of them focused on continuous outcomes and continuous exposures. Though there are several reports for binary outcomes[12-15], most of them investigated how to estimate risk ratio or odds ratio. Little attention has been paid to the estimation of risk difference when both the exposure and outcome are binary. In RCTs, the exposure is typically binary (treatment versus control), and the outcome many times is binary, particularly when with regards to safety. Compared with risk ratio (RR) and odds ratio (OR), risk difference (RD) is more interpretable and straightforward. It is also known as Absolute Risk Reduction

(ARR) and its inverse is equal to the Number-Needed-to-Treat (NTT). In noninferiority studies, RD can often be more easily interpreted in a clinically meaningful way than RR [16].

There are several IV-based methods to estimate RD. Two-stage least square (2SLS) is the most commonly used approach. In the 1st stage, fitted values of the exposure variable are derived from a linear model against the instrument. In the 2nd stage, rather than use the exposure variable itself, the fitted values from the 1st stage are used in a linear model for the outcome to estimate the treatment effect. This method is well established in the setting of continuous outcomes and continuous exposures in the non-randomized data setting. To estimate RD when the outcome is binary, a linear probability model has been suggested for the 2nd stage[17]. Nevertheless when the exposure variable is binary, various models have been proposed for the first stage including a linear model, logistic-model, and a two-step logistic-linear model[18]. However, there is lack of consensus as to which method is the best, particularly in RCTs with non-compliance.

The goal of this study is to empirically investigate the different IV-based approaches to estimate the risk difference as CACE in RCTs with binary outcome and binary treatment group. In the next chapter, we review the literature regarding treatment effect estimation methods in RCTs with noncompliance, mainly focus on IV-based methods. In chapter 3, we describe our simulation studies to empirically investigate the statistical performance of different IV-based approaches in estimating CACE. Lastly, we summarize and discuss the findings, limitations, and future directions.

2.0 LITERATURE REVIEW

2.1 POTENTIAL OUTCOMES FRAMEWORK

In the previous chapter, several potential approaches were mentioned to make causal inference in RCTs with noncompliance. In order to understand and compare these different approaches, it is helpful to think in the context of the Neyman-Rubin potential outcomes framework[4, 19, 20] and discuss Frangakis and Rubin's idea of principal stratification[6].

The potential outcomes framework makes the causal effect inferences more straightforward and practical. This framework argues that there are two potential outcomes for an individual i, one is the value of the outcome the individual would experience if exposed to the treatment ($Y_i(1)$), the other is the value of the outcome the individual would experience if exposed to the control ($Y_i(0)$). The treatment effect for individual i can be expressed as the difference between two potential outcomes, $\big(Y_i(1) - Y_i(0)\big)$ [21]. The average causal effect (ACE) or average treatment effect (ATE) is $E\big(Y_i(1) - Y_i(0)\big) = E\big(Y_i(1)\big) - E\big(Y_i(0)\big)$. This framework is based on one assumption, that is, the treatment assignment of one participant does not influence the outcomes of other participants. This is known as "stable unit-treatment value" assumption (SUTVA). If this assumption does not hold, the potential outcomes for an individual would not be just two, then causal inference is impossible.

However in reality, it is only possible to observe one of the two potential outcomes for any given individual. We do not observe what would have happened if the individual had been randomly assigned to the other arm. Fortunately, randomization allows the causal effect to be estimated relatively easy. When the treatment assignment mechanism is random, the treatment

assignment is independent of outcomes. Also, the randomization ensures the construction of two comparable groups with respect to baseline variables whether measured or unmeasured. The observed average outcome in the experimental treatment group is taken as the estimate of the average outcome in all subjects had everyone received the experimental treatment. Similarly, the observed average outcome in the control group is taken as the estimate of the average outcome in all subjects had everyone received the control. The difference of the two estimates is the unbiased estimate of the ACE. The potential outcome framework and the randomization assignment mechanism provide a useful basis to estimate ACE.

2.2 COMPLIER AVERAGE CAUSAL EFFECT

In the real world, the purpose of randomization is often undermined due to imperfect compliance. Non-adherence happens when patients fail to follow the treatment they were assigned to, such as taking the comparison treatment instead of the assigned treatment or taking the assigned treatment but not according to the study protocol. Non-compliance potentially introduces non-random selection bias to RCTs. The most commonly used statistical methods ITT, PP, and AT often generate biased estimate for treatment efficacy. Angrist et al. [4] proposed that under certain assumptions, the ACE in the subgroup of compliers or CACE represents the treatment efficacy. To better understand and compare these different approaches, it is necessary to first understand Frangakis and Rubin's idea of principal stratification[6].

Frangakis and Rubin[6] proposed the concept of principal stratification, which classify participants in RCT with noncompliance as one of four types, compliers (C), never-takers (NT), always-takers (AT), and defiers (D) (Figure 1). Let T_i be the indicator of randomized treatment

assignment, $T_i = 1$ if the i^{th} subject is randomized to the experimental treatment group, and $T_i = 0$ if randomized to the control group, for $i = 1,2, \ldots, N$. Compliance status is assumed to be all-or-nothing. Let R_i denote the indicator of treatment received if randomized to T_i so that $R_i(T_i) = 1$ if the experimental treatment is received and 0 if the control is received. Compliers are people who would receive whatever treatment is assigned $[R_i(1) = 1, R_i(0) = 0]$; never-takers would only receive the control regardless of the treatment assignment $[R_i(1) = R_i(0) = 0]$; always-takers would only receive the experimental treatment $[R_i(1) = R_i(0) = 1]$; and defiers would take the treatment opposite to the assigned treatment $[R_i(1) = 0, R_i(0) = 1]$. The latent classes are determined prior to randomization, they can be viewed as pre-treatment covariates having the same distribution in the two treatment arms due to randomization assignment mechanism. When we assume that there are no defiers (monotonicity), Figure 1 shows the principal strata in both arms. Since each class of patients are equally distributed across treatment arms, the proportion of AT can be estimated from the observed proportion who received the experimental treatment in the control group. Similarly, the proportion of NT can be estimated in the treatment group. Then, the proportion of compliers can be estimated from the difference in the proportions in each arm who received the assigned treatment, i.e., $P(C) = P(C \cup AT) - P(AT) = P(C \cup NT) - P(NT)$.

For simplicity, we assume a common RCT design in which participants in the control group do not have access to the experimental treatment, i.e. $R_i(0) = 0$ for all i. Thus, always takers do not exist. In this context, the latent classes are reduced to two strata based on binary T and binary R. These two classes are compliers (C) $[R_i(1) = 1, R_i(0) = 0]$ and never takers (NT) $[R_i(1) = 0, R_i(0) = 0]$ which we would refer to as non-compliers (NC). Figure 2 shows the principal strata in each arm, where A and A' denote the average outcome for compliers in the treatment group and the control group respectively; likewise B and B' denote the average outcome for non-compliers

in the treatment group and the control group respectively. Note that A and B are both observable in the treatment arm, but only the combined $(A' + B')$ is observable in the control arm. The principal stratification provides a useful tool to estimate causal effect of treatment in the subgroup compliers, namely the CACE. With the knowledge of principal stratification, it is easier to understand the advantage and disadvantage of different approaches to make casual effect in the setting of RCT with noncompliance.

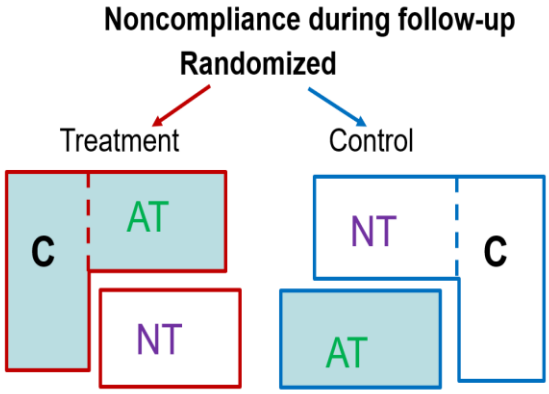

Figure 1. Principal strata in RCTs with noncompliance

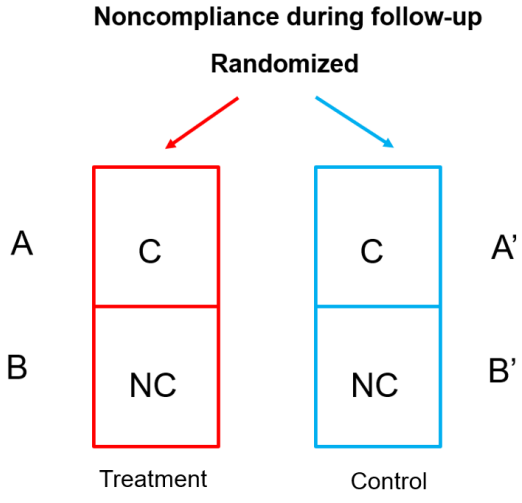

Figure 2. Simplified principal strata in RCTs with noncompliance

One potential solution to estimate average treatment effect is ITT. The ITT principle has long been mandated by the Food and Drug Administration (FDA) as the primary analysis strategy for industry clinical trials[22]. It aims to estimate the effect of treatment as assigned, including all randomized patients in the groups to which they were randomly assigned, regardless of whether or not the patient received the allocated treatment[23] [$(\widehat{A+B})$ - $(\widehat{A'+B'})$]. The advantage of ITT analysis is that it maintains the baseline comparability achieved by the randomizing process. The resulting inference provides an estimate that reflects the effectiveness of the treatment in terms of the public health benefits of administering the treatment in the community, where non-compliance is inevitable in actual clinical practice. Nevertheless, ITT method usually does not generate a valid estimate of the treatment effect on individual patients (the effect of treatment as delivered or as received), especially when the non-compliance rate is high. It usually leads to conservatively biased estimate of as received treatment efficacy, because of dilution due to noncompliance. This could be acceptable for a superiority trial[24] as the bias is towards the null. However, it may be problematic in a non-inferiority trial, since the non-inferiority conclusion may have resulted from dilution of treatment difference due to noncompliance rather than the actual therapeutic similarities between experimental treatment and standard therapy.

Several Non-ITT approaches have been used to estimate the treatment effect through adjustments for non-adherence. Two common non-ITT methods are the As-Treated analysis and the PP analysis. Both approaches aim to estimate the effect of treatment when actually received (as-received treatment effect). The As-Treated analysis compares observed outcomes according to treatment received $E(A) - E(B + A' + B')$, while the PP analysis compares outcomes only in those subjects who were observed to be compliant $E(A) - E(A' + B')$. Both methods however are susceptible to selection bias. Both analyses involve the comparison of two treatment groups with

potentially different underlying patient characteristics[25]. When noncompliance is random and independent of outcomes (ignorable), both solutions could generate valid treatment effect estimate. However, noncompliance cannot usually be assumed to be random. It may be related to many factors such as adverse events and prognosis. These two methods are invalid to estimate treatment efficacy.

Unbiased estimate of the CACE $(A - A')$ is $\left(\hat{A} - \hat{A}'\right)$. However, $\hat{A}'$ cannot be directly calculated from the observed data. Angrist et al. [4] demonstrated that under certain assumptions, the observed data can be used to estimate the CACE.

2.3 INSTRUMENTAL VARIABLE

One way to obtain valid estimates of CACE is through the use of instrumental variables (IV) [4]. IVs are designed to deal with problems of endogeneity (i.e. the explanatory predictor is correlated with the error term.) by isolating the variability in the predictor that is causally related to the outcome[4, 5, 12, 26]. IV has been widely used in RCT with noncompliance[27] or observational studies[12, 28] to estimate the causal effect of treatment or exposure to risk factor when there is unmeasured confounder. In the context of randomized trials, the IV approach can yield unbiased treatment effect estimation in the presence of unmeasured confounders, given certain assumptions are met. This method estimates the CACE in a way that preserve the balance in patient characteristics from randomization[25].

A key component of the IV approach is to identify the instrument. A valid instrument must meet three assumptions [12]. Figure 3 shows a causal diagram that describes the following assumptions that IV must meet [4]:

(1) the IV is associated with the treatment;

(2) the IV is independent of unmeasured confounders;

(3) the IV affects the outcome only through the treatment.

In the case of randomized clinical trials with noncompliance, the randomized treatment assignment satisfies these three assumptions. First, the treatment received (R) is affected but not fully determined by the treatment assignment (T) (assumption 1); Second, the treatment assignment is randomized, so the treatment assignment is independent of unmeasured confounders (U)(assumption 2); Furthermore, because patients and outcome assessors are usually blinded to the randomized treatment assignment, one can assume that it would not have an effect on the outcome (Y) beyond that due to the treatment actually received (assumption 3). Therefore, treatment assignment can generally be used as the instrument to account for unmeasured confounders in randomized clinical trials with non-compliance.

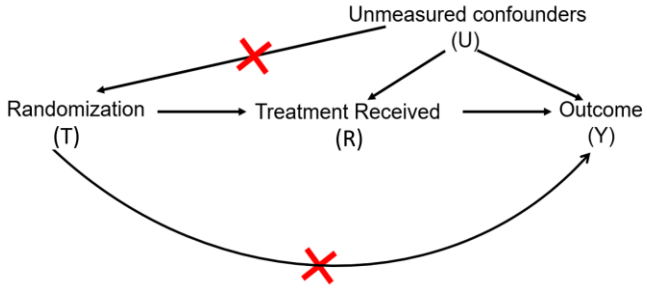

Figure 3. Graphical representation of a trial with noncompliance, with randomization as an instrumental variable for the treatment received

In addition to the identification of instrument, the following assumptions[3] are necessary in order to obtain an unbiased estimate of CACE via the IV-based method.

Assumption 1 (randomization): Treatment assignment is random. Randomization supports the second assumption of IV (the IV is independent of unmeasured confounders). In RCT, treatment assignment is independent of any measured or unmeasured confounders.

Assumption 2 (stable unit treatment value, SUTVA): Potential outcomes for each person are unrelated to the treatment status of other individuals. This assumption allows the application of potential outcome framework, which makes the causal effect estimate straightforward and practical.

Assumption 3 (exclusion restriction): For compliers and noncompliers, the distributions of the potential outcomes are independent of the treatment assignment, that is, the potential outcome depends on treatment assignments only through the actual treatment received. This assumption is equivalent to the third assumption of IV (the IV affects the outcome only through the treatment).

Assumption 4 (monotonicity): There are no defiers. This assumption requires that for every subject who chooses not to take the experimental treatment when randomized to it, he or she will not try to obtain the experimental treatment if randomized to the comparison group.[29] In randomized clinical trials, especially blinded RCT, it is reasonable to assume that there is no defiers.

Assumption 5 (non-zero average causal effect of instrument on treatment): The average causal effect of instrument on treatment is not equal to zero. It means that there is a causal pathway from treatment assignments to treatment received. This assumption is identical to the first assumption of IV (the IV is associated with the treatment).

There has been extensive research on the theory and application of instrumental variables in clinical and epidemiologic studies[8-11]. Most of them applied the IV method in the setting of continuous outcomes and continuous exposures.

When both exposure and outcome are continuous, the most commonly used technique for instrumental variable analysis is the 2SLS method[18]. This method splits the estimation into two stages. In the first stage, the treatment variable is regressed on the instrument and the predicted treatment is obtained from the fitted model; the second stage fits the outcome regression, replacing the treatment variable with their predicted values obtained from the first stage. The mathematical proof of 2SLS is as the following:

Let Y denote the outcome, R is an endogenous exposure, ε is the error term. A simple linear model is:

$$Y_i = \gamma_0 + \gamma_1 R_i + \varepsilon_i \quad (1),$$

When R is correlated with the error term ($Cov(R, \varepsilon) \neq 0$), the estimated $\hat{\gamma}_1$ is biased,

$$\widehat{\gamma_1} = \frac{Cov(R,Y)}{Var(R)} = \gamma_1 + \frac{Cov(R,\varepsilon)}{Var(R)}.$$

Let T denote the instrumental. T is supposed to be linearly related to R, and independent of the unmeasured confounder, that is, $Cov(T, \varepsilon) = 0$, $Cov(T, R) \neq 0$.

$$\begin{aligned} Cov(T, Y) &= Cov(T, \gamma_0 + \gamma_1 R + \varepsilon) \\ &= \gamma_1 Cov(T, R) + Cov(T, \varepsilon) \\ &= \gamma_1 Cov(T, R) \end{aligned}$$

Therefore,

$$\gamma_1 = \frac{Cov(T, Y)}{Cov(T, R)}$$

The 2SLS includes two steps:

$$Step\ 1: R_i = \alpha_0 + \alpha_1 T_i + \eta_i \qquad (2)$$

$$Step\ 2: Y_i = \beta_0 + \beta_1 \hat{R}_i + \nu_i \qquad (3)$$

Since $Cov(\hat{R}, \nu) = 0$, $\widehat{\beta_1}$ is unbiased estimator of β_1.

We can prove that $\beta_1 = \gamma_1$:

$$\beta_1 = \frac{Cov(\hat{R}, Y)}{Var(\hat{R})} = \frac{Cov(\hat{\alpha}_0 + \hat{\alpha}_1 T, Y)}{Var(\hat{\alpha}_0 + \hat{\alpha}_1 T)} = \frac{Cov(T, Y)}{Cov(R, T)} = \gamma_1$$

Hence, the 2SLS method can produce unbiased estimate for the effect of an endogenous exposure.

When the outcome is binary, most of the research investigated how to estimate RR or OR. A number of methods have been reported to estimate RR and OR, such as 2-stage logistic model[30, 31], probit structural equation models[32], 3-stage model[33], and two-stage residual inclusion method[34]. However, little attention has been paid to the estimation of RD when both the exposure and outcome are binary.

We consider 3 IV-based approaches to estimating RD in RCT with imperfect compliance, when the treatment and outcome are both binary: 2SLS, 2-stage logistic/linear approach (LG-LM), and 3-stage approach.

2.3.1 2SLS

The most commonly used technique for instrumental variable analysis is the 2SLS [18]. The 2SLS approach has been justified and widely used in the setting of continuous outcomes [35]. In the case of dichotomous exposures and outcomes, we can mathematically demonstrate that 2SLS can produce a RD estimate. In a linear regression model, the ordinary least squares method can be used to estimate the parameters, and this method produces best linear unbiased estimators of γ_0 and γ_1 [36]. It is known that the procedure, to obtain the unbiased estimator of γ_0 and γ_1, does not require assumptions of normality nor constant variance. Thus, if the research interest is to make estimation about the parameters, the normality and constant variance assumptions are not

required. In the case of binary outcome coded as 0 or 1, the outcome variable is not normally distributed, but the mean is equal to the proportion. Therefore, the OLS method can be used to analyze a binary response[16], and the estimate of RD is the regression coefficient $\hat{\gamma}_1$.

Previously, we have proved that the 2SLS method can produce unbiased estimate for the effect of an endogenous exposure in the setting of continuous exposures and outcomes. Similarly, we can show that the 2SLS method can also generate unbiased estimate of risk difference when both the treatment and outcome are dichotomous.

Suppose that the R, T, and Y are all dichotomous in equation (1), (2) and (3). A simple linear model can be fit between Y and T:

$$Y_i = \kappa_0 + \kappa_1 T_i + \omega_i \qquad (4)$$

We have known that

$$\gamma_1 = \frac{Cov(T,Y)}{Cov(T,R)}$$

It could be expressed as two ordinary least squares estimators:

$$\widehat{\kappa_1} = \frac{Cov(T,Y)}{Var(T)}, \ \widehat{\alpha_1} = \frac{Cov(T,R)}{Var(T)}$$

$$\gamma_1 = \frac{Cov(T,Y)}{Cov(T,R)} = \frac{\widehat{\kappa_1}}{\widehat{\alpha_1}}$$

In the setting of binary instrument, exposure and outcome, the numerator is the difference in mean outcome between $T = 0$ and $T = 1$, that is, $\mathrm{P}(Y = 1|\, T = 1) - \mathrm{P}(Y = 1|\, T = 0)$, and the denominator is the difference in mean exposure between $T = 0$ and $T = 1$, that is, $\mathrm{P}(R = 1|\, T = 1) - \mathrm{P}(R = 1|\, T = 0)$. When the subjects in the control group do not have access to the treatment group, $\mathrm{P}(R = 1|\, T = 0) = 0$.

Therefore,

$$\gamma_1 = \frac{\mathrm{P}(Y=1|\,T=1) - \mathrm{P}(Y=1|\,T=0)}{\mathrm{P}(R=1|\,T=1)}$$

We have proved that $\beta_1 = \gamma_1$, Thus,

$$\beta_1 = \gamma_1 = \frac{\mathrm{P}(Y=1|\,T=1) \; - \; \mathrm{P}(Y=1|\,T=0)}{\mathrm{P}(R=1|\,T=1)}$$

The assumptions of normality or constant variance are assumed when other statistical procedures are made, such as hypothesis testing and confidence interval construction. In the context of binary outcomes and exposures, the statistical inference (hypothesis testing or confidence interval) is not valid since the error structure for a binary outcome does not follow normal distribution. Cheung [16] proposed a modified least-squares regression approach to estimate RD by the OLS regression method accompanied by statistical inference based on the Huber-White robust estimate of variance. He demonstrated that this approach produces valid inferences.

2.3.2 LG-LM

Linear models in the 2SLS may not be natural model choices for binary outcome and exposure since the fitted values are not bounded by (0,1). Even so, in the 2nd stage (outcome regression), a linear model provides a direct estimate of RD through the regression coefficient associated with the exposure, making it a justifiable choice over common models for binary data such as logistic or probit regrssion. In the 1st stage however, we care less about the interpretation of the coefficients and hence we are not forced to use a linear model. A logistic model provides a natural alternative in regressing the exposure on the instrument to obtain fitted exposure values. These can then be used in stage 2 to estimate RD. As with the 2SLS approach, the standard errors could be calculated based on the Huber-White robust estimate of variance. Rassen et al [18]

proposed this two-stage modeling with the first stage logistic and the second stage ordinary least square. The confidence intervals in their study were based on bootstrapped standard errors. They tested this model in three observational studies and observed little difference in point estimate or precision between estimates from 2SLS models and LG-LM model.

2.3.3 3SLS

Considering the model misspecification issue when using 2SLS estimation in the setting of binary exposures and outcomes, that is, the predicted value could be out of the 0-1 range, fitting a logistic model in the first stage may be reasonable. The 2-stage LG-LM offers an alternative solution to model misspecification in the first stage. However, the misspecification of the logistic model in the first stage, i.e. the first stage logistic model is incorrect, would lead to inconsistent estimate in the second stage [18]. Whereas consistent estimates would always be generated from 2SLS. As a result, Angrist [37] proposed a 3-stage method:

(1) use a logistic regression of treatment (R) on instrument (T) to obtain the predicted probability $\hat{p}$ that $R_i = 1$;

(2) a linear first stage model is estimated using $\hat{p}$ as the instrument instead of T;

(3) the outcome linear probability model is fitted to obtain estimate of the risk difference.

3.0 EMPIRICAL STUDY

In this chapter, we conducted two sets of simulation studies to compare the empirical performance of the three previously described IV-based methods and the PP method in estimating the CACE for RCTs with noncompliance. We assumed that the treatment (T) and outcome (Y) are binary. We also assume that subjects assigned to the control group did not have access to the treatment. Therefore, the number of principal stratum is reduced to two: compliers (C) and non-compliers (NC). In the first set of simulation studies, we investigated the performance of IV-based methods when there is no measured confounder in the RCTs, while in the second simulation study, we examined the behavior of IV methods when accounting for a measured confounder. We varied the sample size, compliance probability, and effect size.

3.1 NO MEASURED CONFOUNDER

In this section, we consider the situation when there is no measured confounder between the treatment received and the outcome. This indicates that the noncompliance is due to some unmeasured confounders, which in theory could be corrected by IV approaches. We initially compared the three IV methods and the PP approach using a fixed sample size. We then expanded the simulation scenarios by including wider levels of compliance, effect size, and sample size albeit imposing a simplifying assumption between complier and non-complier outcomes.

3.1.1 IV-based methods and PP with fixed sample size

In this set of simulations, we make a less restrictive assumption that the average outcome between C and NC under control were different, i.e. the noncompliance is not random. This assumption reflects the real-world situation of noncompliance, generally non-adherents are less healthy and less health conscious than adherents. We took the case where the unmeasured confounder (U) is binary. The confounder predicts the compliance stratum (C or NC). Figure 4 shows a causal diagram that illustrates the relationship between these variables. The Y is dependent on U and R. Since individuals in the control group could not access treatment, the compliance strata (S) include two strata: C and NC, and the status of compliance is predicted by U. The R is dependent on the T and S, i.e. only if subjects are randomly assigned to the treatment group and they are compliers, they will receive treatment, otherwise, they will not. In fact, $R = TS$. The parameter of interest, CACE, is represented by γ_1. We used a linear probability model for the outcome so that γ_1 represents the risk-difference.

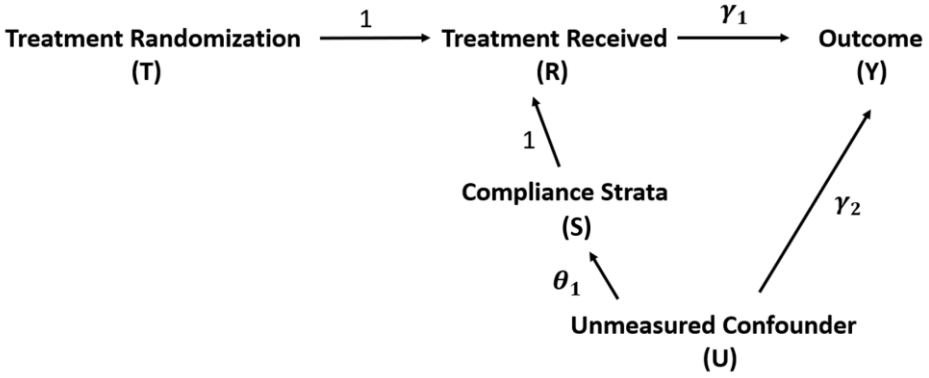

Figure 4. Graphical representation of a RCT with noncompliance, with randomization as an instrumental variable for the treatment received

Data generation

Based on the relationship between the variables in Figure 4, the following factors in the current simulation study were considered:

(1) Sample size (N): 500;

(2) Treatment assignment probability (π_T): the probability to be assigned to the treatment group, π_T was set to 0.5;

(3) Binary unmeasured confounder (U): U was drawn from a Bernoulli distribution with parameter π_U in the treatment group and the control group. Here, π_U was set to 0.5;

(4) To generate the compliance strata (S), the following semi-parametric linear probability model was applied:

$S = \theta_0 + \theta_1 U + \varepsilon_1$, where $E(\varepsilon_1) = 0$; $U = \{0,1\}$; $S = \{0,1\}$.

Then,

$$E(S|U) = \theta_0 + \theta_1 U$$

S was drawn from a Bernoulli distribution with probability equals to $E(S|U)$ in the treatment group and the control group.

(5) The treatment received (R) was generated based on the equation:

$$R = ST$$

that is, $R = 1$ if and only if $S = 1$ and $T = 1$.

(6) The binary outcome (Y) was generated based on the following semi-parametric linear probability model:

$E(Y = 1|R, U) = \gamma_0 + \gamma_1 R + \gamma_2 U$, where $U = \{0,1\}$; $R = \{0,1\}$; $Y = \{0,1\}$.

The outcome Y was drawn from a Bernoulli distribution with probability equals to $E(Y = 1|U, R)$.

We investigated the effect of changes in compliance probability levels (low, high) and risk difference levels (none, low, high) on the estimation properties using four approaches, namely, three IV-based methods and PP approach. The parameter settings for each scenario are shown in Table 1. Scenarios 1-3 represent low compliance (average compliance probability = 0.55) and Scenarios 4-6 represent high compliance (average compliance probability = 0.8). For each compliance setting, we varied the true risk difference (CACE) from no effect ($\gamma_1 = 0$) to as high as 0.3. For each scenario, 1000 simulated datasets were generated.

Table 1. No measured confounder simulation 1 scenarios

Scenario	Compliance	Effect	θ_0	θ_1	γ_0	γ_1	γ_2
1	Low	None	0.8	-0.5	0.4	0.00	-0.2
2		Low	0.8	-0.5	0.4	0.15	-0.2
3		High	0.8	-0.5	0.4	0.30	-0.2
4	High	None	0.9	-0.2	0.4	0.00	-0.2
5		Low	0.9	-0.2	0.4	0.15	-0.2
6		High	0.9	-0.2	0.4	0.30	-0.2

Analysis: Estimating risk difference

RD was estimated for each simulated dataset using three IV-based approaches: (1) 2SLS, (2) 2-stage LG-LM, and (3) 3-stage LG-LM-LM. The ivreg command from R package AER was used for fitting 2SLS. Here, robust standard error (R packages "sandwich" and "lmtest" were applied, type= "HC4") was required in the last stage of each method. Furthermore, we compared RD estimates from the PP method. Because individuals in the control group do not have access to the treatment, RD in the PP was estimated as the difference between the compliers in the treatment group and all the subjects in the control group.

To evaluate each approach's performance, the bias, relative bias, average of standard error (ASE), empirical standard error (ESE), coverage probability (CP), and power for assessing the treatment effect among compliers were examined. Relative bias was calculated as the mean of the

bias in each scenario divided by the corresponding true risk difference. Coverage probability is defined as the proportion of replications where the true parameters values are covered by the nominal 95% confidence interval of the parameter estimates[38]. Power is defined as the proportion of times that each method rejects the null hypothesis (the risk difference is equal to 0) using an $\alpha = 0.05$. To calculate the power in the PP method, chi-square test was conducted.

Results

In terms of the five evaluation parameters, the tested three IV-based methods generated valid estimates with almost identical performances. First, the three IV-based methods led to almost identical unbiased estimates (Figure 5), the relative bias ranged from 0.5% to 1% in all cases. The power (Figure 6) results showed that when the risk difference was zero, the type I error of three IV-based methods was about 5%. As the risk difference increases, the power goes up. When the risk difference was as high as 0.3, the power was close to 1. From Figure 8, we can see that the CP was very close to 95% in all cases when using the IV-based methods. With respect to ESE (Figure 7) and ASE, the values were almost identical in three IV methods, both the ESE and ASE decreased as the compliance probability grew. Also, the ASE generated from all three IV methods were all very close to the corresponding ESE.

In contrast, the performance of PP was not as good as the three IV-based methods. The relative bias was large when PP method was applied, especially when the compliance probability was low. For example, the relative bias was as high as 31% when the compliance probability was 0.55, the risk difference was 0.15. But when the compliance probability increased to 0.8, the relative bias dramatically decreased to 7%. The power results showed that when the risk difference was zero, the type I error of PP method was around 16% when the compliance probability was low. The type I error reduced to 5% when the compliance probability increased to 0.8. The CP was

less than 85% when compliance probability was as low as 0.55. However, when the compliance probability increased to 0.8, the CP was around 95%. The ESE and ASE results were very similar to that of IV methods, whereas PP had smaller ASE and ESE than three IV methods, especially when the compliance probability was low. This indicates that the PP method was more efficient than IV-based methods.

Concerning the impact of compliance probability on each method's performance, we can see that PP method was more sensitive to compliance probability than IV methods. The relative bias of PP greatly decreased as the compliance probability increased, while the relative bias of IV methods did not show this trend, it varied around 0.5% to 1%. The influence of compliance probability on power depends on risk difference, when the risk difference is none or high, the three IV methods' power was almost identical, on the other hand, when the risk difference was low, increasing the compliance probability led to stronger power. However, for the PP method, the compliance probability had greater influence when the risk difference was none, that is, increasing the compliance resulted in smaller type I error. The CP of IV methods keeps constant when the compliance probability change, whereas the CP of PP improved from about 85% to around 95% when the compliance probability changed from low to high. All four methods' ESE and ASE decreased as the compliance probability increased.

Comparing the influence of risk difference on each method's behavior, we can see that relative risk of IV methods was not very sensitive to risk difference when the compliance is low, but when the compliance is high, the relative risk decreased as the risk difference got larger. For PP method, the relative bias showed a decreasing trend as the risk difference increasing. The power of these four methods all revealed an increasing direction as the risk difference increased, the PP

method achieved high power (over 95%) when the risk difference was 0.15. As for ASE, ESE and CP, all four methods did not show strong sensitivity to risk difference.

To sum up, the three IV-based methods can all provide unbiased estimates with the settings in the current simulation, and their performances were almost identical in terms of relative bias, power, ESE, ASE, and CP. However, the estimate obtained from PP method is highly biased when the compliance probability is relative low. On the other hand, as the compliance probability increases, the performance of PP and IV converge.

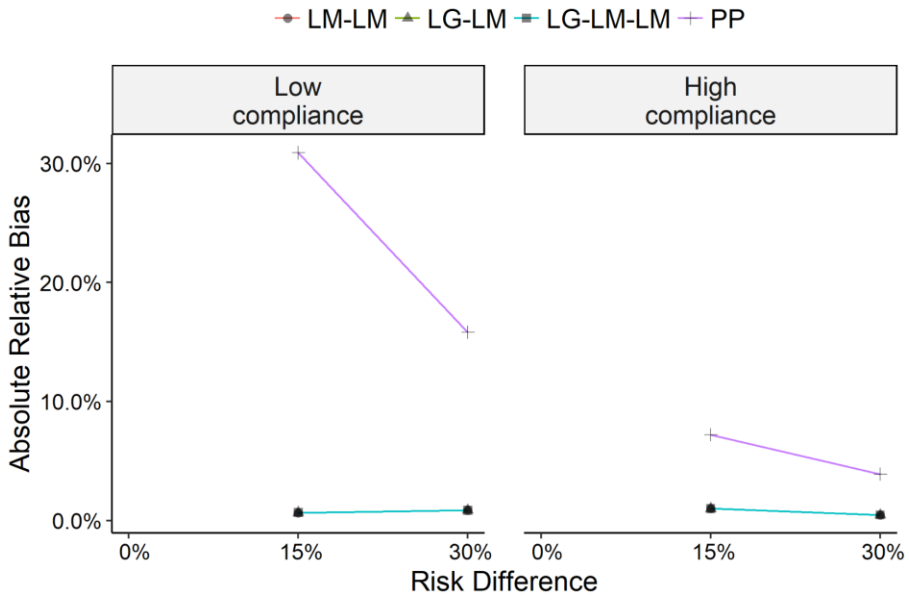

Figure 5. Absolute relative bias of IV-based methods and PP in each scenario

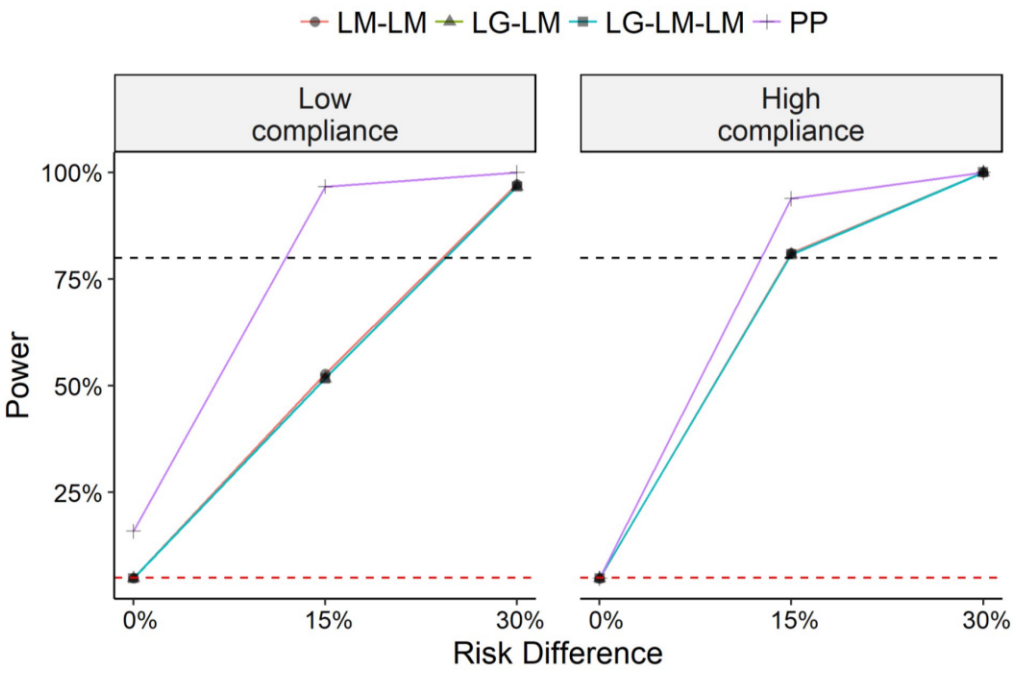

Figure 6. Power of IV-based methods and PP in each scenario

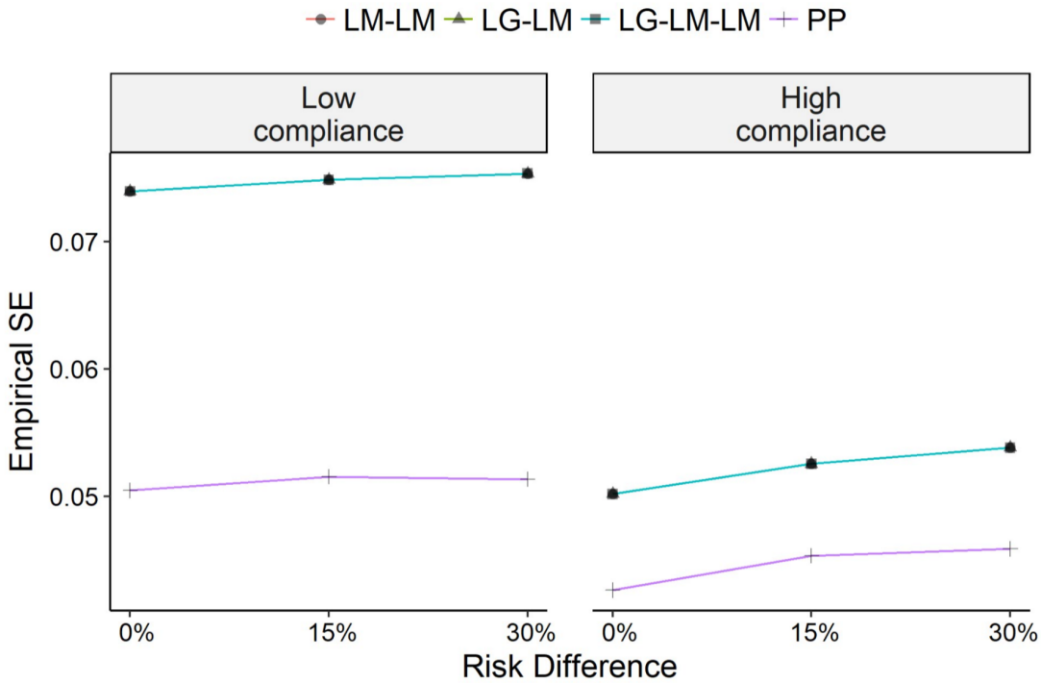

Figure 7. ESE of IV-based methods and PP in each scenario using

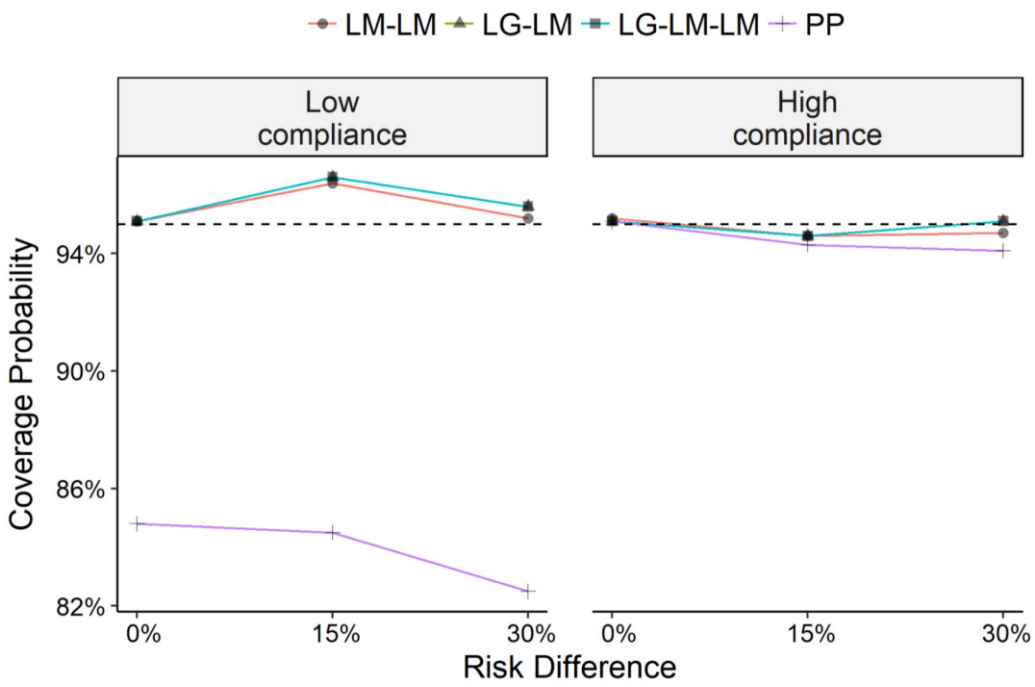

Figure 8. CP of IV-based methods and PP in each scenario

3.1.2 Expanded comparisons for IV-based methods

In this section, we carried out a series of simulation to investigate the sensitivity of the three IV-based methods to the variation of factors, namely, sample size, compliance probability, and risk difference. We varied the sample size and included more levels of compliance and effect sizes. For simplicity, we assume that noncompliance was unrelated with prognosis, that is, the potential outcomes for compliers and noncompliers under control were the same.

Data generation

We considered the following factors in the simulation study design:

(1) Sample size (N): three levels of sample size were considered -- 100, 250, 500;

(2) Treatment assignment probability (π_T): the probability one is assigned to the treatment group. π_T was set to 0.5.

(3) Compliance probability (π_R): the probability of compliance if one is assigned to the treatment group. We assessed four levels of π_R -- 0.50, 0.65, 0.8, 0.95;

(4) Risk difference (Δ): six levels of risk difference were evaluated -- 0.00, 0.05, 0.1, 0.15, 0.2, 0.25;

(5) The baseline effect (π_0): the probability of success when one does not receive treatment. π_0 was set to 0.5.

Thus, we have 72 different combination of scenarios (Table 2.) For each scenario, 1000 simulated datasets were generated. Each dataset was generated as follows:

1. The binary treatment assignment indicator T_i for the ith subject is simulated: T_i ~ Bernoulli (P= π_T), where π_T=0.5.

2. The treatment received indicator R_i for the i^{th} subject is draw from Bernoulli distribution:

$$R_i \sim \text{Bernoulli } (P = \pi_R)\ T_i$$

that is, for the i^{th} subject who is assigned to the control group ($T_i = 0$), the R_i is always 0, if assigned to the treatment group ($T_i = 1$), then R_i ~ Bernoulli ($P = \pi_R$).

3. The binary outcome Y_i for the i^{th} subject is simulated:

$$Y_i \sim \text{Bernoulli } (P = \pi_0)\ \text{I}(R_i = 0)$$

$$Y_i \sim \text{Bernoulli } (P = \pi_0 + \Delta)\ \text{I}(R_i = 1),$$

that is, if the i^{th} subject does not receive the treatment ($R_i = 0$), the subject's probability to have the outcome is π_0, otherwise, the probability is $\pi_0 + \Delta$, where π_0 is 0.5.

Table 2. No measured confounder simulation 2 scenarios

Scenario	N^*	π_R	Δ
1	100	0.5	0
2			0.05
3			0.1
4			0.15
5			0.2
6			0.25
7		0.65	0
8			0.05
9			0.1
10			0.15
11			0.2
12			0.25
13		0.8	0
14			0.05
15			0.1
16			0.15
17			0.2
18			0.25
19		0.95	0
20			0.05
21			0.1
22			0.15
23			0.2
24			0.25

Note: * repeated the process for sample size $N = 250, 500$.

Results

The relative bias, empirical standard errors, power and coverage probabilities are shown in Figures 9, 10, 11, and 12 respectively. In general, we observed the same patterns as in the previous simulations. The three IV-based methods have nearly identical performance. They yield estimates with low relative bias and confidence intervals and type I error rate close to nominal levels. Relative bias, efficiency (in terms of lower ESE), and power improve with higher sample sizes or higher compliance probabilities.

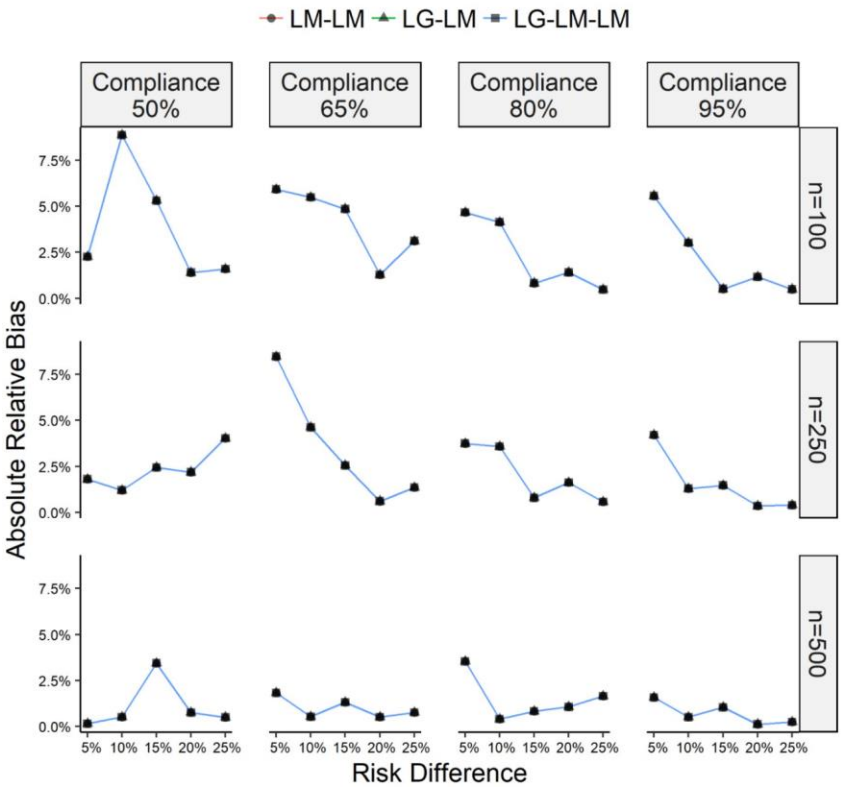

Figure 9. Absolute relative bias of IV-based methods in each scenario

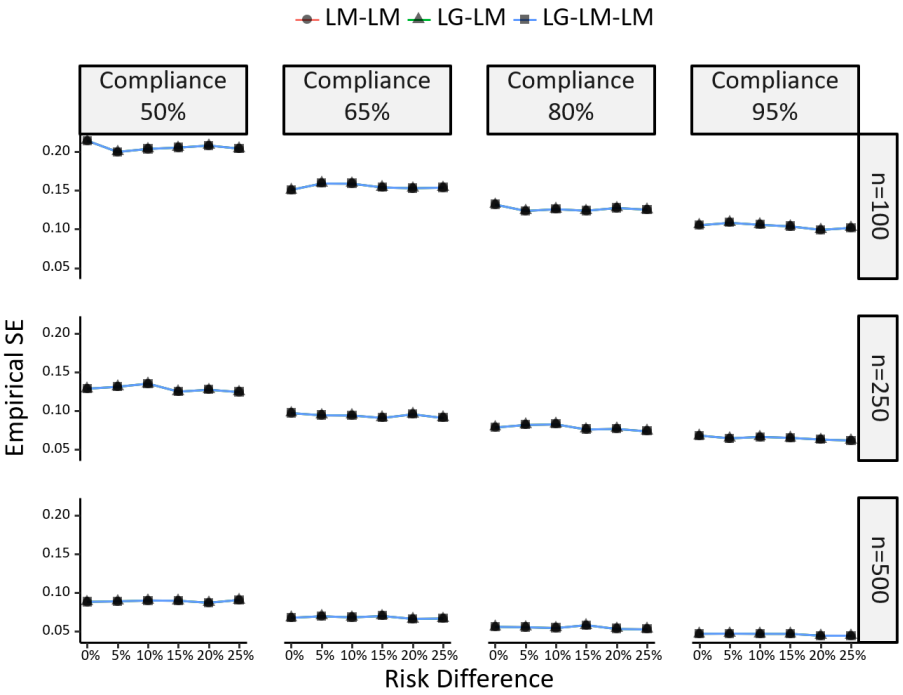

Figure 10. ESE of IV-based methods in each scenario

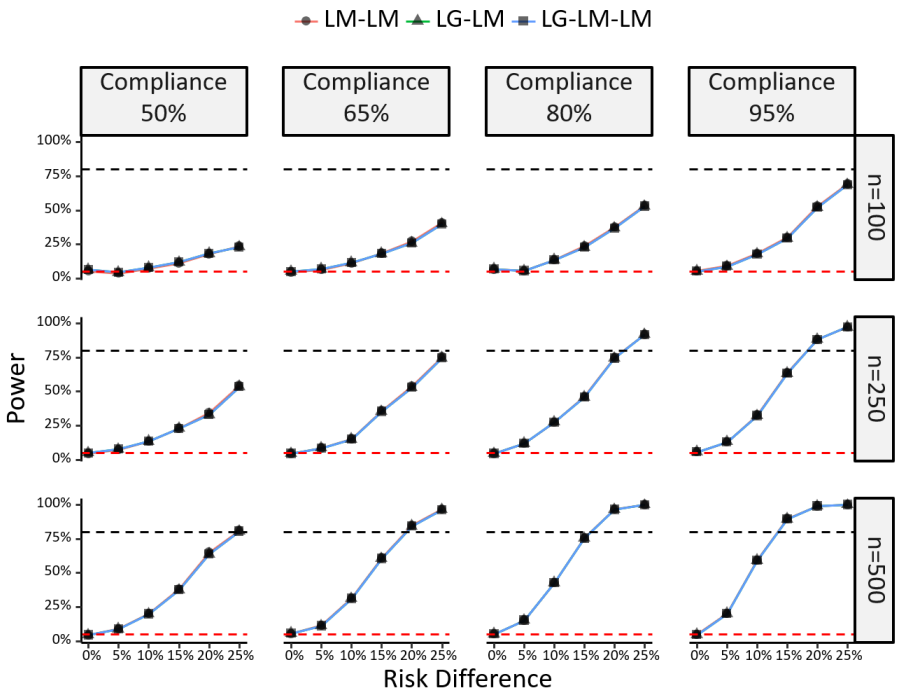

Figure 11. Power of IV-based methods in each scenario

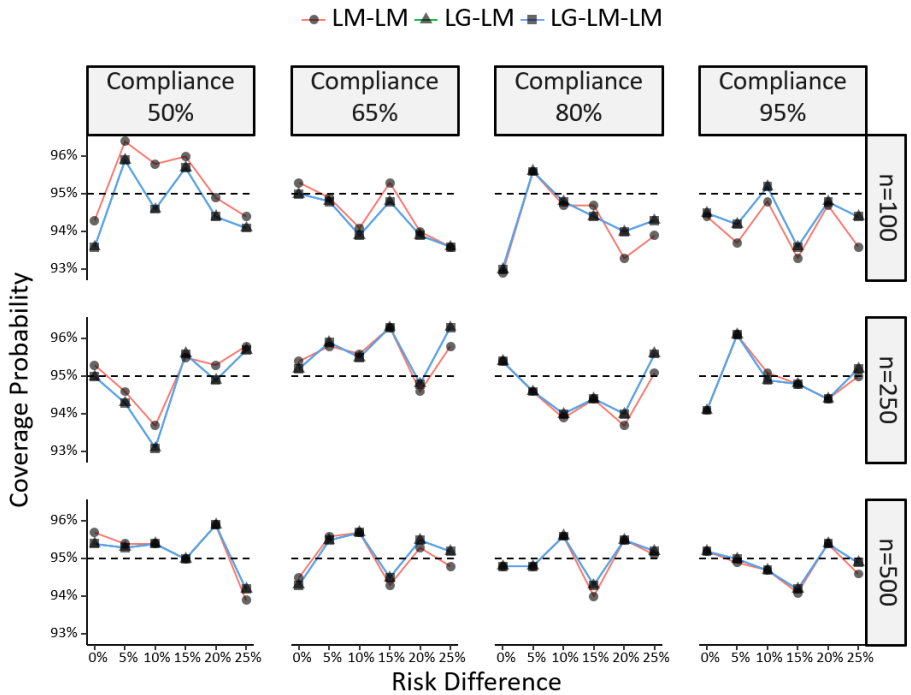

Figure 12. CP of IV-based methods in each scenario

3.2 ACCOUNTING FOR A MEASURED CONFOUNDER

In this section, we assume that a measured confounder between the treatment received and outcome exists. The aim was to investigate the statistical properties of CACE estimation when measured confounder was adjusted for in the three IV-based methods, especially in terms of efficiency. Furthermore, we examined the performance of 2SLS when the measured confounder was adjusted for or not. The causal diagram (Figure 13) is similar to that in the previous section, except for an additional measured confounder X. Here, the compliance strata are predicted by both the measured and unmeasured confounder. Outcome is dependent on X, U, and R.

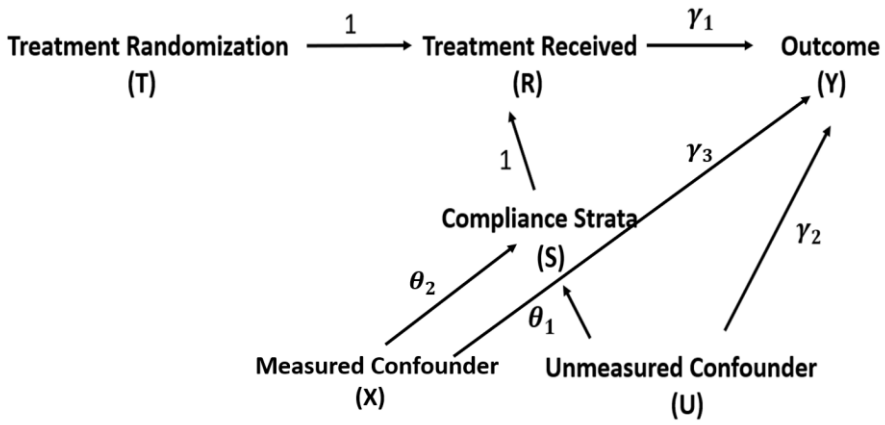

Figure 13. A RCT with noncompliance, when both measured and unmeasured confounders exist

Data generation

Based on the relationship shown in Figure 13 we simulated 24 scenarios to compare these three methods, varying the sample size, compliance probability and risk difference. The simulation procedure was conducted as follows:

(1) Sample size (N): 100, 250, 500, 1000;

(2) Treatment assignment probability (π_T): the probability to be assigned to the treatment group, π_T was set to 0.5;

(3) Binary unmeasured confounder (U): U was drawn from a Bernoulli distribution with parameter π_U in the treatment group and the control group. Here, π_U was set to 0.5;

(4) Continuous measured confounder (X): X was drawn from a uniform distribution:

$$X \sim Uniform(1,3)$$

(5) The compliance strata (S) was draw from a Bernoulli distribution with parameter p_1, which is a function of both X and U.

$$S \sim Bernoulli(p_1 = \frac{\exp(\theta_0 + \theta_1 U + \theta_2 X)}{1 + \exp(\theta_0 + \theta_1 U + \theta_2 X)})$$

(6) The treatment received (R) was generated based on the equation:

$$R = ST$$

that is, $R = 1$ if and only if $S = 1$ and $T = 1$.

(7) The binary outcome (Y) was generated based on the following semi-parametric linear probability model:

$$E(Y = 1|R, U, X) = \gamma_0 + \gamma_1 R + \gamma_2 U + \gamma_3 \log(X)$$

Here, X was log transformed to ensure that the probability is bounded between 0 and 1. Then,

$$E(Y = 1|R, U, X) = \gamma_0 + \gamma_1 R + \gamma_2 U + \gamma_3 \log(X)$$

The outcome Y was drawn from a Bernoulli distribution with probability equals to $E(Y|R, U, X)$.

In this simulation study, we investigated 24 scenarios with different sample size, compliance probability and risk difference. The parameter settings for each scenario are shown in Table 3. Scenario 1-3 represent low compliance (average compliance probability at 0.5), while

scenario 4-6 represent high compliance (average compliance probability at 0.9). For each compliance setting, the risk difference (γ_1) ranged from none (0.00) to low (0.15) to high (0.3). Scenario 1-6 were repeated for different sample size, $N = 250$, 500, 1000. For each scenario, 1000 simulated datasets were generated.

Table 3. Accounting for a measured confounder simulation scenarios

Scenario	Compliance	Effect	N*	θ_0	θ_1	θ_2	γ_0	γ_1	γ_2	γ_3
1	Low	None	100	7	3	-4	0.35	0.00	-0.1	0.25
2		Low	100	7	3	-4	0.35	0.15	-0.1	0.25
3		High	100	7	3	-4	0.35	0.30	-0.1	0.25
4	High	None	100	11	3	-4	0.35	0.00	-0.1	0.25
5		Low	100	11	3	-4	0.35	0.15	-0.1	0.25
6		High	100	11	3	-4	0.35	0.30	-0.1	0.25

Note: * repeated the process for sample size $N = 250$, 500,1000.

Analysis: Estimating risk difference

RD was estimated for each simulated dataset using three IV-based approaches. The measured confounder X was adjusted for at each stage. In the dataset generation procedure, the outcome was linear with the log transformed X.In real data analysis, however, we would not know the exact functional relationship between the covariate and outcome. Therefore, in the data analysis step, we used restricted cubic splines to flexibly model the functional form of the covariate. The "rcs" function from R package "rms" was used to do this. Additionally, we estimated the CACE with the 2SLS method without adjusting for the measured confounder. Robust standard error (R packages "sandwich" and "lmtest" were applied, type= "HC4") was required in the last stage of each method.

As with the previous simulation section, we report the bias, relative bias, average of standard error (ASE), empirical standard error (ESE), coverage probability (CP), and power in assessing the performance of the different approaches.

Results

Overall, after adjusting for the measured confounder, the statistical properties of these three methods are not identical unlike the simulation results from the previous section. This is particularly evident when the sample size is small and compliance probability is low.

In terms of relative bias (Figure14), we can see that the 3-stage model generated the smallest relative bias when the sample size was small, for example, when N=100, compliance probability was low, relative risk at 0.15, the relative risk from 3 stage model was as low as 0.004, while 2SLS was 0.03, and LG-LM was as high as 0.09. As the sample size increases, the difference becomes negligible. Also, the difference between 2SLS and 3 stage model is trivial when the compliance probability is high regardless of the sample size. Although the relative bias from LG-LM approach can decrease to as low as 1.5% when the sample size increase to 1000, the relative bias is consistently higher than other methods, especially when the sample size is small.

Similar trends were observed in terms of efficiency (Figure 15) and power (Figure 16). When the sample size was as small as 100 and compliance was low, the standard errors from the 3-stage approach were considerably smaller than those from 2SLS. As the sample size increased, the difference became negligible. For power, the 3-stage approach has stronger power than the 2SLS when the sample size is small. However, when the compliance was high, the behavior of these three methods were similar in terms of efficiency and power regardless of sample size. Therefore, when the sample size and the compliance probability are both small, the 3-stage method outperforms the other two methods.

In contrast, the three methods displayed almost the same performance in terms of CP (Figure 17). The CP in all cases were around 95%.

Finally, for the 2SLS approach, accounting for the measured confounder did not show any improved efficiency or decreased relative bias. Its performance is nearly the same as the 2SLS approach without covariate adjustment.

In sum, adjusting for the measured confounder, the 3-stage method performs better than the other three methods, especially when the sample size is small. The LG-LM method is likely to generate substantially higher relative bias than the other three methods when the sample size is small.

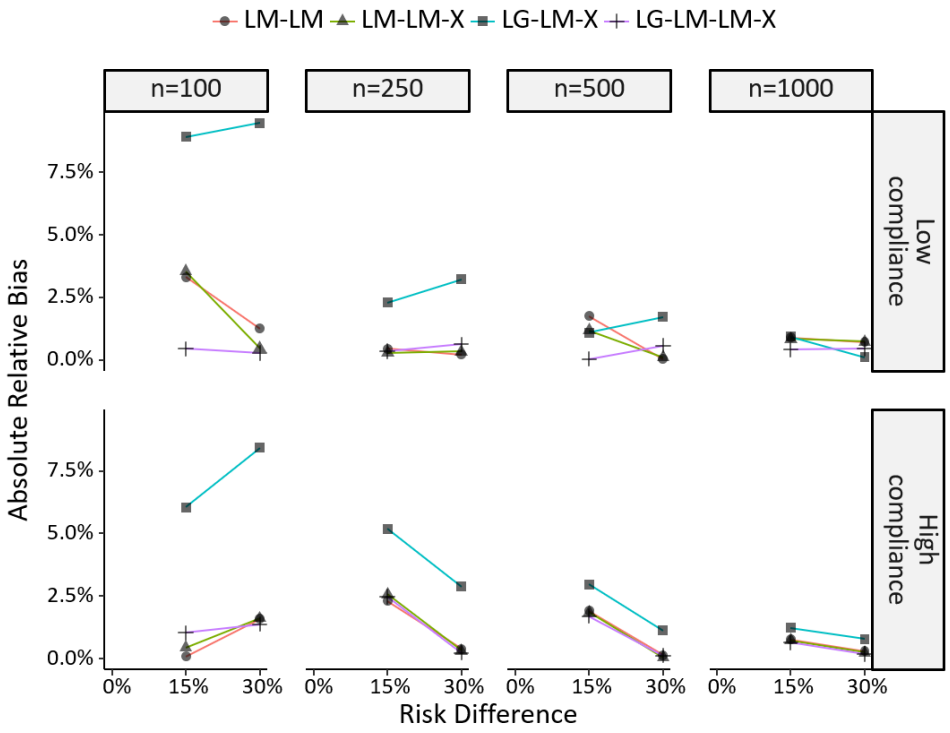

Figure 14. Absolute relative bias of IV-based methods in each scenario

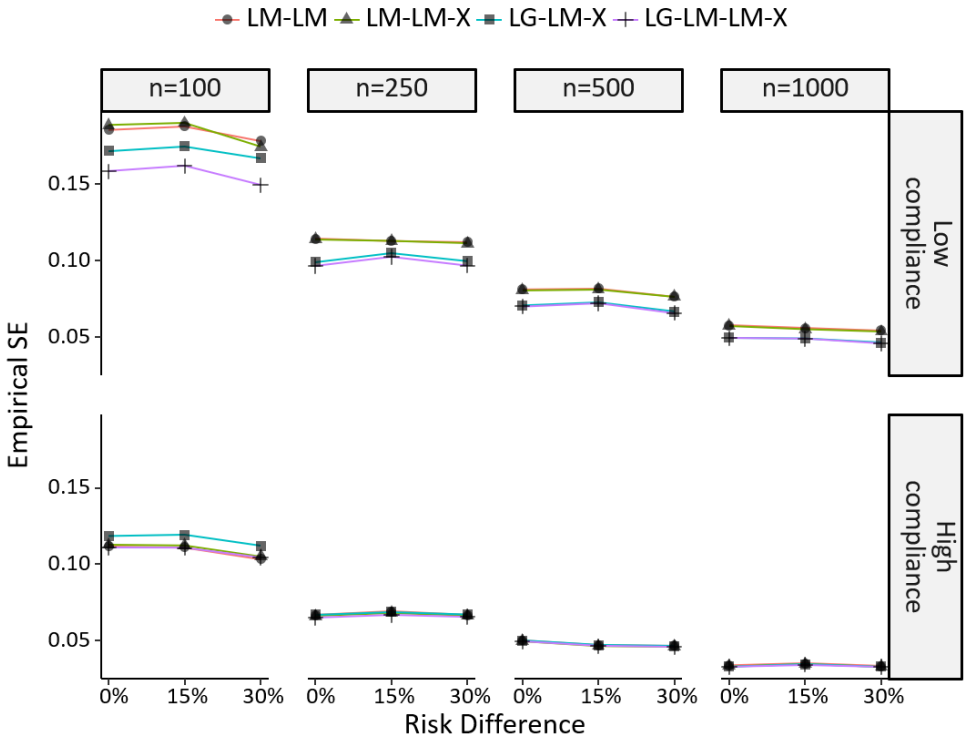

Figure 15. ESE of IV-based methods in each scenario

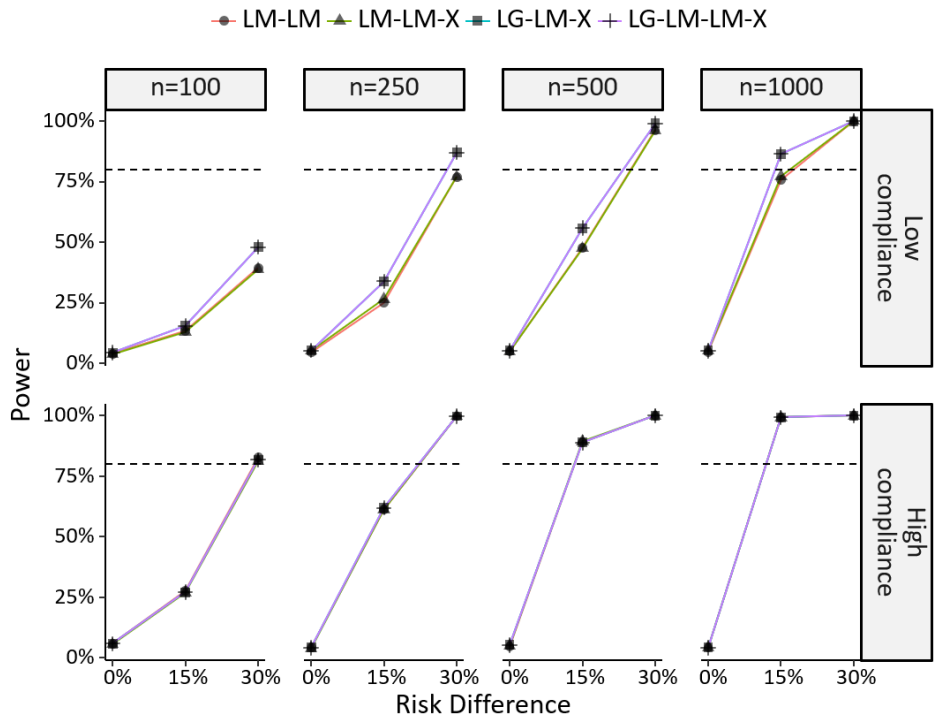

Figure 16. Power of IV-based methods in each scenario

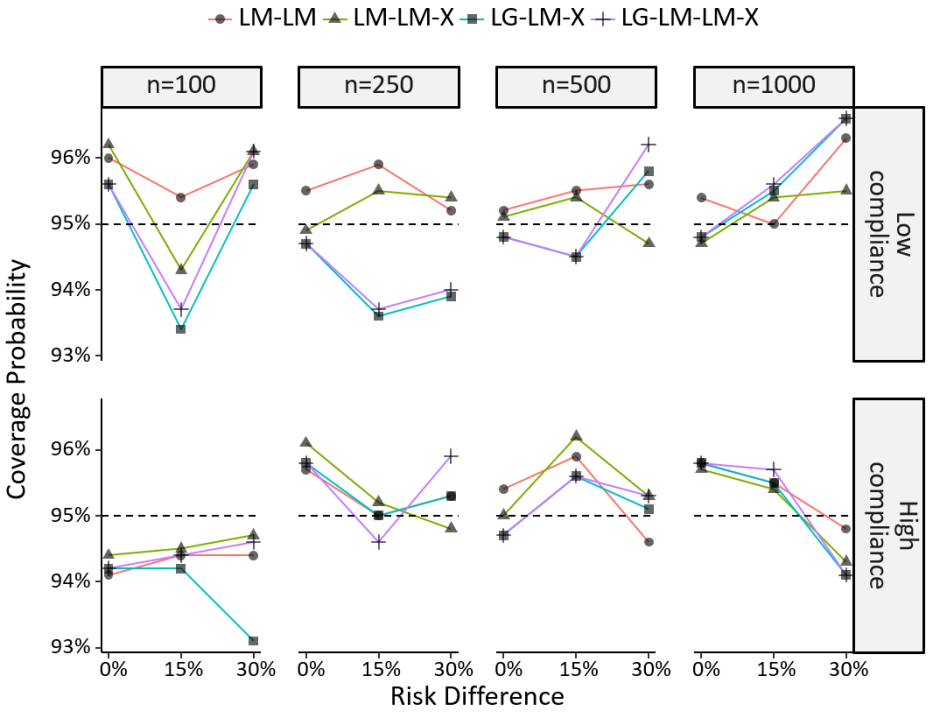

Figure 17. CP of each scenario using IV-based methods

4.0 DISCUSSION AND CONCLUSION

In RCTs, non-compliance to the assigned treatment often threaten the validity of estimation of treatment efficacy. CACE has been known as a potential estimate to represent the average treatment effects in the subgroup of compliers. IV-based method have been proposed to provide valid estimate of CACE. Although there have been many research investigated the IV-based methods in estimating CACE for binary outcome and binary treatment groups, most of them focused on estimating risk ratios or odds ratios. In this study, we studied the effectiveness of IV-based methods for the estimation of risk difference in RCTs with non-compliance, when both the exposure and outcome are binary. We conducted a series of simulation studies to compare three IV-based methods and conventional method PP with respect to bias, efficiency, CP and power in estimating risk difference as the CACE. We also examined how their performance is affected by varying levels of compliance, effect size, and sample size. In addition, we evaluated the behavior of three IV methods when measured confounders exist.

The results from all simulations indicate that compliance probability is a critical factor that influences the performance of different methods. For example, when the compliance probability is high, such as 0.95, each method tends to give valid estimate of treatment efficacy, although the relative bias from the PP method is consistently higher than IV methods. However, as the compliance at low or median level, the estimation obtained from PP method is highly biased, while the three IV-based methods still provide unbiased estimate.

Also, we examined the performance of the three IV-based methods in different levels of compliance, risk difference, and sample size, when there is no measured confounders. We found

that all the IV-based methods generally provide valid and almost the same estimates, efficiency and power in the setting of no measured confounders.

However, when measured confounders exist and we account for them, the 3-stage method performs better than the other two methods, especially when the sample size is small. It provides more efficient estimates and yield higher power. In contrast, the LG-LM method is likely to generate substantially higher relative bias than the other two methods when the sample size is small. However, as the compliance probability goes to 1 or as the sample size increases, the differences between the different IV-based methods become negligible.

When the research interest is in estimating the risk difference and there is no measured confounder, we mathematically and empirically showed that 2SLS is a valid approach to estimate CACE in the setting of binary outcome and treatment groups. We recommend using 2SLS to estimate when there is no measured confounder, as 2SLS is easier to perform and it will generate consistent estimate even when the model is misspecified. However, when there is measured confounder and the sample size is relatively small, we suggest using the 3-stage approach, as it will generate more efficient estimate and provide higher power. On the other hand, we would not suggest the use of LG-LM method, as it will give rise to estimates with high relative bias. The reason could be when the first stage is logistic regression, and the model is misspecified, unlike the 2SLS approach, it will not generate consistent estimate. However, when there is no measured confounders and treatment assignment is the only predictor in the first stage, fitting linear regression or logistic regression in the first stage will not cause inconsistency, as long as the probability is not extremely high or low[39].

In the last simulation, we did not observe significant improvement of efficiency when including measured confounder in the 2SLS compared to the estimate from 2SLS without adjusting

for a measured confounder. The potential reason could be including one confounder in the simulation model may not be sufficient to change the performance of 2SLS. In a future study, additional covariates could be included in the diagram to explore the influence of measured confounders on the performance of 2SLS method.

APPENDIX A: RESULT TABLES

Table 4. Relative bias and power of each scenario using four methods (Figure 5 and 6)

SCENARIO	2SLS		LG-LM		3SLS		PP	
	RELATIVE BIAS	POWER	RELATIVE BIAS	POWER	RELATIVE BIAS	POWER	RELATIVE BIAS	POWER
1	N/A	0.0490	N/A	0.0490	N/A	0.0490	N/A	0.1590
2	0.0066	0.5280	0.0066	0.5170	0.0066	0.5170	0.3092	0.9670
3	0.0086	0.9720	0.0086	0.9670	0.0086	0.9670	0.1583	1.0000
4	N/A	0.0480	N/A	0.0490	N/A	0.0490	N/A	0.0490
5	-0.0100	0.8110	-0.0100	0.8080	-0.0100	0.8080	0.0722	0.9390
6	-0.0046	1.0000	-0.0046	1.0000	-0.0046	1.0000	0.0390	1.0000

Table 5. ESE and ASE of each scenario using four methods (Figure 7)

SCENARIO	2SLS		LG-LM		3SLS		PP	
	ASE	ESE	ASE	ESE	ASE	ESE	ASE	ESE
1	0.0749	0.0739	0.0747	0.0739	0.0747	0.0739	0.0497	0.0505
2	0.0763	0.0748	0.0772	0.0748	0.0772	0.0748	0.0515	0.0515
3	0.0751	0.0753	0.0782	0.0753	0.0782	0.0753	0.0499	0.0513
4	0.0513	0.0502	0.0513	0.0502	0.0513	0.0502	0.0437	0.0426
5	0.0530	0.0525	0.0534	0.0525	0.0534	0.0525	0.0456	0.0453
6	0.0524	0.0538	0.0536	0.0538	0.0536	0.0538	0.0450	0.0459

Table 6. CP of each scenario using four methods (Figure 8)

SCENARIO	2SLS	LG-LM	3SLS	PP
	CP	CP	CP	CP
1	0.9510	0.9510	0.9510	0.8480
2	0.9640	0.9660	0.9660	0.8450
3	0.9520	0.9560	0.9560	0.8250
4	0.9520	0.9510	0.9510	0.9510
5	0.9460	0.9460	0.9460	0.9430
6	0.9470	0.9510	0.9510	0.9410

Table 7. Relative bias and power of each scenario using IV-based methods (Figure 9 and 11)

SCENARIO	2SLS		LG-LM		3SLS	
	RELATIVE BIAS	POWER	RELATIVE BIAS	POWER	RELATIVE BIAS	POWER
1	N/A*	0.055	N/A	0.063	N/A	0.063
2	-0.0226	0.038	-0.0226	0.044	-0.0226	0.044
3	0.0887	0.074	0.0887	0.078	0.0887	0.078
4	-0.0531	0.113	-0.0531	0.119	-0.0531	0.119
5	0.0140	0.180	0.0140	0.183	0.0140	0.183
6	-0.0160	0.236	-0.0160	0.230	-0.0160	0.230
7	N/A	0.046	N/A	0.050	N/A	0.050
8	-0.0592	0.063	-0.0592	0.068	-0.0592	0.068
9	0.0549	0.112	0.0549	0.115	0.0549	0.115
10	0.0486	0.185	0.0486	0.181	0.0486	0.181
11	0.0128	0.272	0.0128	0.260	0.0128	0.260
12	0.0312	0.410	0.0312	0.399	0.0312	0.399
13	N/A	0.067	N/A	0.068	N/A	0.068
14	0.0467	0.055	0.0467	0.057	0.0467	0.057
15	-0.0414	0.135	-0.0414	0.135	-0.0414	0.135
16	-0.0083	0.238	-0.0083	0.229	-0.0083	0.229
17	0.0141	0.376	0.0141	0.369	0.0141	0.369
18	0.0047	0.535	0.0047	0.529	0.0047	0.529
19	N/A	0.055	N/A	0.053	N/A	0.053
20	0.0557	0.092	0.0557	0.088	0.0557	0.088
21	0.0302	0.183	0.0302	0.178	0.0302	0.178
22	-0.0051	0.301	-0.0051	0.295	-0.0051	0.295
23	0.0116	0.529	0.0116	0.520	0.0116	0.520
24	-0.0049	0.693	-0.0049	0.687	-0.0049	0.687
25	N/A	0.046	N/A	0.049	N/A	0.049
26	0.0180	0.074	0.0180	0.077	0.0180	0.077
27	0.0121	0.136	0.0121	0.135	0.0121	0.135
28	0.0244	0.230	0.0244	0.229	0.0244	0.229
29	-0.0219	0.345	-0.0219	0.331	-0.0219	0.331
30	0.0403	0.544	0.0403	0.535	0.0403	0.535
31	N/A	0.046	N/A	0.046	N/A	0.046
32	0.0847	0.086	0.0847	0.085	0.0847	0.085
33	-0.0462	0.153	-0.0462	0.150	-0.0462	0.150
34	0.0255	0.360	0.0255	0.354	0.0255	0.354
35	-0.0061	0.541	-0.0061	0.529	-0.0061	0.529
36	-0.0136	0.755	-0.0136	0.747	-0.0136	0.747
37	N/A	0.046	N/A	0.045	N/A	0.045
38	0.0375	0.120	0.0375	0.119	0.0375	0.119
39	0.0358	0.275	0.0358	0.275	0.0358	0.275
40	-0.0079	0.464	-0.0079	0.458	-0.0079	0.458
41	0.0162	0.749	0.0162	0.744	0.0162	0.744
42	0.0058	0.918	0.0058	0.917	0.0058	0.917
43	N/A	0.058	N/A	0.058	N/A	0.058
44	0.0421	0.130	0.0421	0.131	0.0421	0.131
45	-0.0130	0.328	-0.0130	0.323	-0.0130	0.323
46	0.0147	0.638	0.0147	0.633	0.0147	0.633
47	0.0035	0.881	0.0035	0.881	0.0035	0.881
48	0.0039	0.973	0.0039	0.973	0.0039	0.973

Table 7 Continued

49	N/A	0.043	N/A	0.045	N/A	0.045
50	-0.0015	0.089	-0.0015	0.088	-0.0015	0.088
51	0.0052	0.200	0.0052	0.198	0.0052	0.198
52	-0.0344	0.380	-0.0344	0.374	-0.0344	0.374
53	0.0075	0.650	0.0075	0.638	0.0075	0.638
54	0.0050	0.811	0.0050	0.805	0.0050	0.805
55	N/A	0.053	N/A	0.056	N/A	0.056
56	-0.0182	0.114	-0.0182	0.112	-0.0182	0.112
57	0.0053	0.314	0.0053	0.311	0.0053	0.311
58	0.0132	0.609	0.0132	0.605	0.0132	0.605
59	0.0052	0.848	0.0052	0.842	0.0052	0.842
60	0.0076	0.968	0.0076	0.963	0.0076	0.963
61	N/A	0.052	N/A	0.052	N/A	0.052
62	0.0354	0.152	0.0354	0.152	0.0354	0.152
63	-0.0041	0.427	-0.0041	0.427	-0.0041	0.427
64	-0.0083	0.757	-0.0083	0.754	-0.0083	0.754
65	0.0107	0.967	0.0107	0.964	0.0107	0.964
66	-0.0165	1.000	-0.0165	0.999	-0.0165	0.999
67	N/A	0.048	N/A	0.047	N/A	0.047
68	0.0158	0.201	0.0158	0.201	0.0158	0.201
69	0.0051	0.593	0.0051	0.591	0.0051	0.591
70	0.0104	0.897	0.0104	0.894	0.0104	0.894
71	-0.0012	0.990	-0.0012	0.990	-0.0012	0.990
72	0.0025	1.000	0.0025	1.000	0.0025	1.000

Table 8. ESE and ASE of each scenario using IV-based methods (Figure 10)

SCENARIO	**2SLS**		**LG-LM**		**3SLS**	
	ASE	**ESE**	**ASE**	**ESE**	**ASE**	**ESE**
1	0.2049	0.2143	0.2048	0.2143	0.2048	0.2143
2	0.2061	0.1999	0.2062	0.1999	0.2062	0.1999
3	0.2039	0.2038	0.2042	0.2038	0.2042	0.2038
4	0.2022	0.2055	0.2033	0.2055	0.2033	0.2055
5	0.2016	0.2081	0.2031	0.2081	0.2031	0.2081
6	0.1995	0.2041	0.2024	0.2041	0.2024	0.2041
7	0.1548	0.1507	0.1556	0.1507	0.1556	0.1507
8	0.1554	0.1596	0.1561	0.1596	0.1561	0.1596
9	0.1545	0.1593	0.1555	0.1593	0.1555	0.1593
10	0.1528	0.1541	0.1543	0.1541	0.1543	0.1541
11	0.1517	0.1530	0.1537	0.1530	0.1537	0.1530
12	0.1490	0.1539	0.1516	0.1539	0.1516	0.1539
13	0.1251	0.1321	0.1260	0.1321	0.1260	0.1321
14	0.1251	0.1236	0.1261	0.1236	0.1261	0.1236
15	0.1242	0.1259	0.1253	0.1259	0.1253	0.1259
16	0.1228	0.1239	0.1243	0.1239	0.1243	0.1239
17	0.1210	0.1276	0.1226	0.1276	0.1226	0.1276
18	0.1197	0.1252	0.1218	0.1252	0.1218	0.1252
19	0.1050	0.1056	0.1060	0.1056	0.1060	0.1056
20	0.1043	0.1088	0.1053	0.1088	0.1053	0.1088
21	0.1036	0.1061	0.1047	0.1061	0.1047	0.1061

Table 8 Continued

22	0.1030	0.1039	0.1041	0.1039	0.1041	0.1039
23	0.1007	0.0992	0.1019	0.0992	0.1019	0.0992
24	0.0985	0.1019	0.0998	0.1019	0.0998	0.1019
25	0.1286	0.1289	0.1285	0.1289	0.1285	0.1289
26	0.1274	0.1313	0.1275	0.1313	0.1275	0.1313
27	0.1270	0.1352	0.1272	0.1352	0.1272	0.1352
28	0.1261	0.1252	0.1268	0.1252	0.1268	0.1252
29	0.1257	0.1275	0.1269	0.1275	0.1269	0.1275
30	0.1241	0.1246	0.1260	0.1246	0.1260	0.1246
31	0.0974	0.0974	0.0976	0.0974	0.0976	0.0974
32	0.0977	0.0948	0.0980	0.0948	0.0980	0.0948
33	0.0965	0.0946	0.0969	0.0946	0.0969	0.0946
34	0.0962	0.0915	0.0968	0.0915	0.0968	0.0915
35	0.0952	0.0960	0.0963	0.0960	0.0963	0.0960
36	0.0939	0.0914	0.0954	0.0914	0.0954	0.0914
37	0.0791	0.0792	0.0794	0.0792	0.0794	0.0792
38	0.0789	0.0825	0.0791	0.0825	0.0791	0.0825
39	0.0786	0.0833	0.0790	0.0833	0.0790	0.0833
40	0.0776	0.0766	0.0782	0.0766	0.0782	0.0766
41	0.0763	0.0773	0.0771	0.0773	0.0771	0.0773
42	0.0751	0.0741	0.0762	0.0741	0.0762	0.0741
43	0.0664	0.0681	0.0667	0.0681	0.0667	0.0681
44	0.0663	0.0647	0.0666	0.0647	0.0666	0.0647
45	0.0658	0.0665	0.0661	0.0665	0.0661	0.0665
46	0.0649	0.0652	0.0652	0.0652	0.0652	0.0652
47	0.0639	0.0631	0.0642	0.0631	0.0642	0.0631
48	0.0624	0.0619	0.0629	0.0619	0.0629	0.0619
49	0.0898	0.0883	0.0898	0.0883	0.0898	0.0883
50	0.0897	0.0889	0.0897	0.0889	0.0897	0.0889
51	0.0894	0.0899	0.0897	0.0899	0.0897	0.0899
52	0.0887	0.0896	0.0892	0.0896	0.0892	0.0896
53	0.0882	0.0872	0.0891	0.0872	0.0891	0.0872
54	0.0871	0.0905	0.0885	0.0905	0.0885	0.0905
55	0.0688	0.0677	0.0688	0.0677	0.0688	0.0677
56	0.0687	0.0699	0.0688	0.0699	0.0688	0.0699
57	0.0685	0.0684	0.0687	0.0684	0.0687	0.0684
58	0.0679	0.0703	0.0683	0.0703	0.0683	0.0703
59	0.0673	0.0661	0.0680	0.0661	0.0680	0.0661
60	0.0664	0.0666	0.0674	0.0666	0.0674	0.0666
61	0.0559	0.0562	0.0560	0.0562	0.0560	0.0562
62	0.0559	0.0557	0.0560	0.0557	0.0560	0.0557
63	0.0554	0.0548	0.0556	0.0548	0.0556	0.0548
64	0.0549	0.0580	0.0552	0.0580	0.0552	0.0580
65	0.0541	0.0538	0.0546	0.0538	0.0546	0.0538
66	0.0531	0.0533	0.0538	0.0533	0.0538	0.0533
67	0.0470	0.0471	0.0471	0.0471	0.0471	0.0471
68	0.0469	0.0473	0.0470	0.0473	0.0470	0.0473
69	0.0466	0.0470	0.0467	0.0470	0.0467	0.0470
70	0.0459	0.0470	0.0461	0.0470	0.0461	0.0470
71	0.0452	0.0445	0.0454	0.0445	0.0454	0.0445
72	0.0441	0.0445	0.0444	0.0445	0.0444	0.0445

Table 9. CP of each scenario using IV-based methods (Figure 12)

SCENARIO	2SLS	LG-LM	3SLS
	CP	CP	CP
1	0.943	0.936	0.936
2	0.964	0.959	0.959
3	0.958	0.946	0.946
4	0.960	0.957	0.957
5	0.949	0.944	0.944
6	0.944	0.941	0.941
7	0.953	0.950	0.950
8	0.949	0.948	0.948
9	0.941	0.939	0.939
10	0.953	0.948	0.948
11	0.940	0.939	0.939
12	0.936	0.936	0.936
13	0.929	0.930	0.930
14	0.956	0.956	0.956
15	0.947	0.948	0.948
16	0.947	0.944	0.944
17	0.933	0.940	0.940
18	0.939	0.943	0.943
19	0.944	0.945	0.945
20	0.937	0.942	0.942
21	0.948	0.952	0.952
22	0.933	0.936	0.936
23	0.947	0.948	0.948
24	0.936	0.944	0.944
25	0.953	0.950	0.950
26	0.946	0.943	0.943
27	0.937	0.931	0.931
28	0.955	0.956	0.956
29	0.953	0.949	0.949
30	0.958	0.957	0.957
31	0.954	0.952	0.952
32	0.958	0.959	0.959
33	0.956	0.955	0.955
34	0.963	0.963	0.963
35	0.946	0.948	0.948
36	0.958	0.963	0.963
37	0.954	0.954	0.954
38	0.946	0.946	0.946
39	0.939	0.940	0.940
40	0.944	0.944	0.944
41	0.937	0.940	0.940
42	0.951	0.956	0.956
43	0.941	0.941	0.941
44	0.961	0.961	0.961
45	0.951	0.949	0.949
46	0.948	0.948	0.948
47	0.944	0.944	0.944
48	0.950	0.952	0.952

Table 9 Continued

49	0.957	0.954	0.954
50	0.954	0.953	0.953
51	0.954	0.954	0.954
52	0.950	0.950	0.950
53	0.959	0.959	0.959
54	0.939	0.942	0.942
55	0.945	0.943	0.943
56	0.956	0.955	0.955
57	0.957	0.957	0.957
58	0.943	0.945	0.945
59	0.953	0.955	0.955
60	0.948	0.952	0.952
61	0.948	0.948	0.948
62	0.948	0.948	0.948
63	0.956	0.956	0.956
64	0.940	0.943	0.943
65	0.955	0.955	0.955
66	0.951	0.952	0.952
67	0.952	0.952	0.952
68	0.949	0.950	0.950
69	0.947	0.947	0.947
70	0.941	0.942	0.942
71	0.954	0.954	0.954
72	0.946	0.949	0.949

Table 10. relative bias and power of each scenario using IV-based methods (Figure 14 and 16)

SCENARIO	LM-LM		LM-LM-X		LG-LM-X		LG-LM-LM-X	
	RELATIVE BIAS	POWER	RELATIVE BIAS	POWER	RELATIVE BIAS	POWER	RELATIVE BIAS	POWER
1	N/A	0.0400	N/A	0.0380	N/A	0.0440	N/A	0.0440
2	0.0331	0.1330	0.0354	0.1300	0.0892	0.1550	0.0044	0.1550
3	0.0126	0.3930	0.0046	0.3880	0.0948	0.4800	0.0027	0.4800
4	N/A	0.0590	N/A	0.0560	N/A	0.0580	N/A	0.0580
5	-0.0007	0.2760	-0.0043	0.2690	0.0605	0.2710	-0.0102	0.2710
6	0.0158	0.8270	0.0161	0.8120	0.0844	0.8180	0.0137	0.8180
7	N/A	0.0450	N/A	0.0510	N/A	0.0530	N/A	0.0530
8	0.0044	0.2500	-0.0028	0.2650	0.0229	0.3380	-0.0035	0.3380
9	0.0021	0.7700	0.0032	0.7700	0.0322	0.8700	0.0062	0.8700
10	N/A	0.0430	N/A	0.0390	N/A	0.0420	N/A	0.0420
11	0.0229	0.6130	0.0257	0.6120	0.0519	0.6200	0.0247	0.6200
12	0.0039	0.9980	0.0030	0.9970	0.0287	0.9980	0.0022	0.9980
13	N/A	0.0480	N/A	0.0490	N/A	0.0520	N/A	0.0520
14	0.0174	0.4760	0.0118	0.4740	0.0109	0.5580	-0.0003	0.5580
15	0.0003	0.9620	-0.0010	0.9610	0.0171	0.9900	0.0057	0.9900
16	N/A	0.0460	N/A	0.0500	N/A	0.0530	N/A	0.0530
17	0.0190	0.8890	0.0184	0.8930	0.0296	0.8890	0.0170	0.8890
18	0.0012	1.0000	-0.0004	1.0000	0.0112	1.0000	-0.0012	1.0000
19	N/A	0.0460	N/A	0.0530	N/A	0.0520	N/A	0.0520
20	0.0088	0.7560	0.0083	0.7720	0.0093	0.8640	0.0040	0.8640
21	-0.0071	1.0000	-0.0072	1.0000	0.0010	1.0000	-0.0044	1.0000
22	N/A	0.0420	N/A	0.0430	N/A	0.0420	N/A	0.0420
23	0.0076	0.9930	0.0073	0.9930	0.0123	0.9930	0.0063	0.9930
24	0.0030	1.0000	0.0024	1.0000	0.0078	1.0000	0.0018	1.0000

Table 11. ESE and ASE of each scenario using IV-based methods (Figure 15)

SCENARIO	LM-LM		LM-LM-X		LG-LM-X		LG-LM-LM-X	
	ASE	ESE	ASE	ESE	ASE	ESE	ASE	ESE
1	0.1850	0.1855	0.1864	0.1886	0.1726	0.1717	0.1588	0.1583
2	0.1849	0.1878	0.1864	0.1900	0.1738	0.1748	0.1598	0.1618
3	0.1811	0.1782	0.1822	0.1743	0.1709	0.1670	0.1568	0.1492
4	0.1105	0.1118	0.1118	0.1127	0.1175	0.1183	0.1099	0.1107
5	0.1098	0.1111	0.1114	0.1122	0.1175	0.1192	0.1098	0.1106
6	0.1042	0.1034	0.1052	0.1051	0.1113	0.1119	0.1040	0.1044
7	0.1147	0.1141	0.1142	0.1138	0.1009	0.0989	0.0983	0.0964
8	0.1150	0.1128	0.1144	0.1127	0.1019	0.1051	0.0993	0.1023
9	0.1124	0.1121	0.1117	0.1111	0.0998	0.0996	0.0973	0.0966
10	0.0694	0.0667	0.0693	0.0659	0.0702	0.0669	0.0684	0.0652
11	0.0690	0.0690	0.0688	0.0681	0.0699	0.0689	0.0681	0.0669
12	0.0656	0.0667	0.0654	0.0662	0.0667	0.0672	0.0649	0.0656
13	0.0805	0.0813	0.0797	0.0806	0.0696	0.0708	0.0688	0.0700
14	0.0809	0.0818	0.0802	0.0812	0.0706	0.0729	0.0698	0.0721
15	0.0789	0.0764	0.0782	0.0763	0.0692	0.0666	0.0684	0.0657
16	0.0491	0.0501	0.0487	0.0497	0.0487	0.0499	0.0481	0.0493
17	0.0487	0.0471	0.0484	0.0465	0.0485	0.0470	0.0479	0.0464
18	0.0463	0.0466	0.0459	0.0461	0.0463	0.0463	0.0457	0.0458
19	0.0566	0.0575	0.0560	0.0572	0.0488	0.0495	0.0485	0.0492
20	0.0569	0.0559	0.0563	0.0550	0.0495	0.0490	0.0492	0.0488
21	0.0556	0.0543	0.0551	0.0535	0.0485	0.0461	0.0483	0.0458
22	0.0347	0.0337	0.0343	0.0333	0.0342	0.0331	0.0340	0.0329
23	0.0344	0.0352	0.0341	0.0346	0.0340	0.0345	0.0338	0.0343
24	0.0327	0.0333	0.0323	0.0331	0.0324	0.0331	0.0323	0.0329

Table 12. CP of each scenario using IV-based methods (Figure 17)

SCENARIO	LM-LM	LM-LM-X	LG-LM-X	LG-LM-LM-X
	CP	CP	CP	CP
1	0.9600	0.9620	0.9560	0.9560
2	0.9540	0.9430	0.9340	0.9370
3	0.9590	0.9610	0.9560	0.9610
4	0.9410	0.9440	0.9420	0.9420
5	0.9440	0.9450	0.9420	0.9440
6	0.9440	0.9470	0.9310	0.9460
7	0.9550	0.9490	0.9470	0.9470
8	0.9590	0.9550	0.9360	0.9370
9	0.9520	0.9540	0.9390	0.9400
10	0.9570	0.9610	0.9580	0.9580
11	0.9500	0.9520	0.9500	0.9460
12	0.9530	0.9480	0.9530	0.9590
13	0.9520	0.9510	0.9480	0.9480
14	0.9550	0.9540	0.9450	0.9450
15	0.9560	0.9470	0.9580	0.9620
16	0.9540	0.9500	0.9470	0.9470
17	0.9590	0.9620	0.9560	0.9560
18	0.9460	0.9530	0.9510	0.9530
19	0.9540	0.9470	0.9480	0.9480
20	0.9500	0.9540	0.9550	0.9560
21	0.9630	0.9550	0.9660	0.9660
22	0.9580	0.9570	0.9580	0.9580
23	0.9550	0.9540	0.9550	0.9570
24	0.9480	0.9430	0.9410	0.9410

APPENDIX B: R SAMPLE CODES

```
## n: sample size, Pc: comply% given treat=1, b0=0, b1: treatment effect,
## P0: binary outcome baseline mean P=0.5
######################## Parameters#############################
## generate all kinds of combination
risk <- c(0.00, 0.05, 0.1, 0.15, 0.2, 0.25)
comply.per <- c(0.5, 0.65, 0.8, 0.95)
size<- c(100, 250, 500)
combination <- function(n,Pc,b1){
        x <- matrix(0,nrow = 24, ncol = 3)
        for(j in 1:length(Pc)){
                for (k in 1:length(b1)){
                        x[(k+6*(j-1)),]<- c(n,Pc[j],b1[k])
                }
        }
        return(x)
}
size.100 <- combination(size[1],comply.per,risk)
size.250 <- combination(size[2],comply.per,risk)
size.500 <- combination(size[3],comply.per,risk)
all.combi <- rbind(size.100, size.250, size.500)
new.combi<-   data.frame(cbind(all.combi[,c(1,2)],all.combi[,3],0.5))       ##
generate all kinds of combination
names(new.combi)<-c("size","Pcomply","effect","P0")
write.csv(new.combi,file = 'parameters set_binary.csv')


############################dataset generation#########################
data.gen_bi <- function(n,Pc,b1,P0){
        gen_data <- list()
```

```
trails <- 1000
for (i in 1:trails) {
        # generate treatment assignment
        treat <- rbinom(n, 1, 0.5)  ## n sample size
        # order treat
        data.1 <- cbind(treat)
        data.1 <- data.1[order(-treat),]############ order by treat
        n.treat <- table(treat)
        # n.treat[2] is the number that treat=1
        # compliance: 0.5, generate comply
        comply <- rbinom(n.treat[2], 1, Pc)
        # generate comply status given treat=1
        rest <- rep(0, n.treat[1])
       # generate comply status=0 given treat=0
        comply <- c(comply,rest)     # merge comply and rest
        data.1 <- cbind(data.1,comply)
        ## binary outcome, set baseline p=0.5,
        ##means if subjects do not comply, their chance of event is 0.5
        ## treatment effect b1
        data.1 <- data.frame(data.1[order(comply),])
       # order by comply, increasing
       # generate binary outcome when comply=0 , p=0.5
       binary.y.0 <-rbinom(length(data.1[data.1$comply==0,2]),1,P0)
       set.seed(s_4*trails+i)
       binary.y.1<-rbinom(length(data.1[data.1$comply==1,2]),1,P0+b1)
       binary.y <- c(binary.y.0,binary.y.1)
       data.1<- data.frame(cbind(data.1,binary.y))
       names(data.1) <- c("treat","comply","binary.y")
       gen_data[[i]] <- data.1
}

return(gen_data)
```

```
}

data1_72 <- function(x){
        new.data <- list()
        for (i in 1:72){
                new.data[[i]]                                                <-
data.gen_bi(x[i,1],x[i,2],x[i,3],x[i,4],x[i,5],
                                          x[i,6],x[i,7],x[i,8])
              }
        return(new.data)
}
Set.seed(1567)
all.data_bi <- data1_72(new.combi)
## Binary dataset: 72 scenarios, for each scenarios there are 1000 datasets,
1000 replications.

##################################DATA ANALYSIS#########################
## approach 1: 2SLS, lm_lm

library(AER)
library(sandwich)
library(lmtest)

present.2SLS.bi <- function(all.data,parameter){
        Bias <- c()
        ASE <- c()
        ESE <- c()
        CP <- c()
        median_est <- c()  # output the median of point estimate
        relative_bias <- c()
        #calculate the relative bias, relative_bias=mean(bias)/effect
        power_bi <- c()  #calcualte the power
```

```
        for (i in 1:72) {
                est.bi <- c()    #point estimate
                est.bi.se <- c()   #Robust SE for point estimate
                lower.bi <- c()    # lower 95% CI
                upper.bi <- c()    # upper 95% CI
                bias.bi <- c()         # bias for each point estimate
                converge.p.bi <- c()
                # converge symbol: 1, true treatment effect falls in 95% CI
                p_value <- c()  # extract the p value for each treatment term
                pow <- c()
               # power indicator: equals if the p value of
               #treatment term smaller than 0.05

                for (j in 1:1000){
                        data_1 <- all.data[[i]][[j]]
                        # define outcome, instrument, endogenous variable
                        Y1 <- data_1$binary.y
                        Y2 <- data_1$comply
                        X2 <- data_1$treat
                        # 2SLS estimation (AER Package)
                        ivreg.bi <- ivreg(Y1 ~ Y2|X2)

                        # sandwich_se <- diag(vcovHC(mod, type = "HC"))^0.5
                        # temp.1 <- summary(ivreg.bi, vcov=sandwich)
                        temp.1  <-  coeftest(ivreg.bi,  vcov=vcovHC(ivreg.bi,
type = "HC4"))
                        est.bi[j] <- temp.1[2,1]

                        est.bi.se[j] <- temp.1[2,2]
                        p_value[j] <- temp.1[2,4]
                        pow[j] <- ifelse(p_value[j]<0.05, 1, 0)
```

```
                        ## if the treatment coefficient=0 was rejected,
                        ## then symbol as 1


                         lower.bi[j]  <-coefci(ivreg.bi,  vcov=vcovHC(ivreg.bi,
type = "HC4"))[2,1]  ## calculate CI based on robust SE
                         upper.bi[j]  <-coefci(ivreg.bi,  vcov=vcovHC(ivreg.bi,
type = "HC4"))[2,2]

                         converge.p.bi[j] <- ifelse(parameter[i,3]>lower.bi[j]
& parameter[i,3]<upper.bi[j], 1, 0)
                         bias.bi[j] <-est.bi[j]-parameter[i,3]
                }
                Bias[i] <- mean(bias.bi)
                ASE[i] <- mean(est.bi.se)
                ESE[i] <- sd(est.bi)
                CP[i] <- sum(converge.p.bi)/1000
                power_bi[i] <- sum(pow)/1000
                median_est[i] <- median(est.bi)
                relative_bias[i] <- mean(bias.bi)/(parameter[i,3])
        }
        return(cbind(Bias,ASE,ESE,CP,power_bi,median_est,relative_bias))

}
LM_LM_Bi <- data.frame(present.2SLS.bi(all.data_bi,new.combi))
write.csv(LM_LM_Bi, file="lm_lm_bi.csv")
#########################################################################
## approach 2: lg_lm
present.lg_lm.bi<- function(all.data,parameter){
  Bias <- c()
  ASE <- c()
  ESE <- c()
```

```
CP <- c()
median_est <- c()  # output the median of point estimate
relative_bias <- c()
#calculate the relative bias, relative_bias= mean(bias)/effect
power_bi <- c()  #calcualte the power
for (i in 1:72) {
  est.bi <- c()    #point estimate
  est.bi.se <- c()   #Robust SE for point estimate
  lower.bi <- c()    # lower 95% CI
  upper.bi <- c()    # upper 95% CI
  bias.bi <- c()         # bias for each point estimate
  converge.p.bi <- c()
  # converge symbol: 1, true treatment effect falls in 95% CI
  p_value <- c()  # extract the p value for each treatment term
  pow <- c()
 # power indicator: equals if the p value of treatment term smaller than 0.05
  for (j in 1:1000){
    data_1 <- all.data[[i]][[j]]
    # define outcome, instrument, endogenous variable
    Y1 <- data_1$binary.y
    Y2 <- data_1$comply
    X2 <- data_1$treat
    # 1st stage fit logistic regression
    logisreg.bi <- glm(Y2~X2,family = binomial(link ='logit'), data =data_1,
                       control = list(maxit = 50))
    phat.bi<- predict(logisreg.bi, type="response")

    # 2nd stage fit linear regression estimate treatment effect
    logi.bi <- lm(Y1 ~ phat.bi)

    temp.2 <- coeftest(logi.bi, vcov=vcovHC(logi.bi, type = "HC4"))
    est.bi[j] <- temp.2[2,1]
```

```
      est.bi.se[j] <- temp.2[2,2]
      p_value[j] <- temp.2[2,4]
      pow[j] <- ifelse(p_value[j]<0.05, 1, 0)
      ## if the treatment coefficient=0 was rejected, then symbol as 1
      lower.bi[j] <-coefci(logi.bi, vcov=vcovHC(logi.bi, type = "HC4"))[2,1]
## calculate CI based on robust SE
      upper.bi[j] <-coefci(logi.bi, vcov=vcovHC(logi.bi, type = "HC4"))[2,2]
      converge.p.bi[j]      <-      ifelse(parameter[i,3]>lower.bi[j]      &
parameter[i,3]<upper.bi[j], 1, 0)
      bias.bi[j] <-est.bi[j]-parameter[i,3]
    }
    Bias[i] <- mean(bias.bi)
    ASE[i] <- mean(est.bi.se)
    ESE[i] <- sd(est.bi)
    CP[i] <- sum(converge.p.bi)/1000
    power_bi[i] <- sum(pow)/1000
    median_est[i] <- median(est.bi)
    relative_bias[i] <- mean(bias.bi)/(parameter[i,3])
  }
  return(cbind(Bias,ASE,ESE,CP,power_bi,median_est,relative_bias))
}
LG_LM_Bi <- data.frame(present.lg_lm.bi(all.data_bi,new.combi))
write.csv(LG_LM_Bi, file="lg_lm_bi.csv")
#########################################################################
## approach 3:  3 stage least square
present.lg_lm_lm.bi<- function(all.data,parameter){
        Bias <- c()
        ASE <- c()
        ESE <- c()
        CP <- c()
        median_est <- c()  # output the median of point estimate
        relative_bias <- c()
```

```
        #calculate the relative bias, relative_bias=mean(bias)/effect
        power_bi <- c()  #calcualte the power
        for (i in 1:72) {
                est.bi <- c()    #point estimate
                est.bi.se <- c()   #Robust SE for point estimate
                lower.bi <- c()    # lower 95% CI
                upper.bi <- c()    # upper 95% CI
                bias.bi <- c()         # bias for each point estimate
                converge.p.bi <- c()
          # converge symbol: 1, true treatment effect falls in 95% CI
                p_value <- c()  # extract the p value for each treatment term
                pow <- c()
   # power indicator: equals if the p value of treatment term smaller than 0.05
                for (j in 1:1000){
                        data_1 <- all.data[[i]][[j]]
                        # define outcome, instrument, endogenous variable
                        Y1 <- data_1$binary.y
                        Y2 <- data_1$comply
                        X2 <- data_1$treat
                        # 1st stage fit logistic regression
                        logisreg.bi   <-   glm(Y2~X2,family   =   binomial(link
='logit'), data =data_1,
                                               control = list(maxit = 50))
                        phat.bi<- predict(logisreg.bi, type="response")
                        # 2nd stage linear regression Y2 on X2 hat
                        lg_lm_lm.bi_2 <- lm(Y2~phat.bi)
                        Y2_hat <- fitted(lg_lm_lm.bi_2)

                        # 3rd stage linear regression Y1 on Y2_hat
                        lg_lm_lm.bi_3 <- lm(Y1~Y2_hat)
                        temp.3 <- coeftest(lg_lm_lm.bi_3,
vcov=vcovHC(lg_lm_lm.bi_3, type = "HC4"))
```

```
                        est.bi[j] <- temp.3[2,1]
                        est.bi.se[j] <- temp.3[2,2]
                        p_value[j] <- temp.3[2,4]
                        pow[j] <- ifelse(p_value[j]<0.05, 1, 0)
## if the treatment coefficient=0 was rejected, then symbol as 1
                        lower.bi[j]<-coefci(lg_lm_lm.bi_3,
vcov=vcovHC(lg_lm_lm.bi_3, type = "HC4"))[2,1]
## calculate CI based on robust SE
                        upper.bi[j]<-coefci(lg_lm_lm.bi_3,
vcov=vcovHC(lg_lm_lm.bi_3, type = "HC4"))[2,2]

                        converge.p.bi[j] <- ifelse(parameter[i,3]>lower.bi[j]
& parameter[i,3]<upper.bi[j], 1, 0)
                        bias.bi[j] <-est.bi[j]-parameter[i,3]

                }
                Bias[i] <- mean(bias.bi)
                ASE[i] <- mean(est.bi.se)
                ESE[i] <- sd(est.bi)
                CP[i] <- sum(converge.p.bi)/1000
                power_bi[i] <- sum(pow)/1000
                median_est[i] <- median(est.bi)
                relative_bias[i] <- mean(bias.bi)/(parameter[i,3])
        }
        return(cbind(Bias,ASE,ESE,CP,power_bi,median_est,relative_bias))
}
LG_LM_LM_Bi <- data.frame(present.lg_lm_lm.bi(all.data_bi,new.combi))
write.csv(LG_LM_LM_Bi, file="lg_lm_lm_bi.csv")

################################## PLOT ########################################

require(foreign); require(haven) # to import spss/stata files
```

```
require(readxl); # read excel files
require(tidyverse); require(forcats); # for efficiently manipulating datasets
require(compareGroups) # for descriptive stats
require(reshape2)  # for reshaping data & plotting
require(ggplot2);  require(ggthemes);  require(ggrepel);  require(ggpubr);  #
require(ggeffects) # for graphics
require(gtable);  require(gridExtra);  require(grid);  require(cowplot)  #  for
arranging plots
require(stringr) # for text processing

library(ReporteRs) # for exporting plot to power point
pd1 <- read_csv("combined_simu_1.csv") #6 x 34
pd2 <- read_csv("combined_simu_2.csv") #72 x 27
pd3 <- read_csv("combined_simu_3.csv") #24 x 32

# function to save plot to ppt
exp2ppt <- function(plot,file,lab="") {
  mydoc <- pptx()
  mydoc <- addSlide(mydoc,"Content with Caption")
  # add caption
  mydoc <- addTitle(mydoc,lab)
  # add a plot into mydoc
  mydoc <- addPlot(mydoc,fun=print,x=plot)
  writeDoc(mydoc,file)
}
## First clean up variable names
names(pd3)
pd3_c <- pd3 %>%
  janitor::clean_names() #24 x 32
names(pd3)

## Convert to long format for ggplot2 plotting
```

```
pd3_l <- pd3_c %>%
  select(-x1) %>%
  gather(key=metric, value=value
         ,-n,-pc,-risk_difference
         ) #672 x 5

## remove empty value
pd3_l <- pd3_l %>%
  filter(!is.na(value)) #576 x 5

## convert value to numeric, set inf to NA
pd3_l <- pd3_l %>%
  mutate(value=as.numeric(value)
         ,value=ifelse(value==Inf, NA, value)
         )

## Create a variable to identify method
metriclist <- c("bias","relative_bias","ase","ese","cp","power_bi")
pd3_l <- pd3_l %>%
  mutate(method=case_when(
    metric %in% metriclist ~ "LM-LM"
    ,metric %in% paste0(metriclist,"_1") ~ "LM-LM-X"
    ,metric %in% paste0(metriclist,"_2") ~ "LG-LM-X"
    ,metric %in% paste0(metriclist,"_3") ~ "LG-LM-LM-X"
                          )
         ) #576 x 6
## Standardize the names of the metric
pd3_l <- pd3_l %>%
  mutate(metric=str_replace_all(metric,"[_123]","")) #576 x 6
with(pd3_l,table(metric))

## Recode to factor for nice labels:
```

```
rdlev <- pd3_l$risk_difference %>%
  unique() %>%
  sort()
nlev <- pd3_l$n %>%
  unique() %>%
  sort()
pclev <- pd3_l$pc %>%
  unique() %>%
  sort()
methodlev <- c("LM-LM"
               ,"LM-LM-X"
               ,"LG-LM-X"
               ,"LG-LM-LM-X"
               )

dsp <- pd3_l %>%
  mutate(n_f=factor(n
                    ,levels=nlev
                    ,labels=paste0("n=",nlev)
                    )
         ,pc_f=factor(pc
                      ,levels=pclev
                      ,labels=c("Low\ncompliance"
                                ,"High\ncompliance"
                                )
                      )
         ,method_f=factor(method
                          ,levels=methodlev
                          ,labels=methodlev
                          )
         )
## Plot relative bias (line plot)
```

```
fn <- "simulation3_ARB" #filename
yl <- "Absolute Relative Bias"
xl <- "Risk Difference"
pl <- dsp %>%
  filter(metric=="relativebias") %>%
  ggplot(aes(x=risk_difference,y=abs(value),group=method_f
             )
         ) +
  geom_point(aes(shape=method_f)
             ,alpha=0.7
             ,size=2
             # ,position=position_dodge(width=0.05)
             ) +
  scale_y_continuous(labels = function(x) sprintf("%4.1f%%",x*100)) +
  scale_x_continuous(labels = function(x) sprintf("%3.0f%%",x*100)
                     ,breaks=rdlev
                     ) +
  geom_line(aes(color=method_f)) +
  facet_grid(pc_f~n_f) +
  theme_pubr() +
  theme(legend.title = element_blank()
        # ,legend.direction="vertical"
        # ,legend.position ="right"
        ,legend.text = element_text(size=rel(1.2))
        ,axis.title=element_text(size=rel(1.25))
        ,axis.text = element_text(size=rel(1))
        ,strip.text = element_text(size=rel(1.2))
        ,panel.spacing = unit(1.5, "lines")
        ) +
  labs(x=xl,y=yl)
### Export to powerpoint
exp2ppt(pl,paste0(fn,".pptx"),fn)
```

BIBLIOGRAPHY

1. Perkin, M.R., et al., *Randomized Trial of Introduction of Allergenic Foods in Breast-Fed Infants.* N Engl J Med, 2016. **374**(18): p. 1733-43.
2. Ranganathan, P., C.S. Pramesh, and R. Aggarwal, *Common pitfalls in statistical analysis: Intention-to-treat versus per-protocol analysis.* Perspect Clin Res, 2016. **7**(3): p. 144-6.
3. Jo, B., *Model misspecification sensitivity analysis in estimating causal effects of interventions with non-compliance.* Stat Med, 2002. **21**(21): p. 3161-81.
4. Angrist, J.D., G.W. Imbens, and D.B. Rubin, *Identification of Causal Effects Using Instrumental Variables.* Journal of the American Statistical Association, 1996. **91**(434): p. 444-455.
5. Sussman, J.B. and R.A. Hayward, *An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials.* BMJ, 2010. **340**: p. c2073.
6. Frangakis, C.E. and D.B. Rubin, *Principal stratification in causal inference.* Biometrics, 2002. **58**(1): p. 21-9.
7. Stuart, E.A. and B. Jo, *Assessing the sensitivity of methods for estimating principal causal effects.* Stat Methods Med Res, 2015. **24**(6): p. 657-74.
8. Ertefaie, A., et al., *A tutorial on the use of instrumental variables in pharmacoepidemiology.* Pharmacoepidemiology and Drug Safety, 2017. **26**(4): p. 357-367.
9. Hertogh, E.M., et al., *Noncompliance in lifestyle intervention studies: the instrumental variable method provides insight into the bias.* Journal of Clinical Epidemiology, 2010. **63**(8): p. 900-906.
10. Ye, C., et al., *Estimating treatment effects in randomised controlled trials with non-compliance: a simulation study.* BMJ Open, 2014. **4**(6): p. e005362.
11. Palmer, T.M., et al., *Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses.* Int J Epidemiol, 2008. **37**(5): p. 1161-8.
12. Greenland, S., *An introduction to instrumental variables for epidemiologists.* Int J Epidemiol, 2000. **29**(4): p. 722-9.
13. Nagelkerke, N., et al., *Estimating treatment effects in randomized clinical trials in the presence of non-compliance.* Stat Med, 2000. **19**(14): p. 1849-64.
14. Palmer, T.M., et al., *Nonparametric bounds for the causal effect in a binary instrumental-variable model.* Stata Journal, 2011. **11**(3): p. 345-367.
15. White, I.R., *Uses and limitations of randomization-based efficacy estimators.* Stat Methods Med Res, 2005. **14**(4): p. 327-47.
16. Cheung, Y.B., *A modified least-squares regression approach to the estimation of risk difference.* Am J Epidemiol, 2007. **166**(11): p. 1337-44.
17. Angrist, J.D., *Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice.* Journal of Business & Economic Statistics, 2001. **19**(1): p. 2-16.
18. Rassen, J.A., et al., *Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes.* Am J Epidemiol, 2009. **169**(3): p. 273-84.

19. Rubin, D.B., *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.* Journal of Educational Psychology, 1974. **66**(5): p. 688-701.
20. Splawa-Neyman, J., *On the application of probability theor ot agricultural experiments.* Statistical Science, 1923. **5**: p. 463-480.
21. Imai, K., B. Jo, and E.A. Stuart, *Commentary: Using Potential Outcomes to Understand Causal Mediation Analysis.* Multivariate Behavioral Research, 2011. **46**(5): p. 861-873.
22. Gupta, S.K., *Intention-to-treat concept: A review.* Perspect Clin Res, 2011. **2**(3): p. 109-12.
23. Adewuyi, T.E., G. MacLennan, and J.A. Cook, *Non-compliance with randomised allocation and missing outcome data in randomised controlled trials evaluating surgical interventions: a systematic review.* BMC Res Notes, 2015. **8**: p. 403.
24. Piaggio, G., et al., *Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement.* JAMA, 2012. **308**(24): p. 2594-604.
25. Kim, M.Y., *Using the instrumental variables estimator to analyze noninferiority trials with noncompliance.* J Biopharm Stat, 2010. **20**(4): p. 745-58.
26. Little, R.J., Q. Long, and X. Lin, *A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance.* Biometrics, 2009. **65**(2): p. 640-9.
27. Bellamy, S.L., J.Y. Lin, and T.R. Ten Have, *An introduction to causal modeling in clinical trials.* Clin Trials, 2007. **4**(1): p. 58-73.
28. Hernan, M.A. and J.M. Robins, *Instruments for causal inference: an epidemiologist's dream?* Epidemiology, 2006. **17**(4): p. 360-72.
29. Ten Have, T.R., et al., *Intent-to-Treat vs. Non-Intent-to-Treat Analyses under Treatment Non-Adherence in Mental Health Randomized Trials.* Psychiatr Ann, 2008. **38**(12): p. 772-783.
30. Carroll, R.J., D. Ruppert, and L.A. Stefanski, *Measurement error in nonlinear models*. Monographs on statistics and applied probability. 1998, Boca Raton: Chapman & Hall/CRC. xxiv, 305 p.
31. Rosner, B., D. Spiegelman, and W.C. Willett, *Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error.* Am J Epidemiol, 1990. **132**(4): p. 734-45.
32. JM, W., *Introductory Econometrics: A Modern Approach.* 2006.
33. Angrist, J.D., *Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors.* Journal of Business & Economic Statistics, 2001. **19**(1): p. 2-28.
34. Terza, J.V., A. Basu, and P.J. Rathouz, *Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling.* J Health Econ, 2008. **27**(3): p. 531-43.
35. Wooldridge, J.M., *Econometric analysis of cross section and panel data*. 2002, Cambridge, Mass.: MIT Press. xxi, 752 p.
36. Montgomery, D.C., E.A. Peck, and G.G. Vining, *Introduction to linear regression analysis*. 5th ed. Wiley series in probability and statistics. 2012, Hoboken, NJ: Wiley. xvi, 645 p.
37. Angrist, J.D., *Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors: Simple Strategies for Empirical Practice.* Journal of Business and Economic Statistics, 2001. **19**(1,Jan): p. 2-15.

38. Jo, B. and E.A. Stuart, *On the use of propensity scores in principal causal effect estimation.* Stat Med, 2009. **28**(23): p. 2857-75.
39. RJ, B. and T. DA, *A comparative study of instrumental vriables estimators for nonlinear simultaneous models.* Journal of the American Statistical Association, 1981. **76**(376): p. 988-995.