Performance of Logistic Regression, Propensity Scores, and Instrumental Variable for Estimating Their True Target Odds Ratios in the Absence and Presence of Unmeasured Confounder

Tianqi Yu
West China Hospital of Sichuan University
Chengyu Ke
Southern Methodist University
Wentao Xu
West China Hospital of Sichuan University
Jing Li (lijing68@hotmail.com)
West China Hospital of Sichuan University

Research Article

Keywords: bias, unmeasured confounder, logistic model, propensity score, instrumental variable

Posted Date: December 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-133868/v1

RESEARCH

Performance of logistic regression, propensity scores, and instrumental variable for estimating their true target odds ratios in the absence and presence of unmeasured confounder

Tianqi. Yu[1], Chengyu. Ke[2], Wentao. Xu[1] and Jing. Li[1*]

*Correspondence: lijing68@hotmail.com
[1] Department of Evidence-Based Medicine and Clinical Epidemiology, Sichuan University West China Hospital, Chengdu, China
Full list of author information is available at the end of the article

Abstract

Background: A lot of studies have compared the ability of statistical methods to control for confounding. However, a majority of studies mistakenly assumed these methods estimate the same effect. The aim of this study was to use Monte Carlo simulations to compare logistic regression, propensity scores and instrumental variable analysis for estimating their true target odds ratios in terms of bias and precision in the absence and presence of unmeasured confounder.

Methods: We established the formula allowing us to compute the true odds ratio of each method. We varied the instrument's strength and the unmeasured confounder to cover a large range of scenarios in the simulation study. We then use logistic regression, propensity score matching, propensity score adjustment and two-stage residual inclusion to obtain estimated odds ratios in each scenario.

Results: In the absence of unmeasured confounder, instrumental variable without direct effect on the outcome could produce unbiased estimates as propensity score did, but the mean square errors of instrumental variable were greater. When unmeasured confounder existed, no other method could produce unbiased estimation except instrumental variable, provided that the proposed instrument is not directly related to the outcome. Using the defined instrument, which affected the outcome directly, resulted in positive biased estimation of the treatment effect and this bias was greater compared to that from other methods.

Conclusions: Overall, with good implementation, instrumental variable can lead to unbiased results. However, the bias caused by violating the required assumptions of instrumental variable can overweigh the positive effect of its ability to control for unmeasured confounder.

Keywords: bias; unmeasured confounder; logistic model; propensity score; instrumental variable

Background

Randomized controlled trials are considered as the gold standard for clinical evaluation but are difficult to conduct because of many practical considerations. Well-designed observational studies are also helpful to enhance and confirm the findings of randomized studies, although they cannot be regarded as a replacement for RCT [1].

Logistic regression is an alternative method in observational studies for dichotomous outcomes. It is often used to reduce the bias caused by measured confounders. However, when modeling, if too many variables need to be included in a model relative to the number of observed events, the estimates from these models can be incorrect [2]. To address these limits, Rosenbaum and Rubin [3] proposed "propensity scores"(PS), which is the conditional probability of a subject receiving a particular treatment given the set of confounders. It allows simultaneously control for multiple variables in situations where conventional multivariate models might perform badly, owing to the insufficient observed events. A central concern with observational data, however, is bias by unmeasured or uncontrolled confounding which might explain away the observed association between the treatment and the outcome [4].

Instrumental variable (IV) analysis is an approach to obtain unbiased estimates even in the presence of unmeasured confounders, provided that certain assumptions are met. An valid IV should satisfy the following assumptions: it is associated with treatment; it has no direct effect on the outcome (exclusion restriction); and it is independent of all (unmeasured) confounders of the treatment–outcome relationship. In practice, however, it is hard to meet all these assumptions, and some of which are even not testable. Consequently, we can hardly be certain that the proposed instrument is valid to adjust for unmeasured confounding [5].

In recent years, a lot of studies have compared the ability of statistical methods to control for confoundings, but these studies have some limitations. First, in most studies, researchers mistakenly assumed that these methods estimate the same effect, while in fact they do not. For example, logistic regression intends to estimate the conditional treatment effect; propensity score matching allows us to estimate the average treatment effet in treated; propensity score adjustment intends to estimate the average treatment effect; and IV aims to estimate the true compiler average causal effect. As these studies did not distinguish between these different effects when compared the performance of different methods, we suspect whether the results of these studies can be a true representation of the performance of them. Second, it has been argued that IV and logistic regression or propensity score methods are applicable to different scenarios. The former is applicable to the presence of unmeasured confounders, while the latter two are applicable to the absence of unmeasured confounders in the study. However, we should bear in mind that we can hardly be sure if there are any unmeasured confounders in the study in practice; In addition, since some of the assumptions required for IV are not directly testable, it is hard to tell whether all of the assumptions for IV are met. Therefore, we have doubts about whether IV outperforms the other two methods in the presence of unmeasured confounders.

Accordingly, the purpose of this study was to compare the performances of different methods for estimating their true target treatment effects when odds ratio is used as a measure of treatment effect by using Monte Carlo simulations. We examined three kinds of different methods for estimating treatment effects: logistic regression, propensity score method and instrumental variable analysis; We calculated two values to assess their performances: bias and mean squared error; Finally, we sought to investigate the consequences associated with instruments of different strengths and to compare IV with other methods when unmeasured confounder exists and the exclusion restriction is violated.

Methods

Definitions

Eleven variables related to the treatment and the outcome were considered; N denotes the sample size; Treatment selection variable $Z_i(i = 1 \dots N)$ depends on$(X_1, \dots, X_7)$and R_i, while outcome variable Y_i depends on $(X_4, \dots, X_{10})$ and R_i; R_i is a defined instrument. In the current study, we defined the $Z_i(1)$,$Z_i(0)$ which was a treatment status conditional on the subject having been assigned to treatment ($R_i = 1$) and a treatment status conditional on the subject having been assigned to control ($R_i = 0$). We also defined two potential probabilities of occurrence of an event for each subject: a probability conditional on the subject having been treated $P_i(1)$ and a probability conditional on the subject having not been treated $P_i(0)$.

Data-generation process

We generated eleven independent covariates x_1–x_{10} and R_i for each of N subjects. Each of the x_1–x_{10} covariates was assumed to have a Bernoulli distribution with parameter 0.5. R_i was assumed to have a Bernoulli distribution with parameter 0.7.

Treatment status

We then generated a treatment status for each of N subjects by

$$Z_i \sim Bernoulli(p_{i,treatment})$$

where $logit(p_{i,treatment}) = \alpha_0 + \alpha_1 x_{i,1} + \dots + \alpha_7 x_{i,7} + \alpha_r R_i$.

Outcomes

For each subject we randomly generated a dichotomous outcome (1=occurrence of an event; 0=absence of an event) using a logistic model:

$$Y_i \sim Bernoulli(p_{i,outcome})$$

where $logit(p_{i,outcome}) = \beta_0 + \beta_4 x_{i,4} + \dots + \beta_{10} x_{i,10} + \beta_{treat} Z_i + \beta_r R_i$

Parameter values for data generation

In the data generation process, the regression coefficients took the values displyed in Table 1.

Table 1 Data Generation Parameters for the Simulation Study

Variable	Value
α_1, β_4	ln(1.5)
$\alpha_2, \alpha_3, \alpha_4, \beta_5, \beta_6, \beta_7$	ln(2.5)
$\alpha_5, \alpha_6, \alpha_7, \beta_8, \beta_9, \beta_{10}$	ln(5)
β_{treat}	ln(1), ln(1.5), ln(2.5), ln(5)

Allowing α_r to be $\ln(2)$, $\ln(5)$ and $\ln(10)$, indicating a weak, moderate and strong instrument respectively; β_r was set to be $\ln(1.2)$ and $\ln(1)$, implying the proposed instrument had a weak and had no direct association with the outcome.

We wanted to have approximately 50 per cent of the subjects who were assigned to the control to be exposed to treatment. The value of α_0 was set to -4 so that the treatment would be assigned to approximately half of these subjects.

We want the outcome occurs for approximately 50 per cent of the untreated subjects. Then for $\ln(1)$, the value of β_0 was set to -4; for $\ln(1.2)$, the value of β_0 was set to -4.1.

The true odds ratos of each method

Using the data-generating process, first,we randomly generated a treatment status for each subject that was conditional on the subject's baseline covariates and the proposed instrument. We then used the second data-generating process to randomly generate an outcome that was conditional on both the actual treatment assigned and on the subject's baseline covariates. We also generated two potential treatment status for each subject:$Z_i(1)$,$Z_i(0)$ and two potential probabilities: $P_i(1)$,$P_i(0)$ .These two potential treatment status and outcomes were used to determine the true treatment effects on the odds ratio.

The true conditional treatment effects(CTE)
the ture OR of CTE is defined as:

$$OR_{CTE} = e^{\beta_{treat}}$$

The true average treatment effect (ATE)
the ture OR of ATE is calculated by:

$$OR_{ATE} = \frac{\frac{\bar{P}_i(1)}{1-\bar{P}_i(1)}}{\frac{\bar{P}_i(0)}{1-\bar{P}_i(0)}}$$

$$\bar{P}_i(1) = \frac{1}{N}\sum_{i=1}^{N} P_i(1)$$

$$\bar{P}_i(0) = \frac{1}{N}\sum_{i=0}^{N} P_i(0)$$

The true average treatment effect in treated (ATT)
the true OR of ATT is calculated by:

$$OR_{ATT} = \frac{\frac{\bar{P}_i(1)}{1-\bar{P}_i(1)}}{\frac{\bar{P}_i(0)}{1-\bar{P}_i(0)}}$$

$$\bar{P}_i(1) = \frac{1}{N_{Z=1}}\sum_{i=1}^{N_{Z=1}} P_i(1)$$

$$\bar{P}_i(0) = \frac{1}{N_{Z=1}} \sum_{i=0}^{N_{Z=1}} P_i(0)$$

where $N_{Z=1}$ denotes the number of population who actually receive the treatment.

The true compiler average causal effect (CACE)
the true OR of CACE is calculated by:

$$OR_{CACE} = \frac{\dfrac{\bar{P}_i(1)}{1-\bar{P}_i(1)}}{\dfrac{\bar{P}_i(0)}{1-\bar{P}_i(0)}}$$

$$\bar{P}_i(1) = \frac{1}{N_c} \sum_{i=1}^{N_c} P_i(1)$$

$$\bar{P}_i(0) = \frac{1}{N_c} \sum_{i=0}^{N_c} P_i(0)$$

where N_c denotes the number of the compilers population. The compilers are subjects who adhere to the assignment of treatment but do not take it when not assigned to it ($Z_i(1) = 1$ and $Z_i(0) = 0$).

The mean of OR_{ATE}, OR_{ATT}, OR_{CACE} are determined across the simulated datasets and will serve as the true target OR.

Estimating the treatment effects

Two different scenarios were considered:

Scenario 1 no unknown confounder exists, which means all $(X_1, \ldots, X_{10})$ variables can be involved in when fiting models.

Scenario 2 X_5 was viewed as an unknown confounder in this scenario, which means it would not be involved in when fitting models.

Logistic regression
We used logistic regression to estimate the conditional treatment effect(CTE). Two different specifications were considered. In the first specification, we used logistic regression model (logistic model 1) to regress outcomes on the treatment status and four baseline covariates that affected the treatment status and the outcome$(X_4, \ldots, X_7)$. The second (logistic model 2) controlled for all the covariates related to the outcome $(X_4, \ldots, X_{10})$. In scenario 2, X_5 was not be involved in the logistic model in both specifications.

Propensity score matching
We used propensity score matching to create a matched sample of treated and untreated patients. For each subject, we computed the logit of the estimated propensity score by regressing treatment status on the seven baseline covariates

$(X_4, \ldots, X_{10})$ in scenario 1 and six baseline covariates$(X_4, X_6, \ldots, X_{10})$ in scenario 2. We employed all variables related to the outcome as it has been shown to lead to better estimation compared to selecting only those variables that affect treatment status [6]. We then used a greedy-matching algorithm to match subjects with calipers of 0.2 standard deviations of the logit of the estimated propensity score. We thus obtained the estimated odds ratio of ATT from a matched-pairs design.

Adjustment using the propensity score

Covariate adjustment using the propensity score was commonly used form of the propensity score in the clinical literature [1, 7],and it has shown perform as well as PS matching [8]. In this method, the propensity score (on the probability scale) and a variable denoting treatment status, are entered in the logistic regression model. The estimated OR of ATE is obtained from the nature exponential of regression coefficient for treatment status.

Two stage residual inclusion(2SRI)

Terza et al.[9] showed the consistency and the superiority of the 2SRI method, they recommended applied researchers to employ 2SRI estimation when they are trying to address endogeneity in nonlinear models. In our study for dealing with binary treatment status and outcomes, logistic regression is used for both the first and second stages of the 2SRI procedure. In the first stage of 2SRI, regression of treatment received on the treatment assignment R_i as an instrument, and the results are used to generate predicted values for calculating the residual which is $\hat{X}_u = Z_i - \hat{p}_{i,treatment}$. In the second-stage regression, the first-stage residuals are included as additional regressors in second-stage estimation:

$$logit(\hat{p}_{i,outcome}) = \hat{\beta}_0 + \hat{\beta}_4 x_{i,4} + \ldots + \hat{\beta}_7 x_{i,7} + \hat{\beta}_{treat} Z_i + \hat{\beta}_u \hat{X}_u$$

and then the $e^{\hat{\beta}_{treat}}$ was viewed as the estimated OR of CACE.

Monte Carlo simulations

For each of the combinations of α_r,β_r and β_{treat}, we randomly generated 1000 datasets using the data-generating process. Each randomly generated dataset consisted of 10 000 subjects. Using each of the 1000 datasets, we estimated the CTE, ATE, ATT and CACE on the odds ratio by using each method. We then determined bias and mean square error (MSE) on the odds scale as:

$$Bias = \overline{OR} - OR_{true}$$

$$MSE = \frac{\sum_{i=1}^{1000} (\widehat{OR} - OR_{true})^2}{1000}$$

where $\widehat{OR}$ is the estimated odds ratio of each method in one simulated dataset; $\overline{OR}$ is the average of the $\widehat{OR}$ over the 1000 simulated datasets; OR_{true} is the true target treatment effect on the odds scale. Simulations were conducting by using R, version 4.0.3, software.

Results

In this section we examined the bias and the MSE in the estimated odds ratio when there is no and there is unobservable confounder existing. The estimated values are displayed in Table 2. and Table 3.. Figure 1. and Figure 2. contain the bias (in the odds scale) for each of the statistical methods used and for each of the true target treatment effects. Figure 3. and Figure 4. show the MSE (in the odds scale) for each of the statistical methods used.

Among Figure 1-4., A, B and C shows the bias for each method when the defined IV had no association with the outcome, whereas D, E and F shows that when the defined IV had a direct effect on the outcome. In addition, A and D , B and E, C and F employed the weak, moderate and strong IV respectively.

For comparative purposes, the initial crude or unadjusted estimate of the treatment effect was also calculated.

Table 2 Ture and Estimated Odds Ritios of Different Methods Under Different Parameter Values in the Absence of Unmeasurable Confounders

OR	β_r	α_r	OR_{ATE}	OR_{ATT}	OR_{CACE}	Crude	lg1	lg2	PSA	PSM	IV
1.00	ln(1)	ln(2)	1.00	1.00	1.00	1.77	1.00	1.00	1.00	1.00	1.05
		ln(5)	1.00	1.00	1.00	1.74	1.00	1.00	1.00	1.00	1.02
		ln(10)	1.00	1.00	1.00	1.69	1.00	1.00	1.00	1.00	1.01
	ln(1.2)	ln(2)	1.00	1.00	1.00	1.80	1.02	1.03	1.02	1.02	2.94
		ln(5)	1.00	1.00	1.00	1.79	1.04	1.06	1.04	1.04	1.74
		ln(10)	1.00	1.00	1.00	1.77	1.06	1.08	1.06	1.06	1.53
1.50	ln(1)	ln(2)	1.31	1.32	1.32	2.34	1.34	1.50	1.35	1.33	1.38
		ln(5)	1.31	1.32	1.32	2.30	1.34	1.50	1.35	1.32	1.35
		ln(10)	1.31	1.32	1.32	2.23	1.34	1.50	1.35	1.32	1.35
	ln(1.2)	ln(2)	1.31	1.32	1.32	2.37	1.37	1.54	1.37	1.35	3.69
		ln(5)	1.31	1.32	1.32	2.36	1.39	1.59	1.40	1.38	2.32
		ln(10)	1.31	1.32	1.32	2.33	1.42	1.63	1.43	1.40	2.04
2.50	ln(1)	ln(2)	1.86	1.89	1.88	3.35	1.95	2.51	1.96	1.89	1.93
		ln(5)	1.86	1.88	1.88	3.27	1.95	2.51	1.96	1.89	1.93
		ln(10)	1.86	1.87	1.88	3.18	1.95	2.51	1.97	1.89	1.95
	ln(1.2)	ln(2)	1.86	1.89	1.88	3.40	1.98	2.57	2.00	1.93	5.44
		ln(5)	1.86	1.88	1.88	3.38	2.03	2.65	2.04	1.97	3.39
		ln(10)	1.86	1.88	1.88	3.32	2.06	2.71	2.08	1.99	2.91
5.00	ln(1)	ln(2)	3.01	3.11	3.08	5.51	3.27	5.00	3.29	3.10	3.14
		ln(5)	3.01	3.08	3.07	5.37	3.26	5.03	3.30	3.08	3.19
		ln(10)	3.01	3.06	3.06	5.20	3.26	5.03	3.31	3.07	3.24
	ln(1.2)	ln(2)	3.01	3.11	3.08	5.60	3.33	5.14	3.36	3.17	9.27
		ln(5)	3.01	3.09	3.07	5.54	3.39	5.29	3.44	3.21	5.55
		ln(10)	3.01	3.07	3.06	5.45	3.45	5.44	3.51	3.24	4.82

Abbreviations: lg1, logistic model 1 (including the covariates related to both treatment and outcome); lg2, logistic model 2 (including the all covariates related to outcome); PSA, propensity score adjustment; PSM, propensity score matching.

Bias of each method when no unmeasured confounder exists

For each of the true odds ratios of ATE, the crude estimate was biased upwards.

In examining the estimated effects of conditional treatment when logistic regression was used, the estimated treatment effects were biased towards the null when only the true confounders (related to both the treatment and the outcome) were included in the model (logistic model 1),and as the true odds ratio was greater than one, it resulted in greater bias; the estimated treatment effects were almost unbiased when all variables related to the outcome were included in the model (logistic model 2).

In examining the OR of ATE when covariate adjustment using the estimated propensity score was used, the estimated effect was slightly positively biased, except

Table 3 Ture and Estimated Odds Ritios of Different Methods Under Different Parameter Values in the Presence of Unmeasurable Confounders

OR	β_r	α_r	OR_{ATE}	OR_{ATT}	OR_{CACE}	Crude	lg1	lg2	PSA	PSM	IV
1.00	ln(1)	ln(2)	1.00	1.00	1.00	1.77	1.25	1.36	1.25	1.24	1.06
		ln(5)	1.00	1.00	1.00	1.74	1.23	1.33	1.24	1.23	0.96
		ln(10)	1.00	1.00	1.00	1.70	1.21	1.30	1.22	1.21	0.95
	ln(1.2)	ln(2)	1.00	1.00	1.00	1.80	1.27	1.38	1.27	1.26	3.57
		ln(5)	1.00	1.00	1.00	1.80	1.28	1.40	1.28	1.27	1.75
		ln(10)	1.00	1.00	1.00	1.77	1.28	1.40	1.28	1.27	1.48
1.50	ln(1)	ln(2)	1.31	1.32	1.32	2.35	1.67	2.01	1.67	1.65	1.35
		ln(5)	1.31	1.32	1.32	2.29	1.64	1.97	1.65	1.62	1.27
		ln(10)	1.31	1.32	1.32	2.23	1.61	1.92	1.62	1.59	1.27
	ln(1.2)	ln(2)	1.31	1.32	1.32	2.37	1.69	2.05	1.69	1.67	4.80
		ln(5)	1.31	1.32	1.32	2.36	1.70	2.06	1.71	1.67	2.30
		ln(10)	1.31	1.32	1.32	2.34	1.71	2.07	1.72	1.68	1.97
2.50	ln(1)	ln(2)	1.86	1.89	1.88	3.34	2.40	3.29	2.41	2.34	1.93
		ln(5)	1.86	1.88	1.88	3.27	2.37	3.23	2.38	2.30	1.81
		ln(10)	1.86	1.87	1.88	3.18	2.33	3.14	2.34	2.26	1.82
	ln(1.2)	ln(2)	1.86	1.89	1.88	3.39	2.44	3.36	2.45	2.38	6.91
		ln(5)	1.86	1.88	1.88	3.37	2.45	3.38	2.46	2.38	3.32
		ln(10)	1.86	1.88	1.88	3.32	2.45	3.38	2.47	2.38	2.81
5.00	ln(1)	ln(2)	3.01	3.11	3.08	5.53	4.01	6.44	4.03	3.85	3.08
		ln(5)	3.01	3.08	3.06	5.37	3.93	6.31	3.96	3.75	2.98
		ln(10)	3.00	3.06	3.06	5.18	3.85	6.15	3.89	3.65	2.97
	ln(1.2)	ln(2)	3.01	3.11	3.08	5.58	4.06	6.54	4.08	3.90	11.67
		ln(5)	3.01	3.09	3.07	5.55	4.08	6.64	4.11	3.90	5.41
		ln(10)	3.01	3.07	3.06	5.45	4.07	6.62	4.11	3.87	4.57

Abbreviations: lg1, logistic model 1 (including the covariates related to both treatment and outcome); lg2, logistic model 2 (including the all covariates related to outcome); PSA, propensity score adjustment; PSM, propensity score matching.

when the true conditional odds ratio was equal to one. However, when estimating the OR of ATT by using propensity score matching, the estimated effect was nearly unbiased and it showed the best performace in each condition.

Finally, we examined the OR of CACE when instrumental variable analysis was used. Using the defined instrument which had no direct association with the outcome resulted in at most negligible bias(Figure 1A-C). In contrast, using the defined instrument, which affected the outcome directly, resulted in positive biased estimation and this bias was greater compared to that for other methods. Further, when IV was weak, the bias it caused was even greater than the crude estimates(Figure 1D). It is important to note that employing an strong IV can still lead to significantly biased results when the instrument had a direct effect on the outcome, even though this effect is weak (Figure 1F).

Bias of each method when unmeasured confounder exists

Different results were observed in this scenario, although the bias of the crude treatment effect was always positive and great.

As shown in Figure 2., when the true conditional odds ratio was less than 5, logistic model 1 resulted in less biased results compared to logistic model 2. In addition, when the true conditional OR was equal or greater than 2.5, logistic model 2 always lead to positive bias, while the estimated treatment effect of logistic model 1 biased towards the null .

Each of the two propensity score methods resulted in positive biased estimation and the bias increased as the true treatment effect increased. As expected, the bias of IV analysis was negligibly different from zero when there is no directly association between the instrument and the outcome. The situation, however, would

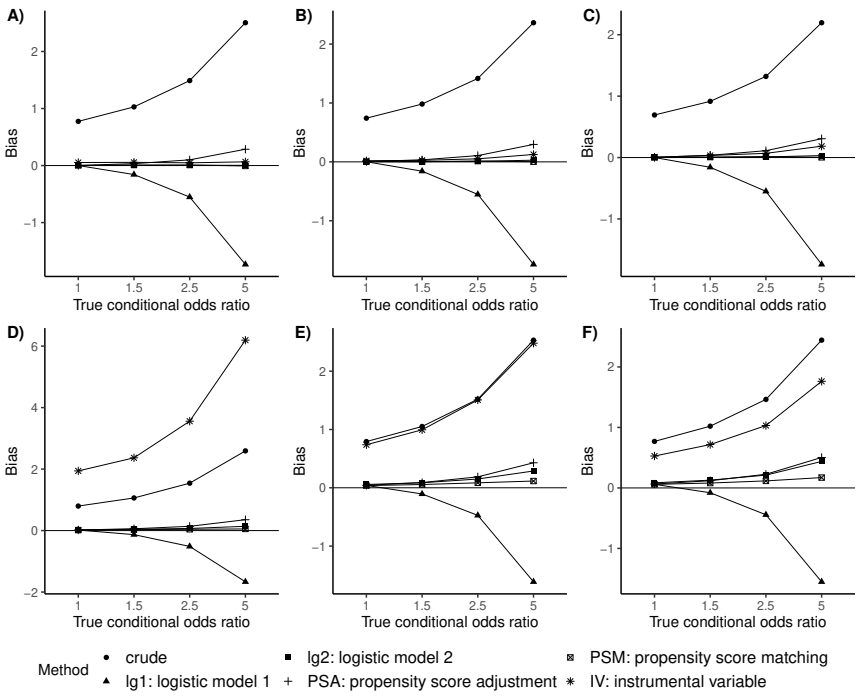

Figure 1 Bias under different conditional odds ratios in the absence of unmeasurable confounders. In parts A, B and C, the defined instrument has direct effect on the outcome; in parts D, E and F, the defined instrument has no direct effect on the outcome ; in parts A and D, the instrument-treatment odds ratios are 2.0; in parts B and E, the instrument-treatment odds ratios are 5.0; In parts A and D, the instrument-treatment odds ratios are 10.0. Baseline treatment prevalence was 50%, and baseline outcome probability was 50%.

be completely different when IV had only a weak effect on the outcome. As the true conditional OR increased, IV overestimated the true OR of CACE more.

MSE of each method when no unmeausred confounder exists

Figure 3. shows the mean squared error of different estimation methods when there is no unobservable confounder. The crude estimate of the treatment effect had the greatest MSE except in the situation when the weak or the moderate instrument had a direct effect on the outcome. Logitic model 1 resulted in minor MSE when the true conditional OR was equal to or below 2.5, but the MSE of it increased significantly when the true conditional OR was equal to 5. When IV analysis was used, the MSE decreased as the defined IV became stronger. When we employed the strong instrument without direct association to the outcome, the MSE of IV were comparable to that of logistic model 2, covariate adjustment using the propensity score and propensity score matching, although the MSE of IV was negligibly greater. When we employed the weak instrument with direct effct on the outcome, the MSE was significantly greater than other methods. As the defined instrument became stronger, the MSE of IV was improved but still larger than that of other methods.

MSE of each method when unmeasured confounder exists

The mean squared error of different estimation methods when unobsevable confounder exists are shown in Figure 4.. In this scenario, the situation was not as

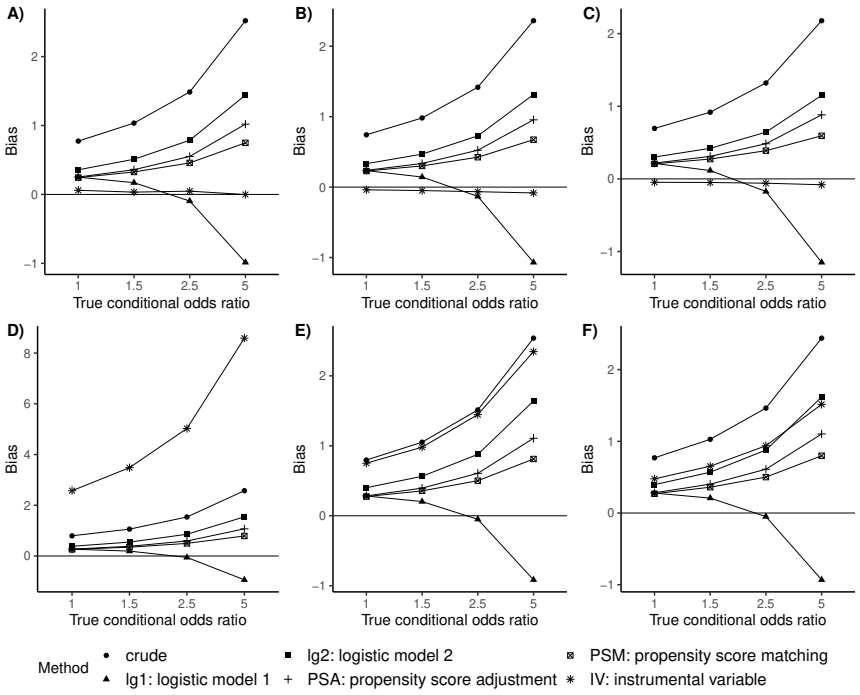

Figure 2 Bias under different conditional odds ratios in the presence of unmeasurable confounders. In parts A, B and C, the defined instrument has direct effect on the outcome; in parts D, E and F, the defined instrument has no direct effect on the outcome ; in parts A and D, the instrument-treatment odds ratios are 2.0; in parts B and E, the instrument-treatment odds ratios are 5.0; In parts A and D, the instrument-treatment odds ratios are 10.0. Baseline treatment prevalence was 50%, and baseline outcome probability was 50%.

same as the above. First of all, the MSE of logistic model 2, covariate adjustment using the propensity score and propensity score matching were larger than that in scenario 1. It is worth noting that the MSE of logistic model 1 is smaller than that of logistic model 2 which showed the opposite of the above results. When we employed the strong instrument which had no direct effect on the outcome, the MSE of it was the smallest. However, the MSE of IV became much larger once the direct association between the instrument and the outcome showed up, even a strong instrument cannot make its MSE less than that of other methods.

Discussion

We conducted an extensive series of Monte Carlo simulations to examine the performance of logistic regression, propensity score and instrumental variable to estimate their target odds ratios. The present analysis illustrates the challenges faced in determining which methods actually produce the most valid results in different settings. We summarize our findings as follows.

First, we demonstrated that when there is no unobservable confounder existing, propensity score matching had the best performance of controling for confounders, and led to the most accurate estimation. Logistic regression including all variables related to the outcome and instrumental variable had comparable performances provided that the assumptions of IV were satisfied. When the defined instrument affected the outcome directly, unsurprisingly, it caused great positive bias even when

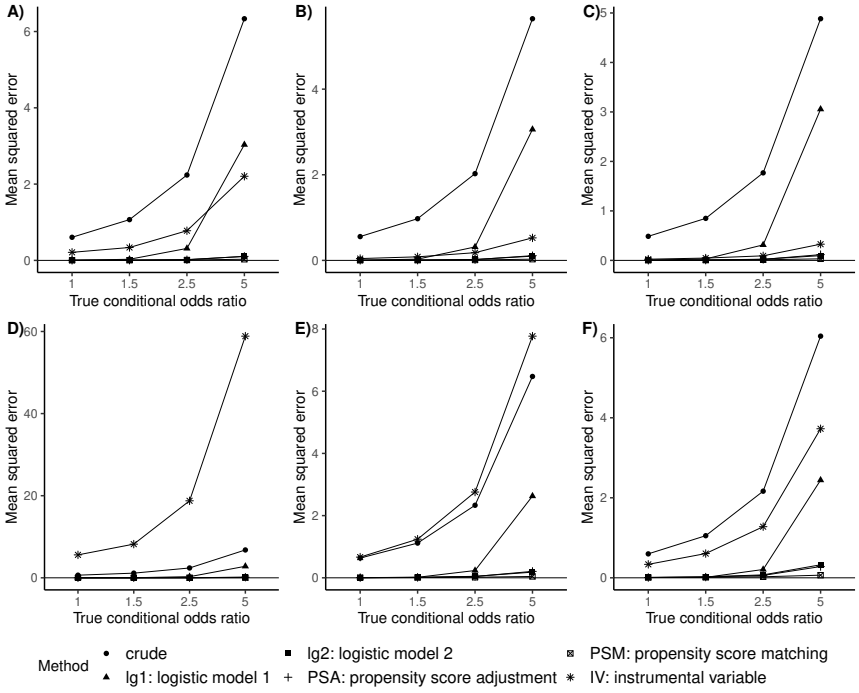

Figure 3 Mean suqared error under different conditional odds ratios in the absence of unmeasurable confounders. In parts A, B and C, the defined instrument has direct effect on the outcome; in parts D, E and F, the defined instrument has no direct effect on the outcome ; in parts A and D, the instrument-treatment odds ratios are 2.0; in parts B and E, the instrument-treatment odds ratios are 5.0; In parts A and D, the instrument-treatment odds ratios are 10.0. Baseline treatment prevalence was 50%, and baseline outcome probability was 50%.

the instrument had a strong association with the treatment, which implies that an strong IV cannot offset the negative effect brought by the weakly direct association between the instrument and the outcome.

In prior researches[10, 6], it was shown that matching on the propensity score can eliminate a greater degree of treatment selection bias than does covariate adjustment on the propensity score. Our study is consistent with this result. When canparing the propensity score methods and logistic regression, a lot of empirical researches [11, 12] has shown that propensity score methods gave similar results to traditional logistic regression, but Cepeda et al.[2] concluded that propensity scores are a better multivariable technique when there are equal or below 7 events per confounder. However, most of these studies compared their results without considering that propensity score methods and logistc regression to be estimating the different treatment effect. Logistic regression allows one to obtain conditional estimated OR of treatment effect, while propensity score methods allow one to estimate the OR of ATE/ATT. These two estimated OR coincide when the true conditional treatment effect is null [13]. We calculated the bias by using the inherently target OR of each method in our study which showed propensity matching had an ignorable better performance than the logistic model including all variables related to the outcome.

Second, we compared the bias caused by each method when unmeasured confounder exists. In this scenario, instruments without direct association to the outcome can lead an uncomparable performace. This situation was unfortunately re-

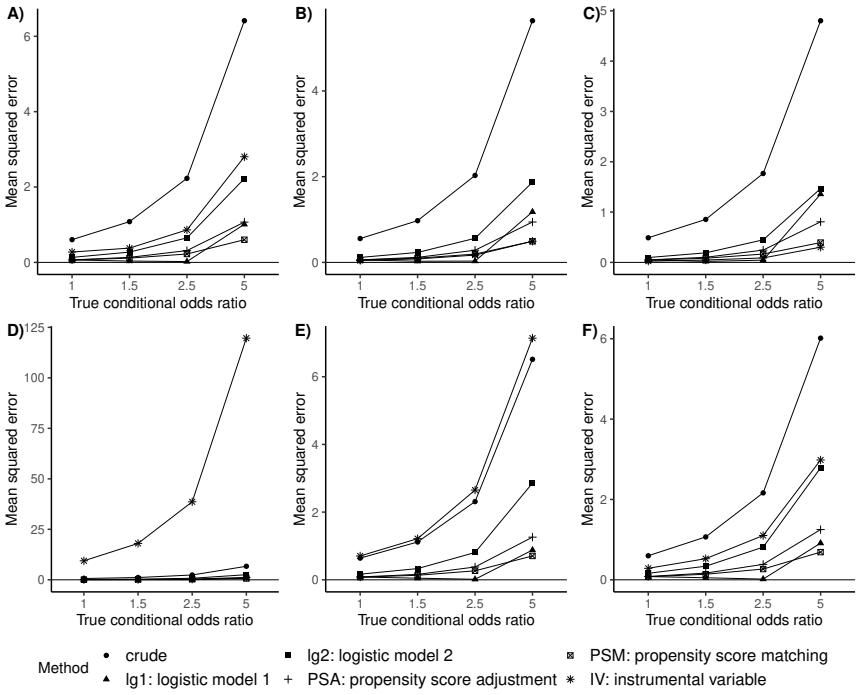

Figure 4 Mean suqared error under different conditional odds ratios in the presence of unmeasurable confounders. In parts A, B and C, the defined instrument has direct effect on the outcome; in parts D, E and F, the defined instrument has no direct effect on the outcome ; in parts A and D, the instrument-treatment odds ratios are 2.0; in parts B and E, the instrument-treatment odds ratios are 5.0; In parts A and D, the instrument-treatment odds ratios are 10.0. Baseline treatment prevalence was 50%, and baseline outcome probability was 50%.

versed by the direct association between the instrument and the outcome. This implies that the ability of IV for controlling unmeasured confounders can be disrupted by the violation of IV assumptions. Another finding that's interesting is that logistic model only including true confounders perform better than logistic model including all outcome-related covariates in this case.

Previous studies have demonstrated that traditional logistic regression models and all propensity scoring methods could only control for measurable confounders, the main limitation in such methods, namely their inability to account for unmeasured confounders. Drake et al. [14] suggested that PS may not be superior to conventional multivariable models in controlling bias from unobserved confounders. Nonetheless, instrumental variable analysis retains a key role in clinical research, given its superior performance in adjusting for unmeasured confounders [15]. However, there is no definitive better IV method for dichotomous data. Terza et al. [9] compared 2SRI with two stage predictor substitution. They found that 2SRI performed better, although in some cases neither of them yielded unbiased estimates [16]. Very few studies compared IV to other methods. We found the only empirical study comparing the results of IV and PS [17], which showed that the results of IV and PS were different and that values obtained from IV were higher than those from PS. Our study showed that when the IV had derictly effect on the outcome, the estimated value from PS was going to be less than that from IV, which implies

that the assumption that IV must not directly affect the outcome might be violated in these studies included in this empirical study.

Third, we compared the MSE of each method in two scenarios. IV led undesirable performances when the defined instrument was weak or when it had direct effect on the outcome, while both propensity score matching and adjustment performed well in all situations. These results are consistent with previous studies [18, 19] which implies that propensity score led the smallest MSE, and IV usually led greater MSE and thus need larger sample size [20, 21]. We also conducted simulations to see what happened when the sample size is small, we find that IV could lead significantly greater MSE, which made its estimation very unstable especially when the sample size was lower than 5000.

In practical studies, investigators are unable to directly prove the existence of unmeasured confounders and the association between IV and outcome, although there are some measures of making indirect explorations [22, 23, 24]. Based on our findings, we believe that the comparison of the results obtained from these methods can provide clues as to which of the results of these methods is more reliable. As shown in Table 2. and Table 3., the odds ratio obtained by IV was likely to be greater than the crude value only if the instrumental variable is weak and directly affect the outcome; Similarly, the odds ratio obtaind from IV was greater than that from PS only if IV is directly associated with the outcome. If IV has no direct effect on the outcome and unmeasured confounder exists, the value of PS is larger than that of IV. Therefore, in practice, when the distribution of the total population, the actual treatment population and the compliance population is similar, if the odds ratio of PS and IV are comparable, it may imply that there is no unmeasured confounding, and IV is not directly associated with the outcome, and then each estimation of the three methods may be reliable; If the value obtained from PS is much larger than IV, it may indicate that there is unmeasured confounding, and IV does not directly affect the outcome. In this case, the results of IV may be more reliable; If the value of IV is much larger than the crude value or than the value of PS, then we should be highly cautious about the value of IV. It is likely that the assumption of no direct association between the instrument and the outcome may be violated, and the bias of the IV results will be greater than that of the other two methods even when unmeasured confounders exist.

Finally, we have to warn that the comparison of results obtained by each method can only provide clues as to which method is likely to be more reliable, rather than making an accurate assessment. Researchers should always bear in mind that the three methods themselves intend to estimate different effects, and that reliable results depend on the good conduct of these methods. Therefore, the first step for the investigator is to find out which treatment effect the study would like to evaluate, determine whether the methods are appropriate for the data and whether the testable assumptions have been tested and satisfied.

Our study has some limitations. First of all, our study is based on dichotomous data and may not generalize to other types of data; Secondly, the given covariates were independent of each other, and there is no multicollinearity and no interaction effect, so we can't say in these cases whether we can draw conclusions that are consistent with our current study; Thirdly, the original data set by us was subject

to logistic distribution. If the real data conforms to other distributions such as Possion distribution, the results may become different.

Conclusion

In conclusion, investigators should keep in mind that there is no magic bullets against all potential causes of bias in analysis. With good implementation, IV can lead to unbiased results. However, weak instrument will bring great MSE, and a little direct association between the instrument and the outcome can cause greater bias than propensity score and logistic model, even when unmeasured confounder exists. The negative effect caused by violating the required assumptions of IV can overweigh the positive effect of IV's ability to control for unmeasured confounder. However, comparing the results of each method can provide us with some clues about the reliability of each.

Acknowledgements
Not applicable

Funding
This work was supported by The National Key Research and Development Program of China. The funding body had no role in the contents and the writing of the manuscript.

Abbreviations
ATE, the average treatment effect; ATT, the average treatment effect in treated ; CACE, the compiler average causal effect; IV, instrumental variable; MSE, mean square error; OR, odds ratio; PS, propensity score; 2SRI, two-stage residual inclusion.

Availability of data and materials
The R programm for simulation are available from the corresponding author.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Authors' contributions
YT designed the study and directed its implementation, including quality assurance and control. KC helped in model selection, especially in instrumental variable model selection, data generation of Monte Carlo simulation, and writing R code. XW helped visualize the results, including generating tables and figures and prepare the Discussion sections of the text. Author LJ helped conduct the literature review and prepare the Discussion sections of the text.

Author details
[1] Department of Evidence-Based Medicine and Clinical Epidemiology, Sichuan University West China Hospital, Chengdu, China. [2]Department of Engineering Management, Information, and Systems, Southern Methodist University, Texas, United States.

References

1. Yao, X.I., Wang, X., Speicher, P.J., Hwang, E.S., Cheng, P., Harpole, D.H., Berry, M.F., Schrag, D., Pang, H.H.: Reporting and Guidelines in Propensity Score Analysis: A Systematic Review of Cancer and Cancer Surgical Studies. J Natl Cancer Inst **109**(8) (2017). doi:10.1093/jnci/djw323. djw323
2. Cepeda, M.S., Boston, R., Farrar, J.T., Strom, B.L.: Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol **158**(3), 280–287 (2003)
3. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
4. VanderWeele, T.J., Ding, P.: Sensitivity analysis in observational research: introducing the e-value. Ann Intern Med **167**(4), 268–274 (2017)
5. Greenland, S.: An introduction to instrumental variables for epidemiologists. Int J Epidemiol **29**(4), 722–729 (2000)
6. Austin, P.C., Grootendorst, P., Anderson, G.M.: A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. Stat Med **26**(4), 734–753 (2007)
7. Weitzen, S., Lapane, K.L., Toledano, A.Y., Hume, A.L., Mor, V.: Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiol Drug Saf **13**(12), 841–853 (2004)

8. Elze, M.C., Gregson, J., Baber, U., Williamson, E., Sartori, S., Mehran, R., Nichols, M., Stone, G.W., Pocock, S.J.: Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. J Am Coll Cardiol **69**(3), 345–357 (2017)
9. Terza, J.V., Basu, A., Rathouz, P.J.: Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. J Health Econ **27**(3), 531–543 (2008)
10. Austin, P.C., Mamdani, M.M., Juurlink, D.N., Alter, D.A., Tu, J.V.: Missed opportunities in the secondary prevention of myocardial infarction: an assessment of the effects of statin underprescribing on mortality. Am Heart J **151**(5), 969–975 (2006)
11. Stürmer, T., Joshi, M., Glynn, R.J., Avorn, J., Rothman, K.J., Schneeweiss, S.: A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol **59**(5), 437–1 (2006)
12. Shah, B.R., Laupacis, A., Hux, J.E., Austin, P.C.: Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol **58**(6), 550–559 (2005)
13. Greenland, S.: Interpretation and choice of effect measures in epidemiologic analyses. Am J Epidemiol **125**(5), 761–768 (1987)
14. Drake, C.: Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics, 1231–1236 (1993)
15. Rassen, J.A., Schneeweiss, S., Glynn, R.J., Mittleman, M.A., Brookhart, M.A.: Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. Am J Epidemiol **169**(3), 273–284 (2009)
16. Basu, A., Coe, N.B., Chapman, C.G.: 2sls versus 2sri: A ppropriate methods for rare outcomes and/or rare exposures. Health Econ **27**(6), 937–955 (2018)
17. Laborde-Castérot, H., Agrinier, N., Thilly, N.: Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review. J Clin Epidemiol **68**(10), 1232–1240 (2015)
18. Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res **46**(3), 399–424 (2011)
19. Austin, P.C.: The performance of different propensity score methods for estimating marginal odds ratios. Stat Med **26**(16), 3078–3094 (2007)
20. Martens, E.P., Pestman, W.R., de Boer, A., Belitser, S.V., Klungel, O.H.: Instrumental variables: application and limitations. Epidemiology, 260–267 (2006)
21. Ionescu-Ittu, R., Delaney, J.A., Abrahamowicz, M.: Bias–variance trade-off in pharmacoepidemiological studies using physician-preference-based instrumental variables: a simulation study. Pharmacoepidemiol Drug Saf **18**(7), 562–571 (2009)
22. Baiocchi, M., Cheng, J., Small, D.S.: Instrumental variable methods for causal inference. Stat Med **33**(13), 2297–2340 (2014)
23. Ertefaie, A., Small, D.S., Flory, J.H., Hennessy, S.: A tutorial on the use of instrumental variables in pharmacoepidemiology. Pharmacoepidemiol Drug Saf **26**(4), 357–367 (2017)
24. Brookhart, M.A., Rassen, J.A., Schneeweiss, S.: Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf **19**(6), 537–554 (2010)

Figures

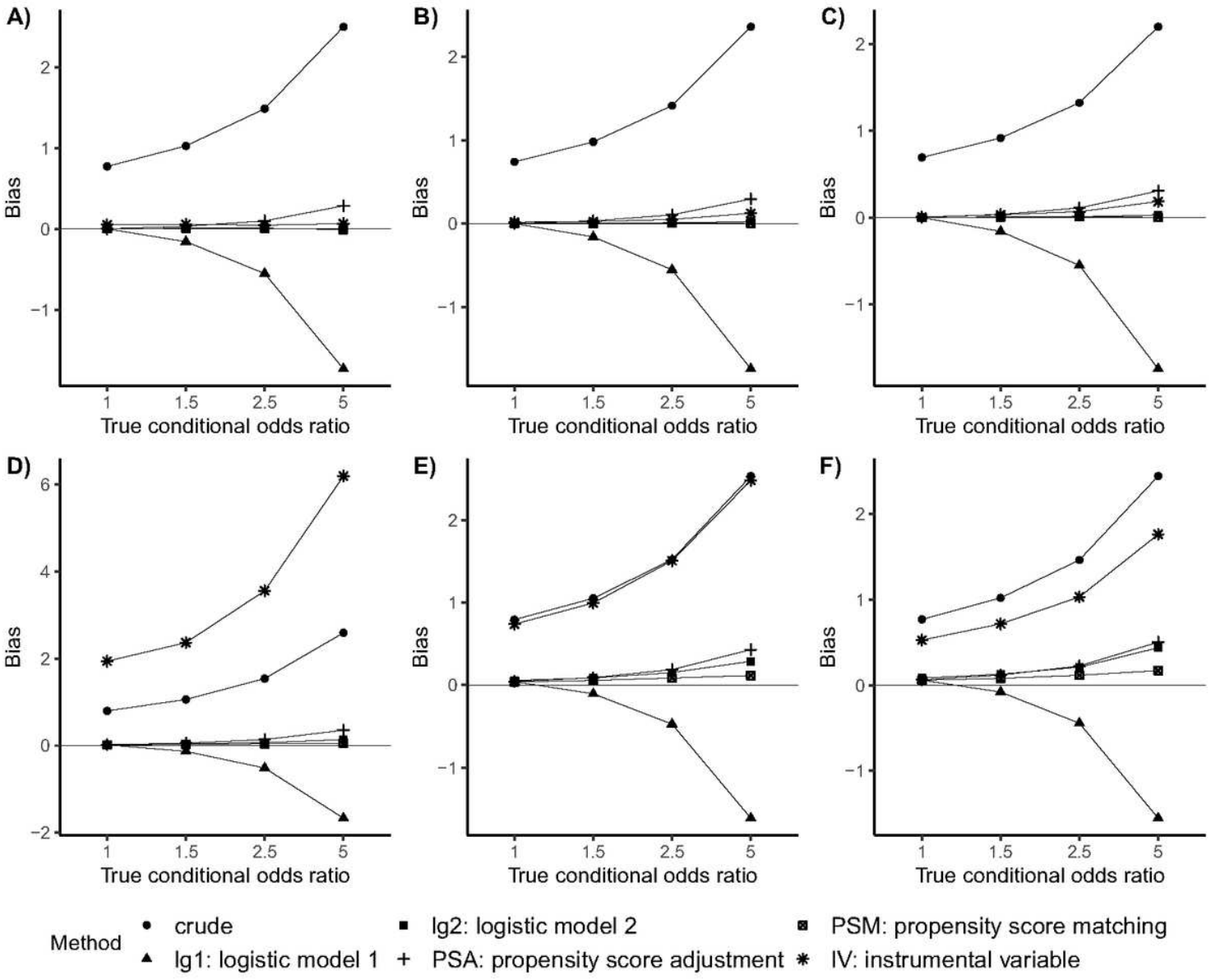

Figure 1

Bias under different conditional odds ratios in the absence of unmeasurable confounders. In parts A, B and C, the defined instrument has direct effect on the outcome; in parts D, E and F, the defined instrument has no direct effect on the outcome ; in parts A and D, the instrument-treatment odds ratios are 2.0; in parts B and E, the instrument-treatment odds ratios are 5.0; In parts A and D, the instrument-treatment odds ratios are 10.0. Baseline treatment prevalence was 50%, and baseline outcome probability was 50%.

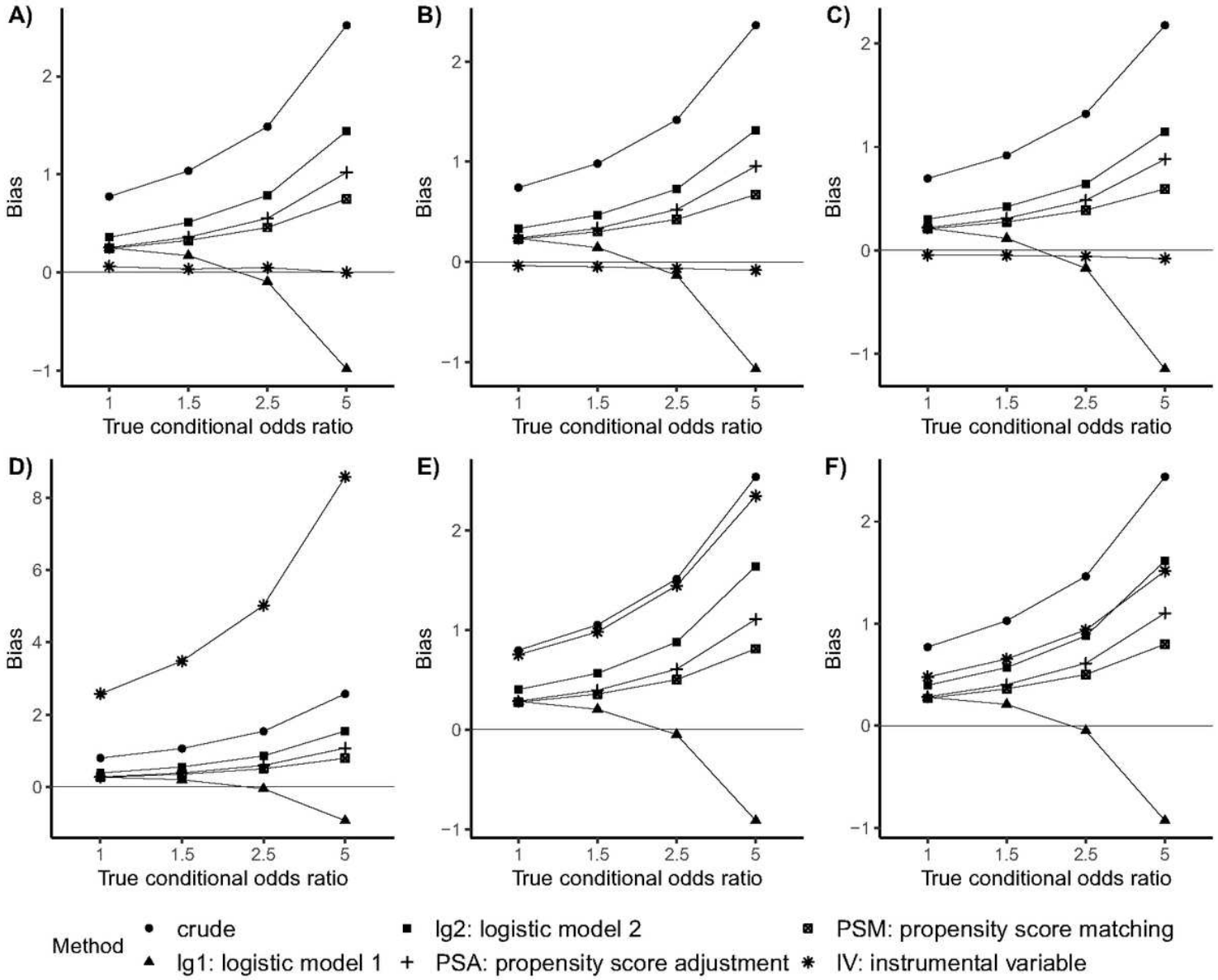

Figure 2

Bias under different conditional odds ratios in the presence of unmeasurable confounders. In parts A, B and C, the defined instrument has direct effect on the outcome; in parts D, E and F, the defined instrument has no direct effect on the outcome ; in parts A and D, the instrument-treatment odds ratios are 2.0; in parts B and E, the instrument-treatment odds ratios are 5.0; In parts A and D, the instrument-treatment odds ratios are 10.0. Baseline treatment prevalence was 50%, and baseline outcome probability was 50%.

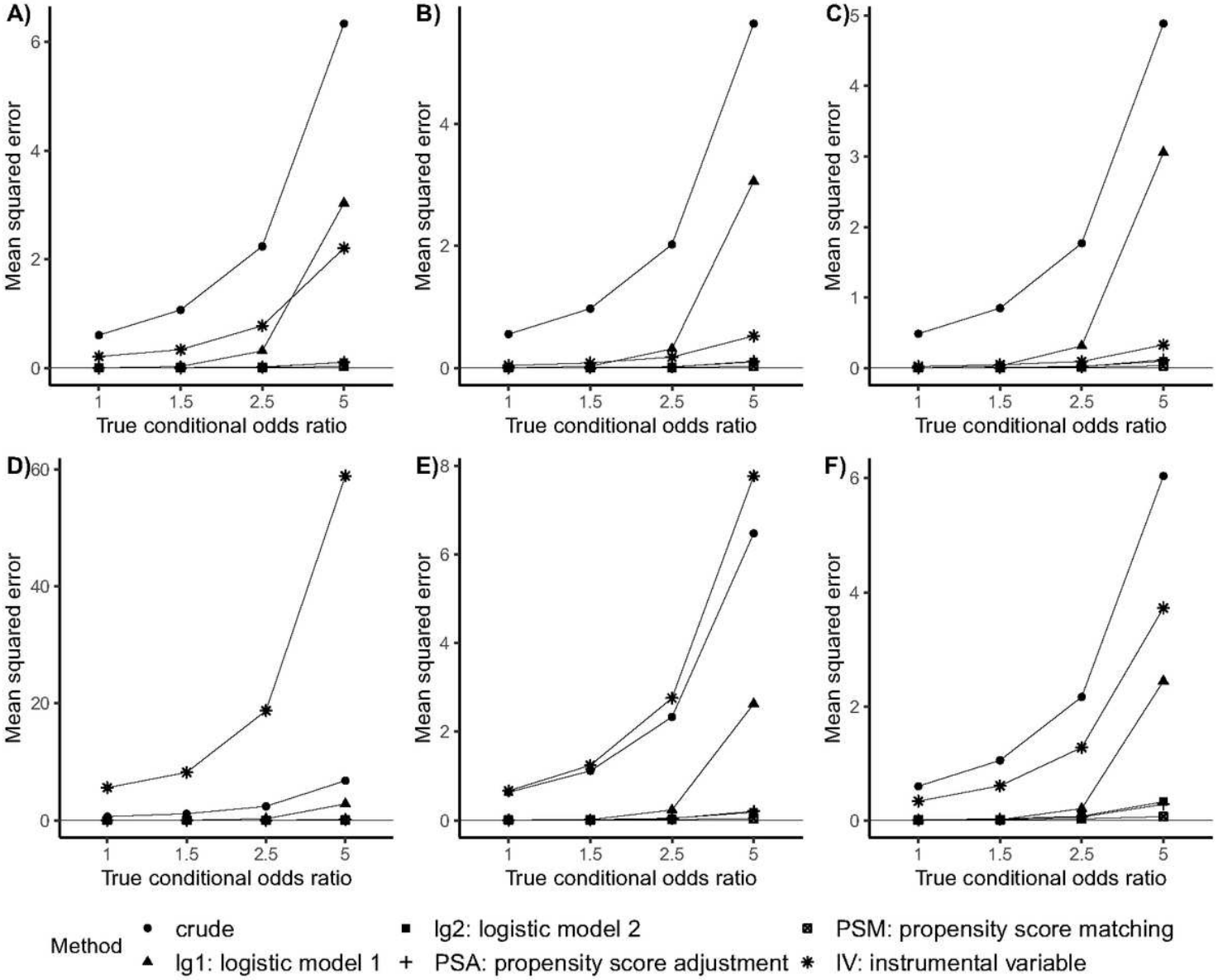

Figure 3

Mean suqared error under different conditional odds ratios in the absence of unmeasurable confounders. In parts A, B and C, the defined instrument has direct effect on the outcome; in parts D, E and F, the defined instrument has no direct effect on the outcome ; in parts A and D, the instrument-treatment odds ratios are 2.0; in parts B and E, the instrument-treatment odds ratios are 5.0; In parts A and D, the instrument-treatment odds ratios are 10.0. Baseline treatment prevalence was 50%, and baseline outcome probability was 50%.

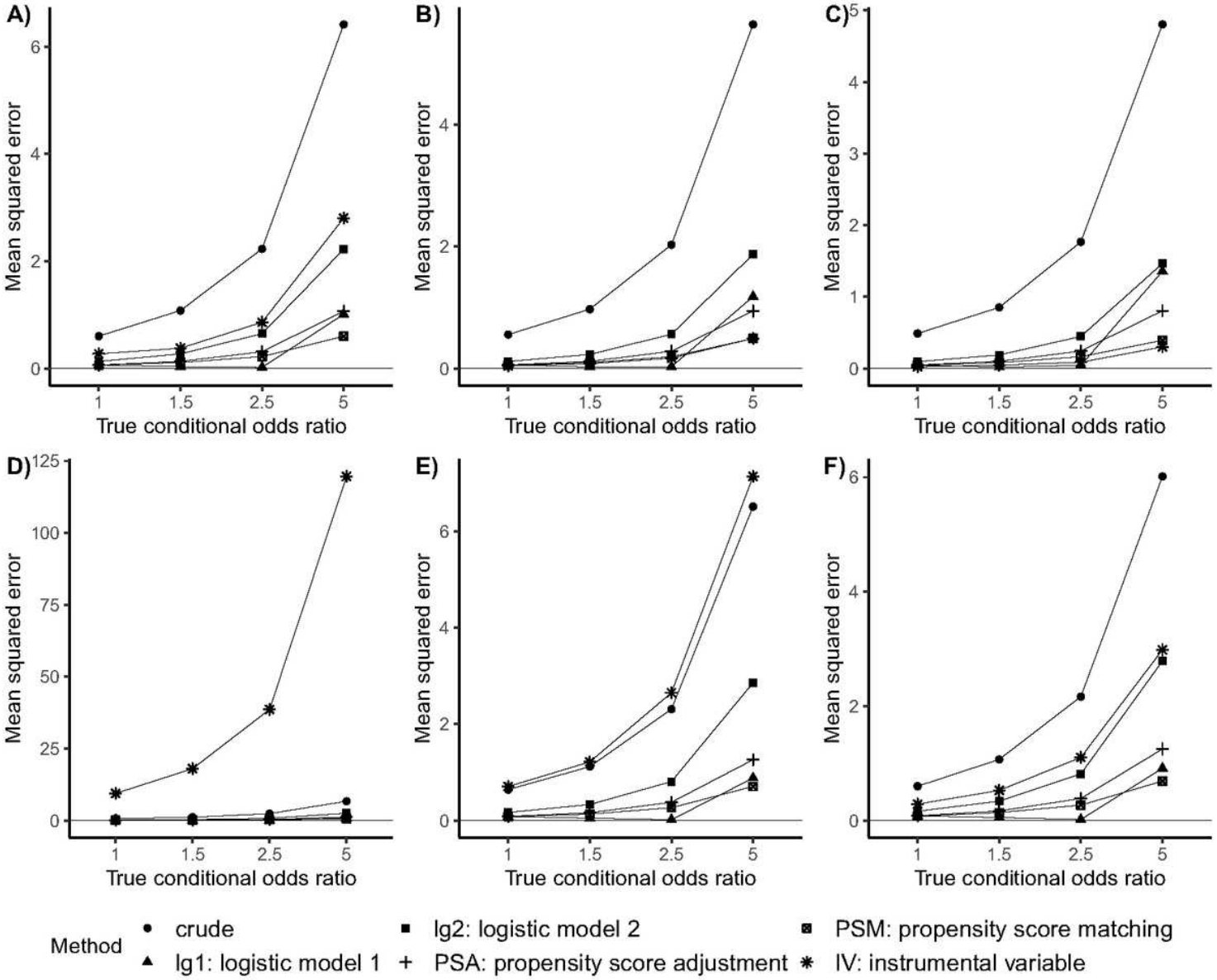

Figure 4

Mean suqared error under different conditional odds ratios in the presence of unmeasurable confounders. In parts A, B and C, the defined instrument has direct effect on the outcome; in parts D, E and F, the defined instrument has no direct effect on the outcome ; in parts A and D, the instrument-treatment odds ratios are 2.0; in parts B and E, the instrument-treatment odds ratios are 5.0; In parts A and D, the instrument-treatment odds ratios are 10.0. Baseline treatment prevalence was 50%, and baseline outcome probability was 50%.