
ICML Comments and Responses of “Identification and Estimation of the Bi-Directional MR with Some Invalid Instruments”

Anonymous Author(s)

Affiliation

Address

email

1 Main Concerns

2 1.1 More descriptions and examples for the bi-directional MR model

3 We have rephrased our Introduction Section in the main paper and offered Appendix A to elaborate
4 detailedly our bi-directional MR causal model, with motivating examples.

- 5 • **(Basic idea of Mendelian Randomization)**. The basic goal of epidemiologic studies is
6 to assess the effect of changes in exposure on outcomes. The causal effect of exposure
7 on outcome is often different from the observed correlation due to confounding factors.
8 Correlations between exposure and outcome cannot be used as reliable evidence for inferring
9 causality. In contrast, Mendelian Randomization (MR) studies use genetic variants to infer
10 the causal effect of exposure on outcome. The idea behind MR studies is to find a genetic
11 variant (or multiple genetic variants) that is associated with exposure but not with other
12 confounders and that does not directly affect the outcome. Such a genetic variant in principle
13 fulfills the basic assumptions of instrumental variables (IVs). It is then used to assess
14 the causal effect of the exposure on the outcome [Thomas and Conti, 2004]. Since the
15 assumptions of IVs cannot be fully tested and may be violated during IVs’ selection, a
16 number of methods have emerged to identify valid IVs.
- 17 • **(Two-sample MR & One-sample MR)**. MR studies can be conducted using a single sample,
18 i.e., **one-sample Mendelian Randomization**, in which individuals are tested for genetic
19 variation, exposure, and outcome in the same population. On the contrary, in **two-sample**
20 **Mendelian Randomization**, the association between genetic variation and exposure is
21 estimated in one dataset while that between genetic variation and outcome is estimated in
22 another dataset. Compared with the Two-sample MR, one-sample MR provides more direct
23 evidence by eliminating potential biases that can arise from individual-level data, which
24 turns out to be more meaningful yet challenging.
- 25 • **(Bi-directional Mendelian Randomization)**. Most existing MR methods assume a one-
26 directional causal relationship between exposure and outcome, whereas bi-directional re-
27 lationships are ubiquitous in real-life scenarios. For instance, there exist bi-directional
28 relationships between obesity and vitamin D status [Vimalaswaran et al., 2013], body mass
29 index (BMI) and type 2 diabetes (T2D), diastolic blood pressure and stroke [Xue and
30 Pan, 2022], insomnia and five major psychiatric disorders [Gao et al., 2019], smoking and
31 BMI [Carreras-Torres et al., 2018], etc. It is thus desirable to take further research on
32 bi-directional MR. **Bi-directional Mendelian Randomization** assesses not only the causal
33 effect of the exposure on the outcome but also the effect of the outcome on the exposure. To
34 achieve this, valid IVs for both the exposure and the outcome are needed. However, this is
35 tough because genetic variants associated with the exposure are often also associated with

the outcome, and vice versa. Our goal here is to identify and differentiate valid IVs for each direction within a one-sample MR framework and infer bi-directional causal effects.

1.2 More experimental results.

The reviewers were interested in the robust performances of our method in other different scenarios. Hence, we added more powerful experimental results in the main paper and Appendix, summarized below.

- **(Baselines' results for $Y \rightarrow X$ direction).** We adopted the helpful suggestions of Reviewer 2 and provided results for $Y \rightarrow X$ direction of those baselines that are only applicable to one-directional MR models. It is not surprising to see that without prior knowledge, these baselines can only perform well in one direction, while would fail in another direction. Results are illustrated in Table 1, Table 3, Table 4, Table 5, and Table 6.
- **(Small sample size).** We performed additional experiments with a sample size of 200. Detailed experimental results are shown in Appendix H.3. We found that though the performances of all methods are not satisfactory, ours still outperforms other baselines. It also reflects an open problem of our work with small sample sizes. We leave it as our future in-depth direction.
- **(Asymmetric IVs & All IVs being valid).** Reviewer 2 was interested in the method's performances in cases where: 1) the number of IVs in both directions is not equal under the bi-directional model; 2) and all IVs in the model are valid. We added these experimental scenarios in Appendix H.4 (Different Numbers of Valid IVs), and Appendix H.5 (All IVs Being Valid). The results demonstrated the applicability and superiority of our method in both cases.

1.3 Violation of Assumption A3 [Randomness].

Reviewer 2 and Reviewer 4 pointed out that our theories are all based on the fact that genetic variant G is uncorrelated with unmeasured confounders U (Assumption A3). It is noteworthy that in an MR model, the genetic variant is usually randomized, that is to say, unmeasured confounders U can not cause G . Thus, we focus on the case that G may cause unmeasured confounders U . According to their suggestions, we extend our theory to allow for the violation of Assumption A3 [Randomness], where G may cause unmeasured confounders U . Please see Proposition 2 and 3 for more details.

2 Detailed comments and reponses

2.1 Reviewer 1 (Rating: 6 & Confidence: 4)

Q1: The bidirectional structure between the exposure X and the outcome Y , which leads to the core contribution of this paper, should be motivated more, with different real-world examples. Currently, I find it hard to think what it means and am inclined to think of time-dependent relations between X and Y , where they influence each other in turn, but that is closer to "control problem" and might not be the case here. In any case, I think it is crucial to motivate why that bidirectional structure is relevant, where does it appear & is commonplace etc, for the overall value of this paper and its potential impact in practice.

A1: Thank you for your important questions. Bidirectional relationships are ubiquitous in real-life scenarios, as seen in various examples such as obesity and vitamin D status [Vimalaswaran et al., 2013], body mass index and type 2 diabetes (T2D), diastolic blood pressure and stroke[Xue and Pan, 2022], insomnia and five major psychiatric disorders[Gao et al., 2019], as well as smoking and BMI[Carreras-Torres et al., 2018], etc. Understanding and analyzing bidirectional relationships can provide deeper insights into complex systems, leading to more effective interventions and solutions. We will add more background information about these bidirectional relationships in Introduction.

Q2: After a quick skim over the experiments Table 1, I find the results for different sample sizes to be similar (Please correct me if I'm wrong). At that point I was wondering given that authors test 3 different sample sizes, it might make more sense to choose as sample sizes, say 200, 2k, 10k, rather than 2k, 5k, 10k, to give an idea for the performance & limitations for small sample sizes.

85 **A2:** Thank you very much for your suggestion, below are the experimental results with 200 samples.
86 We find that though the performances of all four methods are not satisfactory, ours still outperforms
87 other baselines. It also reflects an open problem of our work with small sample sizes. We leave it
88 as our future in-depth direction. We will add this in Discussion. *Detailed experimental results are
89 shown in the Appendix.

90 **2.2 Reviewer 2 (Rating: 7 & Confidence: 3)**

91 **Q1:** Could you provide a more grounded example (specific Gs, Xs, Ys, and Us) for the IV situation
92 discussed in this paper?

93 **A1:** Thanks for your suggestion. A simple example studied by Vimalleswaran et al. [2013] involves the
94 use of specific genetic variants, known as single nucleotide polymorphisms (SNPs), as instrumental
95 variables (IVs) to infer the causal effect of obesity on Vitamin D status. Specifically, for obesity (X),
96 the study employed a BMI allele score derived from 12 SNPs associated with BMI. For Vitamin D
97 status (Y), two allele scores were created, one related to genes involved in the synthesis of 25(OH)D,
98 i.e., 25-hydroxyvitamin D, and another related to genes involved in its metabolism. The potential
99 confounders U for Vitamin D could be lifestyle factors such as diet and outdoor activities, which can
100 influence both BMI and 25(OH)D levels. For instance, individuals who lead sedentary lifestyles and
101 have limited sun exposure may be at risk for both obesity and Vitamin D deficiency.

102 Bidirectional relationships are ubiquitous in real life, as seen in various examples such as obesity
103 and vitamin D status[Vimalleswaran et al., 2013], diastolic blood pressure and stroke[Xue and Pan,
104 2022], insomnia and five major psychiatric disorders[Gao et al., 2019], as well as smoking and
105 BMI[Carreras-Torres et al., 2018], etc. Understanding and analyzing bidirectional relationships can
106 provide deeper insights into complex systems, leading to more effective interventions and solutions.

107 We will include more specific examples in the article to illustrate the motivation for such a study of
108 bidirectional scenarios.

109 **Q2:**In the experimental results, is there are reason the number of IVs in both the $X \rightarrow Y$ and $Y \rightarrow X$
110 directions are always set to be the same? Is performance affected at all by having asymmetry in the
111 number of valid IVs?

112 **A2:** Thanks for raising this point. In fact, whether the number of valid IVs in these two directions
113 is the same or not does not affect the performance of our algorithm theoretically. The following
114 experimental results also empirically verify it. Overall, In cases where the number of IVs in both the
115 $X \rightarrow Y$ and $Y \rightarrow X$ directions are different, our PReBiM algorithm outperforms other baselines in terms
116 of CSR and MSE.

117 Detailed experimental results are shown in the Appendix.

118 **Q3:** Minor note, but Section 6 (Discussion) feels out of place. When I see "Discussion" after an
119 experimental results, I assume it's a discussion of the results. I feel like the contents of Section 6
120 should be moved to before the experimental results.

121 **A3:** Thank you for your helpful suggestion. We have moved the content of Section 6 before the
122 experimental results.

123 **Q4:** In Section 5.2, you say "It has been noted that under the IV-TETRAD method for scenario
124 $S(2,0,6)$, the metrics CSR and MSE outperform our method." "It has been noted" is a strange way to
125 introduce this point (do you mean that you can see this in Figure 4? You don't actually say the total
126 number of genetic variants in these experiments, so it's unclear if Figure 4 corresponds to $S(2,0,6)$ or
127 if you're talking about some other experiments you did that just aren't reported here. If you're talking
128 about Figure 4 here, I'm not sure why you're specifically calling out IV-TETRAD outperforming
129 PReBiM at 2 valid IVs but not sisVIVE outperforming PReBiM at 4 valid IVs.

130 **A4:** Here, Figure 4 corresponds to scenario $S(2,0,6)$. Sorry for the confusion.

131 Regarding "sisVIVE outperforming us", one possible reason is that their method is not very demanding
132 in terms of sample sizes. However, we can observe that when the sample size is sufficient (10k), our
133 method still outperforms all other baselines at 4 valid IVs.

134 **Q5:** The authors mention a few approaches that handle MR studies with invalid IVs (though that
 135 have stronger assumptions than PReBiM). I wish at least one of these had been included in the
 136 experimental results...

137 **A5:** Thank you for your valuable suggestions. We have compared our proposed methods with
 138 MR-Egger method [Bowden et al., 2015], which addresses the problem of invalid IVs. Compared
 139 to our method, MR-Egger has no need to select invalid IVs in advance, but it requires a stronger
 140 assumption (InSIDE assumption). Thus, we here only report the metric MSE. The results are shown
 141 in the following table. We can observe from the table that the mean squared error (MSE) of our
 142 method, calculated from the estimates after selecting the correct instrumental variables (IVs), is
 143 significantly smaller than that of the MR-Egger method in all three cases, across all sample sizes.
 144 Detailed experimental results are shown in the Appendix.

145 **Q6:** In Section 1, you say that "our work does not restrict those assumptions" (referring to Assumption
 146 A3, that G are uncorrelated with U). Is this discussed anywhere in the paper? None of your examples
 147 feature correlation between G and U , and I can't find anywhere that it's mentioned after that line in
 148 the intro.

149 **A6:** Thanks so much for the valuable comment. Regarding the assumption A3, when genetic variants
 150 G are correlated with unmeasured confounders U , our approach still works out. Please see the
 151 following example and experimental illustrations:
 152 A simple example:

153 For an invalid IV set (G_1, G_2) in a bi-directional scenario, G_1 is a valid IV for direction $X \rightarrow Y$
 154 and G_2 is an invalid IV that affects both X and U . Following the procedure of TSLS(), we obtain
 155 $\text{corr}(Y - X\omega_{G_2}, G_1) \neq 0$, that is,

$$\begin{aligned} \omega_{G_2} &= \frac{\text{cov}(G_2, Y)}{\text{cov}(G_2, X)} = \beta_{X \rightarrow Y} + \beta_{\text{bias}}, Y - X\omega_{G_2} \\ &= \varepsilon_Y - \beta_{\text{bias}} \Delta(G_1 \gamma_{X,1} + G_2 \gamma_{X,2} + \varepsilon_X + \beta_{Y \rightarrow X} \varepsilon_X), \text{corr}(Y - X\omega_{G_2}, G_1) \\ &= -\beta_{\text{bias}} \Delta \gamma_{X,1} \neq 0. \end{aligned}$$

159 where $\beta_{\text{bias}} = \frac{\text{cov}(G_2, \varepsilon_Y)}{\text{cov}(G_2, X)} \neq 0$, $\Delta = \frac{1}{1 - \beta_{X \rightarrow Y} \beta_{Y \rightarrow X}} \neq 0$.

160 Our experimental results also show that our method effectively filters invalid IVs. In particular, with
 161 different sample sizes and different numbers of IVs, when assumption A3 is violated, our method
 162 PReBiM still surpasses other baselines by a large margin.

163 Detailed experimental results are shown in the Appendix.

164 **Q7:** My understanding is that X and Y are essentially interchangeable, since the direction of the edge
 165 is unknown or could be directional. So why, in Table 1 don't algorithms besides PReBiM (and naive
 166 for MSE) have results for $Y \rightarrow X$? Couldn't you just rerun the algorithms in the other direction? I know
 167 they aren't designed for that, but it seems like that would be the naive approach that a practitioner
 168 would try in the absence of a method designed for bidirectional MR models (like PReBiM), so it
 169 seems worth comparing against, unless I'm missing something.

170 **A7:** You are right! We can re-run these algorithms to obtain results in the other direction. However,
 171 as you mentioned, these algorithms are not specifically designed for bidirectional MR models. As
 172 such, they may not handle bi-directional structures well, resulting in results that may not be reliable
 173 or valid. Following your suggestion, we re-run these algorithms, and the results are given in the
 174 following table. As we expected, the performance of the other comparison methods is not adorable,
 175 which verifies the above point. We will include these results in the revision.

176 **Q8:** How does PReBiM perform in situations without any invalid IVs? I realize Table 1 is a bit big
 177 as-is, but I'd be interested in viewing a setting such as $S(2,2,4)$, to help disentangle the effects of
 178 accounting for bidirectional effects and the effects of avoiding using invalid instruments.

179 **A8:** Thanks for your intriguing comments. The results in the table below show the settings you
 180 mentioned for $S(2,2,4)$. To ensure the thoroughness of the experiment, we've added another setting,
 181 $S(3,3,6)$. As expected, our method gives the best results across all sample sizes and directions. As the
 182 sample size increases, the metric CSR approaches 1, indicating that our method can identify valid
 183 IVs and also validating our theoretical guarantees. *Detailed experimental results are shown in the
 184 Appendix.

185 **Q9:** Is there a reason why you don't report MSE for Figure 4? Do those results tell the same story
186 as CSR? Looking for MSE results for Figure 4, I looked in the Supplementary Material and found
187 Figure 7. However, from the description in the text and the caption, I can't find what the difference is
188 supposed to be between Figures 4 and 7. Also, the description of Figure 7 says that it shows CSR and
189 MSE, but it clearly only shows CSR. What is the difference between these two sets of experiments?

190 **A9:** You're right, these two measures speak to the same issue, as shown in table 1. This is why we
191 omitted these results. The difference between Figure 4 and Figure 7 is whether the instrumental
192 variables are correlated or not. Specifically, Figure 7 shows the results when the genetic variants
193 (instrumental variables) are correlated, a topic explored in Section 6. As we can see, our method is
194 capable of identifying valid genetic variants even in scenarios where the variants are correlated.

195 **Q10:** Is there any reason that no other bidirectional MR methods were included in the experimental
196 results? Were there implementation issues?

197 **A10:** To the best of our knowledge, no existing method can identify valid instrumental variables
198 under the one-sample MR framework. Hence, our focus is specifically on developing methods for
199 identifying instrumental variables in bidirectional MR to address this significant gap in the research
200 landscape. Given the lack of in-depth exploration of this issue, we believe that this direction is
201 valuable and desirable for the current MR research.

202 **2.3 Reviewer 3 (Rating: 6 & Confidence: 3)**

203 **Q1:** I am confused that this method is only designed for the bidirectional/cyclic case, or can be
204 applied for the one-directional case. All theories are based on the bidirectional data-generating
205 processes in Equation (1), but the authors claimed that the method can work for the one-directional
206 case (e.g., Section 5.2). A detailed clarification can help readers understand the problem. More
207 motivations or examples of the cyclic relation between treatments and outcomes in the IV settings are
208 needed since this setting is quite unfamiliar to the community. Without such motivations, it's hard to
209 assess the significance of this paper.

210 **A1:** Thank you for your suggestions. We would like to clarify that our method is capable of addressing
211 the standard one-directional case by simply setting $\beta_{X \rightarrow Y} \neq 0$ and $\beta_{Y \rightarrow X} = 0$. The experimental
212 results in Section 5.2 also verify this point. Besides, we would like to mention that Bidirectional
213 relationships are ubiquitous in real-life scenarios, as seen in various examples such as obesity and
214 vitamin D status[Vimalaewaran et al., 2013], body mass index and type 2 diabetes (T2D), diastolic
215 blood pressure and strokeXue and Pan [2022], insomnia and five major psychiatric disordersGao
216 et al. [2019], as well as smoking and BMI[Carreras-Torres et al., 2018], etc. Understanding and
217 analyzing bidirectional relationships can provide deeper insights into complex systems, leading to
218 more effective interventions and solutions.

219 We will add more specific examples to the article to illustrate the motivation for such a study of
220 bidirectional scenarios.

221 **Q2:** More statistical analysis (such as the convergence rate, unbiasedness, consistency, etc.) is needed.
222 Currently, all the theories are stated with the condition where $n \rightarrow \infty$.

223 **A2:** Thanks for the insightful comments. In this work, we focus on the identifiability of valid
224 instrumental variables in the bi-directional MR. Analyzing the convergence rate and consistency
225 for our method presents a significant challenge, especially within bi-directional causal models. The
226 methodology of [Windmeijer et al., 2021, Zhang et al., 2022] may offer promising avenues to tackle
227 this issue. We will address this in future research.

228 **Q3:** "Once given a valid IV G, one may estimate the causal effect of risk factor X on the outcome Y
229 of interest consistently by using two-stage least squares (TSLS) (Burgess et al., 2017)." I think this
230 statement misleads, since still we need additional assumptions to identify the causal effect.

231 **A3:** Thank you for your thoughtful suggestion. Here we have defaulted to estimating causal effects
232 under a linear model, which is indeed misleading. We have corrected this statement in the revision.

233 **Q4:** In Figure 1, I think the arrow A3 should be U to G, since the arrow G to U are already
234 encapsulated in the arrow G to X and G to Y.

235 **A4:** Thank you for your comment. The directionality in Figure 1 follows the approaches outlined in
236 some relevant papers, such as [Bowden et al., 2015] and [Xue et al., 2021], which employ a similar
237 graphical structure. In our depiction, we aimed to maintain consistency with these prior studies and
238 adhere to their directional conventions.

239 **A5:** What's the condition that the solution $\beta_{Y \rightarrow X}$ and $\beta_{X \rightarrow Y}$ exists? It's a cyclic graph, so it's possible
240 that such solution may not exist.

241 **A5:** This is a great point, and you are right! Generally speaking, without additional conditions, one
242 can not obtain the solution for $\beta_{Y \rightarrow X}$ and $\beta_{X \rightarrow Y}$. However, if one is provided with at least one valid
243 IV relative to $X \rightarrow Y$ and at least one valid IV relative to $Y \rightarrow X$, the solutions for $\beta_{Y \rightarrow X}$ and
244 $\beta_{X \rightarrow Y}$ not only exist but can also be estimated using the Two-Stage Least Squares (TSLS) estimator.
245 This is why we need to identify valid IVs from candidate genetic variants in our paper. For more
246 detailed information on the identification conditions of the linear simultaneous equation model, one
247 may refer to pages 402-407 in [Hausman, 1983].

248 **Q6:** Is the method reducible to the case where there are no cyclic relation between treatments and
249 outcomes?

250 **A6:** Yes. If $\beta_{Y \rightarrow X} = 0$ (or $\beta_{X \rightarrow Y} = 0$), this model will be the standard one-directional MR model,
251 which does not include the cyclic relation between treatments and outcomes.

252 **Q7:** What is the meaning of "inferring causal direction"? Based on the data generating process in
253 Equation 1, two directions $X \rightarrow Y$ and $Y \rightarrow X$ both look valid.

254 **A7:** Thanks for your comments. Here, the term 'inferring causal direction' refers to the process of
255 identifying which instrumental variables are used to infer the direction $X \rightarrow Y$ and vice versa. In
256 reality, the nature of the relationship between exposure and outcome—whether it is unidirectional or
257 bidirectional—is not predetermined. However, our approach enables us to distinguish between these
258 possibilities, yielding results that are both more accurate and reliable.

259 **2.4 Reviewer 4 (Rating: 5 & Confidence: 3)**

260 **Q1:** This paper only focuses on a linear setting; however, the relationships between phenotypes often
261 involve nonlinear complex models.

262 **A1:** Thank you for your insightful comments.

263 (i) Our contributions mainly lie in that, to the best of our knowledge, no existing method can identify
264 valid instrumental variables under the one-sample MR framework. Hence, our focus is specifically
265 on developing methods for identifying instrumental variables in bidirectional MR to address this
266 significant gap in the research landscape. It is a desirable but challenging research topic.

267 (ii) Linearity model settings enjoy some remarkable characteristics, which could be employed to
268 entail the theoretical identifiability of the bi-directional MR model. In particular,

269 First, linear models provide a clear explanation of causal relationships, allowing us to infer causal
270 direction more easily. Second, linear models have a simpler theoretical foundation that is natural to
271 understand and implement. Finally, linear models have also been widely explored and used in many
272 practical situations, often providing meaningful results [Kang et al., 2016, Windmeijer et al., 2021,
273 Silva and Shimizu, 2017, Li and Ye, 2022]. Hence, we focus primarily on linear models, other than
274 nonlinear ones. And we will leave the nonlinearity as future directions, further delving into their
275 application in instrumental variable identification and causal inference.

276 **Q2:** As shown in Figure 1, invalid instrumental variables (IVs) might be correlated with unmeasured
277 confounders or have a direct pathway to the outcome. However, the premise on which all the theories
278 in this paper are based is that the genetic variants G are uncorrelated with unmeasured confounders
279 U . When this assumption is violated, the theories presented in this paper cannot be guaranteed to
280 hold. The authors have not fully addressed the issue of invalid instruments, nor have they discussed
281 violations of assumption A3, which represents the biggest weakness of the paper.

282 **A2:** Thanks so much for the valuable comment. Regarding the assumption A3, when genetic variants
283 G are correlated with unmeasured confounders U , our approach still works out. Please see the
284 following example and experimental illustrations:

285 A simple example:
 286 For an invalid IV set (G_1, G_2) in a bi-directional scenario, G_1 is a valid IV for direction $X \rightarrow Y$
 287 and G_2 is an invalid IV that affects both X and U . Following the procedure of TSLS(), we obtain
 288 $\text{corr}(Y - X\omega_{G_2}, G_1) \neq 0$, that is,

$$\begin{aligned} \omega_{G_2} &= \frac{\text{cov}(G_2, Y)}{\text{cov}(G_2, X)} = \beta_{X \rightarrow Y} + \beta_{bias, Y - X\omega_{G_2}} \\ &= \varepsilon_Y - \beta_{bias} \Delta (G_1 \gamma_{X,1} + G_2 \gamma_{X,2} + \varepsilon_X + \beta_{Y \rightarrow X} \varepsilon_X), \text{corr}(Y - X\omega_{G_2}, G_1) \\ &= -\beta_{bias} \Delta \gamma_{X,1} \neq 0. \end{aligned}$$

292 where $\beta_{bias} = \frac{\text{cov}(G_2, \varepsilon_Y)}{\text{cov}(G_2, X)} \neq 0$, $\Delta = \frac{1}{1 - \beta_{X \rightarrow Y} \beta_{Y \rightarrow X}} \neq 0$.

294 Experimental illustrations. As shown in the table below, our method effectively filters out invalid
 295 IVs. In particular, with different sample sizes and different numbers of IVs, when assumption A3 is
 296 violated, our method PReBiM still surpasses other baselines by a large margin.
 297 Detailed experimental results are shown in the Appendix

298 **Q3:** The abstract does not clearly present the studied problem and its challenges. In the Abstract and
 299 Introduction sections, the author needs to introduce what bi-directional Mendelian randomization
 300 is, along with some practical examples to illustrate it. When I first read through it, I found myself
 301 quite confused about what bi-directional MR means. Does it imply the presence of cycles within
 302 causal diagrams? What specific challenges does this introduce? How does it differ from traditional
 303 treatment effect estimation?

304 **A3:** Thank you for your helpful comments and suggestions to improve the comprehensibility of the
 305 article.

306 Bi-directional MR model & practical example In many real-world situations, the causal relationship
 307 between exposure and outcome may not be clear. In such cases where causal relationship between
 308 two related pheontypes is unknown, Bi-directional Mendelian randomization model studies how to
 309 estimate not only the causal effect of exposure on outcome but also the causal effect of outcome on
 310 exposure.

311 There are many practical examples in real-world scenarios, including obesity and vitamin D sta-
 312 tus[Vimaleswaran et al., 2013], body mass index and type 2 diabetes (T2D), diastolic blood pressure
 313 and stroke[Xue and Pan, 2022], insomnia and five major psychiatric disorders[Gao et al., 2019], as
 314 well as smoking and BMI[Carreras-Torres et al., 2018].

315 Specifically, a simple example studied by Vimaleswaran et al (2013)[1] involves the use of specific
 316 genetic variants, known as single nucleotide polymorphisms (SNPs), as instrumental variables (IVs)
 317 to infer the causal effect of obesity on Vitamin D status. Specifically, for obesity (X), the study
 318 employed a BMI (Body Mass Index) allele score derived from 12 SNPs associated with BMI. For
 319 Vitamin D status (Y), two allele scores were created, one related to genes involved in the synthesis of
 320 25(OH)D, i.e., 25-hydroxyvitamin D, and another related to genes involved in its metabolism. The
 321 potential confounders U for Vitamin D could be lifestyle factors such as diet and outdoor activities,
 322 which can influence both BMI and 25(OH)D levels. For instance, individuals who lead sedentary
 323 lifestyles and have limited sun exposure may be at risk for both obesity and Vitamin D deficiency.

324 Challenges:

325 Within a causal inference framework, MR can be implemented as a form of instrumental variables
 326 analysis where genetic variants serve as IVs. There exist various methods that through identifying
 327 instrumental variables achieve estimating causal effects, e.g., the baselines sisVIVE, IV-TETRAD,
 328 TSHT, in our paper. However, current methods deduce some limitations: i) circular structure in
 329 the graph is not allowed, and they only applies to unidirectional identification. When it comes to
 330 bi-directional identification, correlations between IVs and latent confounders become complicated,
 331 rendering a major challenge in theory. ii) they focus on the two-sample MR model that assumes the
 332 homogeneity across samples, while we consider the one-sample MR model. Ignoring the one-sample
 333 heterogeneity between sample could easily induce estimation bias for identifying IVs.

334 To the best of our knowledge, no existing method can identify valid instrumental variables under the
 335 one-sample bi-directional MR framework. Hence, our focus is specifically on developing methods for

336 identifying instrumental variables in bidirectional MR to address this significant gap in the research
337 landscape. It is a desirable but challenging research topic.

338 **Q4:** Is the equation (1) studied in this paper a fixed-point model? Is it a stable system?

339 **A4:** In the context of Equation (1), it can conceptually be regarded as a fixed-point model to the extent
340 that it aims to identify a state of equilibrium in causal estimations between two variables. Given its
341 linear framework, the model theoretically reaches a point where further iterations do not alter the
342 causal estimates, aligning with the idea of a fixed point. The stability of Equation (1) depends on
343 whether its causal estimations are consistent and resistant to changes in methodology or data. If these
344 estimations remain unchanged under various analytical conditions, then the model can be considered
345 stable.

346 We will further discuss the nature of equation (1) and the stability of the system in the manuscript for
347 better understanding.

348 **Q5:** As shown in Section 2.2, Bi-directional Mendelian Randomization does not seem to introduce
349 additional bias, allowing one to use standard Two-Stage Least Squares (TSLS) to analyze the one-
350 directional causal effect $X \rightarrow Y$ or $Y \rightarrow X$.

351 **A5:** Thanks for your comments. We would like to clarify the following two facts:

352 (i) Given a valid instrumental variable, one can obtain consistent estimates using the TSLS estimator
353 [Wooldridge, 2010]. It's worth noting that, for the two different models (unidirectional and bidirec-
354 tional), there is no difference in estimation, and consistent estimates can be obtained for both of them
355 [1].

356 (ii) Given an invalid instrumental variable, the TSLS estimator results for both models will be biased.
357 Please see [Bowden et al., 2015] and Remark 1 in our paper for detailed discussions.

358 Overall, it is only with the provision of valid instrumental variables (IVs) that we can obtain unbiased
359 causal effects. Therefore, identifying valid IVs from observational data becomes a prerequisite, which
360 is the main issue our paper addresses: how to identify valid IVs from the data in order to achieve
361 unbiased causal effects.

362 [1] Wooldridge, Jeffrey. 2009a. Introductory econometrics: A Modern Approach. Chapter 15:
363 Instrumental variables estimation and two stage least squares. South-Western, Nashville, TN. (Cited
364 on pages 62, 70, and 176)

365 References

366 Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid
367 instruments: effect estimation and bias detection through egger regression. *International journal*
368 *of epidemiology*, 44(2):512–525, 2015.

369 Robert Carreras-Torres, Mattias Johansson, Philip C Haycock, Caroline L Relton, George Davey
370 Smith, Paul Brennan, and Richard M Martin. Role of obesity in smoking behaviour: Mendelian
371 randomisation study in uk biobank. *bmj*, 361, 2018.

372 Xue Gao, Ling-Xian Meng, Kai-Li Ma, Jie Liang, Hui Wang, Qian Gao, and Tong Wang. The
373 bidirectional causal relationships of insomnia with five major psychiatric disorders: a mendelian
374 randomization study. *European Psychiatry*, 60:79–85, 2019.

375 Jerry A Hausman. Specification and estimation of simultaneous equation models. *Handbook of*
376 *econometrics*, 1:391–448, 1983.

377 Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation
378 with some invalid instruments and its application to mendelian randomization. *Journal of the*
379 *American statistical Association*, 111(513):132–144, 2016.

380 Sai Li and Ting Ye. A focusing framework for testing bi-directional causal effects with gwas summary
381 data. *arXiv preprint arXiv:2203.06887*, 2022.

382 Ricardo Silva and Shohei Shimizu. Learning instrumental variables with structural and non-
383 gaussianity assumptions. *Journal of Machine Learning Research*, 18(120):1–49, 2017.

384 Duncan C Thomas and David V Conti. Commentary: the concept of ‘mendelian randomization’.
385 *International journal of epidemiology*, 33(1):21–25, 2004.

386 Karani S Vimalaswaran, Diane J Berry, Chen Lu, Emmi Tikkanen, Stefan Pilz, Linda T Hiraki,
387 Jason D Cooper, Zari Dastani, Rui Li, Denise K Houston, et al. Causal relationship between
388 obesity and vitamin d status: bi-directional mendelian randomization analysis of multiple cohorts.
389 *PLoS medicine*, 10(2):e1001383, 2013.

390 Frank Windmeijer, Xiaoran Liang, Fernando P Hartwig, and Jack Bowden. The confidence interval
391 method for selecting valid instrumental variables. *Journal of the Royal Statistical Society: Series*
392 *B (Statistical Methodology)*, 83(4):752–776, 2021.

393 Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. 2010.

394 Haoran Xue and Wei Pan. Robust inference of bi-directional causal relationships in presence of
395 correlated pleiotropy with gwas summary data. *PLoS genetics*, 18(5):e1010205, 2022.

396 Haoran Xue, Xiaotong Shen, and Wei Pan. Constrained maximum likelihood-based mendelian
397 randomization robust to both correlated and uncorrelated pleiotropic effects. *The American Journal*
398 *of Human Genetics*, 108(7):1251–1269, 2021.

399 Xinyi Zhang, Linbo Wang, Stanislav Volgushev, and Dehan Kong. Fighting noise with noise: Causal
400 inference with many candidate instruments. *arXiv preprint arXiv:2203.09330*, 2022.