**##Title: MR-TRYX: Exploiting horizontal pleiotropy to infer novel causal pathways**

We appreciate the reviewer for the critical comments and constructive suggestions. We have carefully addressed the comments and changed the manuscript accordingly. We hope that the revised manuscript has adequately addressed the reviewer's comments.

**Reviewer #1**

**Major comments**

1**. My main comment is on how reliable the last part of the MR-TRYX approach –adjusting the MR estimates – is. In particular, it relies on the strong assumption that you can obtain an unbiased estimate of all of the relevant candidate traits P on the outcome Y. To their credit, on page 12 and in the discussion section the authors discuss many caveats and are upfront about the potential limitations. However, I feel the authors could be even more prudent and present the last part of the analysis as a way to tentatively gauge the direction of bias in the original MR estimate rather than presenting it as a bias-corrected MR estimate. MR-base is an extremely rich resource, and is expanding, but the MR-TRYX approach relies heavily on candidate traits being measurable and for which GWAS results exist. Therefore I feel a bit like you’re back in the world of observational studies with unmeasured confounding: you’re never sure whether you’ve adjusted for all the relevant candidate traits that bias the effect of X on Y.**

**One idea that came to mind is whether you could use negative controls or MRGxE to have some sense whether the adjustment through observable candidate traits is sufficient to eliminate bias. I mean, one way to purge the effect of X on Y from possible pleiotropic effects is to measure the relevant pleiotropic effects through observed candidate traits as the authors suggest.**

**An alternative approach is to purge the effect of X on Y from pleiotropic effects using no-relevance groups for whom X=0 by definition as in Chen et al. (2008), Van Kippersluis & Rietveld (2018), and Spiller et al. (2019). If both methods suggest a similar correction of the original IVW estimate, this raises confidence in the correction part of MR-TRYX.**

**Answer)** Thanks for your suggestions. We agreed that it is difficult to conclude that we’ve adjusted for all pleiotropic traits. Given the complexity of human genetics and its corresponding phenotypes, a realistic goal for our study is to obtain less biased estimates. We therefore propose the adjustment method as an additional MR method and recommend using this in combination with other approaches in a sensitivity analysis framework. In the simulation, our method outperformed over other existing methods for outliers in two sample MR design. Even though the performance of our method relies on the availability of GWAS summary statistics, we believe that our method provides a new perspective in addition to the existing methodology. For example, our LASSO-MVMR and graph tool in MR-TRYX would make it easier for users to estimate the direction of bias due to pleiotropic traits in the original MR estimate. We have also described this in the manuscript clearly so that the reader will be able to determine.

As you suggested, the use of negative control can validate our method. However, it is difficult to find the relevant exposure. For example, in case of higher level of BMI, the strength of association with outcomes of interest can be compared with that observed for lower level of BMI, which has different effect on the outcome. Other assumption includes that the negative control should be related to genetic factors in the same way as the original exposure without measurement errors [1, 2]. Therefore, it wouldn’t be ideal to use negative control in two sample setting that MR-TRYX used. We are currently working on to make MR-TRYX available in one sample setting, where use of negative control can be applied with individual data.

MR-TRYX is integrated with MR-Base, which includes GWAS results for the extensive range of phenotypes (n=605) traits from UK Biobank and 342 other complex traits and diseases obtained from 123 GWA studies. Furthermore, we’ve been collaborating with large biobank to collect data and increase the utilisation of MR-Base. With the increasing availability of complete summary results from GWAS consortia and MR-Base, we expect that the capability of MR-TRYX will increase.

**References**

[1] Sanderson E, Macdonald-Wallis C, Davey Smith G. Negative control exposure studies in the presence of measurement error: implications for attempted effect estimate calibration. Int J Epidemiol. 2018;47(2):587-596.

[2] Taylor AE, Davey Smith G, Munafò MR. Re: "Exposure to maternal smoking during pregnancy as a risk factor for tobacco use in adult offspring". Am J Epidemiol. 2014;180(9):959-60.

**2. My second main comment is that I believe the authors should talk more about the hypothesized data generating process under which their approach works. To me it seems that the approach works best when one has an additive model in which X and all the candidate traits P have additive effects, and where none of the candidate traits P is a mediator or collider in the relationship between X and Y. But what if P is a mediator of the relationship between X and Y? What if P is a collider? What if there are interaction effects between X and P?**

**Related, redundancy of candidate traits is currently determined by a statistical approach (LASSO) but shouldn’t redundancy not also be based on some theory or at least some idea of how the candidate trait relates to X and Y in e.g., a Directed Acyclical Graph (DAG)?**

**Answer)** MR-TRYX is based on two-sample MR [1]. Therefore, the basic setting of this framework (i.e. hypothesis generating and data handling) are the same as two-sample MR analysis.

In this study, we assumed that outliers in two sample MR setting should be associated with other traits than exposure, which have associations with the outcome. Therefore, we employed multivariable MR (MVMR) method to estimate the effect of candidate traits P on the outcome (pleiotropic pathway). MVMR provides less biased estimates where genetic variants are associated with several traits [2]. A recent study proved that MVMR does not introduce bias into the results when P is a collider of the relationship between the original exposure (X) and the outcome (Y) [3]. This is because the predicted value of Pis not dependent on the outcome in the MVMR analysis. In case of mediation, however, our adjustment method may not generate entirely reliable estimates as MVMR estimates the direct effect of the exposure on the outcome that doesn’t act via the mediator [3]. We have described this in Discussion as a potential limitation of MR-TRYX. Two-step MR can be applied to MR-TRYX to provide quantifiable estimates of the proportion of mediation in the future [4].

We used LASSO to avoid over correction for similar traits as MR-TRYX includes traits from the UK Biobank (UKB) that appear similar to each other (e.g. the same measurements but measured in the right / left side of body). Therefore we used LASSO to remove those redundant traits from the adjustment analysis model. The previous study showed that the Lasso method for variable selection can be used in linear IV models [5]. We provided additional simulations to prove the validity of LASSO step in MR-TRYX (see the answer to the reviewer 3).

**References**

[1] Lawlor DA. Commentary: Two-sample Mendelian randomization: opportunities and challenges. Int J Epidemiol. 2016;45(3):908-15.

[2] Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. Am J Epidemiol. 2015 Feb 15;181(4):251-60.

[3] Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. Int J Epidemiol. 2018. [Epub ahead of print]

[4] Relton CL, Davey Smith G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. Int J Epidemiol. 2012;41(1):161-76.

[5] Kang H, Zhang A, Cai TT, Small DS. Instrumental Variables Estimation with some Invalid Instruments and its Application to Mendelian Randomization. J Am Stat Assoc. 2016;111:132-144. Journal of the American Statistical Association, 111, 132–144.

**3. I think the description of multivariable MR is a bit confusing. In the introduction it is mentioned that the causal influence of the candidate trait on the outcome is obtained using MR excluding the original outlier; then on page 7 the authors talk about a set of T clumped instruments for both X and P; and then finally on page 10 it is mentioned that additional SNPs are introduced to instrument the candidate traits. Please fill me in on how exactly the multivariable MR is performed.**

**Answer)** Sorry for the confusion. In the two-sample MR approach, it is important to ensure that instruments for an exposure are independent. We clumped the extracted SNPs from GWAS (clumping r2 cut-off=0.001 and clumping window=10,000kb) against a reference dataset of similar ancestry to the samples where the GWAS was conducted. The clumping procedure has been implemented in MR-Base [1]. After we defined the candidate traits and performed LASSO regression, we applied MVMR to estimate the effect of remaining traits from LASSO on the outcome. MVMR analysis included the instruments for remaining traits [2]. Therefore, additional SNPs were introduced into MVMR analysis as the instruments for the candidate traits. We have introduced outlier removal methods in the introduction to suggest that those methods can be a form of cherry picking.

**References**

[1] Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, Tan VY, Yarmolinsky J, Shihab HA, Timpson NJ, Evans DM, Relton C, Martin RM, Davey Smith G, Gaunt TR, Haycock PC. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018.

[2] Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. Am J Epidemiol. 2015 Feb 15;181(4):251-60.

**4. Two comments on the selection of candidate traits and the differential use of false discovery rates:**

1. **Candidate traits are defined as those having an association with the outlier SNP at genome-wide significance level. However, since the pvalue in the GWAS is a function of both the effect size but also the sample size, this way of selecting traits is partly on basis of the accidental sample size of the discovery GWAS and not on basis of the potential influence on the MR estimate under study. Would it not be better to select traits on basis of the effect size (and hence the potential influence on the MR estimate)?**

**Answer)** Thanks for your suggestion. We agreed that the use of replicated GWAS effect size is ideal to reduce the impact of winner’s curse. As you pointed, the summary data registered in MR-Base database is currently biased towards discovery GWAS, with few results from replication studies. We added this in discussion part. Although we note that the use of the P value is not ideal as we mentioned in Method, we use them here to make high dimensional investigations more manageable. We believe GWAS significance is more widely accepted whilst cut-off for the effect size can be subjective. Also, the basic concepts of P value and effect size are almost similar, whilst SNP-traits effect sizes are difficult to be considered in statistical method as those may vary between age or gender as well as their unit.

1. **Then candidate traits are further reduced by looking at the effect of these traits on the outcome. Here a p-value of 0.05 is used. In the outlier detection part a p-value of 0.05 divided by the number of SNPs is used as a correction for multiple testing. For consistency, wouldn’t it be better to apply a similar correction to the 0.05 FDR here too?**

**Answer)** Once the candidate traits are identified, we tested the effect of one candidate traits on the original outcome. As it is a single test for single hypothesis (e.g. is the candidate trait A associated with the outcome?), we didn’t use Bonferroni’ correction.

**5. I don’t find Figure 2 very illuminating. Could you also present the actual estimates in a Table such that readers can judge the bias? How is “substantially different” on line 665 defined?**

Answer) Sorry for the confusion. Now the actual estimates are provided in Supplementary Table 3.

Thank you for the suggestion. We wrestled with how best to illustrate the bias in these simulations. Normally one would just present the average bias across a number of repeated simulations, but the problem with this is that we are not simulating directional pleiotropy, we are allowing the outliers to go in any direction, so across many simulations the average bias will tend towards 0 even if all individual simulations are themselves biased differently. In Figure 2 we are instead presenting the proportion of the simulations that are ‘significantly’ biased, i.e. has an effect estimate with CI not overlapping the true simulated effect. We believe this is the most appropriate way to depict these scenarios. Other ways are to show e.g. the mean squared error or the average absolute bias, but these are not easily interpretable so we opted against those options.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Estimates (S.E)** | | | | | | | | | |
|  | **Number of pleiotropic variants** | | | | | | | | | |
|  | 2 | 6 | 10 | 11 | 15 | 19 | 20 | 24 | 28 |
| **Simulate effect size (beta = 0)** | | | | | | | | | | |
| Raw | 0.000 (0.045) | 0.000 (0.080) | 0.002 (0.102) | 0.000 (0.108) | 0.001 (0.124) | 0.000 (0.136) | -0.001 (0.138) | 0.001 (0.148) | -0.003 (0.156) |
| Pleiotropy detected | | | | | | | | | | |
| Outlier adjusted | 0.000 (0.021) | 0.000 (0.065) | 0.002 (0.09) | 0.001 (0.099) | 0.001 (0.116) | 0.001 (0.128) | -0.002 (0.131) | 0.000 (0.143) | 0.000 (0.151) |
| Outlier removed (all) | 0.000 (0.022) | 0.000 (0.046) | 0.001 (0.065) | 0.001 (0.067) | 0.001 (0.081) | 0.001 (0.093) | -0.001 (0.095) | 0.000 (0.106) | 0.002 (0.114) |
| Outliers removed (candidates) | 0.000 (0.021) | 0.000 (0.041) | 0.000 (0.057) | 0.001 (0.061) | 0.001 (0.073) | 0.000 (0.085) | -0.002 (0.087) | 0.001 (0.098) | 0.002 (0.106) |
| Pleiotropy known | | | | | | | | | | |
| Outlier adjusted | 0.000 (0.021) | 0.000 (0.066) | 0.001 (0.091) | 0.000 (0.099) | 0.001 (0.116) | -0.001 (0.129) | -0.001 (0.131) | 0.001 (0.143) | -0.005 (0.152) |
| Outlier removed (all) | 0.000 (0.021) | -0.001 (0.046) | 0.001 (0.066) | 0.000 (0.068) | 0.000 (0.086) | 0.001 (0.107) | -0.002 (0.113) | -0.001 (0.138) | -0.004 (0.172) |
| Outliers removed (candidates) | 0.000 (0.021) | -0.001 (0.041) | 0.001 (0.058) | 0.000 (0.062) | 0.000 (0.082) | 0.000 (0.108) | -0.001 (0.116) | -0.001 (0.165) | -0.001 (0.349) |
| **Simulated effect size (beta = 0.2)** | | | | | | | | | |
| Raw | 0.201 (0.045) | 0.200 (0.081) | 0.200 (0.103) | 0.201 (0.108) | 0.201 (0.124) | 0.199 (0.136) | 0.200 (0.138) | 0.201 (0.148) | 0.201 (0.156) |
| Pleiotropy detected |  |  |  |  |  |  |  |  |  |
| Outlier adjusted | 0.200 (0.021) | 0.200 (0.066) | 0.199 (0.09) | 0.201 (0.099) | 0.201 (0.116) | 0.199 (0.129) | 0.198 (0.131) | 0.200 (0.142) | 0.202 (0.151) |
| Outlier removed (all) | 0.200 (0.022) | 0.200 (0.046) | 0.199 (0.065) | 0.201 (0.067) | 0.201 (0.081) | 0.198 (0.094) | 0.200 (0.096) | 0.200 (0.106) | 0.203 (0.115) |
| Outliers removed (candidates) | 0.199 (0.022) | 0.200 (0.042) | 0.200 (0.057) | 0.201 (0.061) | 0.201 (0.074) | 0.198 (0.086) | 0.199 (0.088) | 0.200 (0.098) | 0.203 (0.108) |
| Pleiotropy known |  |  |  |  |  |  |  |  |  |
| Outlier adjusted | 0.200 (0.021) | 0.198 (0.066) | 0.199 (0.092) | 0.200 (0.099) | 0.200 (0.116) | 0.199 (0.129) | 0.201 (0.131) | 0.201 (0.142) | 0.199 (0.151) |
| Outlier removed (all) | 0.200 (0.022) | 0.199 (0.046) | 0.201 (0.067) | 0.200 (0.068) | 0.200 (0.086) | 0.200 (0.107) | 0.201 (0.113) | 0.200 (0.137) | 0.195 (0.173) |
| Outliers removed (candidates) | 0.200 (0.021) | 0.200 (0.041) | 0.201 (0.058) | 0.200 (0.062) | 0.201 (0.082) | 0.200 (0.108) | 0.200 (0.116) | 0.200 (0.165) | 0.187 (0.365) |

**Minor comments**

1. **On page 1, line 56, you talk about ‘outliers’, but at this stage in the introduction it is not clearly defined what you mean by outliers. Perhaps you can help the reader by defining outliers as ‘suspicious SNPs’ or ‘SNPs exhibiting possible pleiotropic effects’?**  
   Answer)Revised.
2. **Page 4, lines 111-113, this part is very hard to understand for non-experts. Could you provide a bit more intuition or at least a good reference as you did before on line 102.**Answer) Sorry for the confusion. The ‘NO Measurement Error (NOME)” assumption indicates that the SNP-exposure association is equal to the true associations without error, rather than estimated [1]. Strong violation of this assumption may induce bias in ‘1st order’ IVW estimate towards the null [2]. Therefore, we used ‘2nd order’ weights to better acknowledge the full uncertainty in the SNP-exposure association estimates. The reference has been added.

**Reference**

[1] Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. Int J Epidemiol. 2016;45(6):1961-1974.

[2] Bowden J, Del Greco M F, Minelli C, Zhao Q, Lawlor DA, Sheehan NA, Thompson J, Davey Smith G. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. Int J Epidemiol. 2018. [Epub ahead of print]

1. **In empirical example 1, isn’t ibuprofen use a potential collider?**

**Answer)** It could be a potential collider as people who are prescribed aspirin, which is used to treat myocardial infarction, might be advised to avoid chronic ibuprofen. However, it has been reported that ibuprofen may affect cardiovascular risk. Therefore, we assumed that ibuprofen can be a putative risk factor for cardiovascular disease rather than a collider. Also, the estimate was obtained from MVMR, which is not subject to collider bias (see the answer to #Reviewer 1. Q-2.

**References**

[1] Cross PL. Effect of ibuprofen on cardioprotective effect of aspirin. Lancet. 2003 May 3;361(9368):1560-1.

1. **To fairly compare the confidence intervals across methods, and given that 1000 bootstrap replications is fair but not very large, would it be useful to also bootstrap the standard errors of the IVW approach under the same number of replications, to make sure that the differences in confidence intervals are due to the method of estimation and not the method of computing standard errors?**

**Answer)**

1. **I don’t understand the sentence on line 434-435 on the shrinkage step.**

**Answer)** Although heterogeneity is bad when we test the validity of instrument, it is necessary when we assess the strength of instrument. Suppose that Q statistics for exposure X1 and X2 are Qx1 and Qx2. Q statistics here indicates instrument relevance, which detects the level of heterogeneity to determine if a set of instruments can predict all exposures of interest. If both Qx1 and Qx2 are larger than the chosen critical value for X2 distribution, detection of heterogeneity suggests that the instruments can predict variation in both exposures. As multivariable MR requires more than one trait, SNP-exposure effect heterogeneity is necessary to obtain valid estimates.

1. **The legend and labels of the figures could be improved.**

**Answer)** Revised.

**Reviewer #2**

**1. The authors selected a P-value threshold of 5x10-8 for selecting SNPs exhibiting horizontal pleiotropy in associations with secondary traits and outcomes, which may be too conservative especially given GWAS with smaller sample sizes.**

**Answer)** Thanks for your comment. We used a conservative threshold because we assumed that each trait is independent, and the number of association test of all possible traits is more than 500,000. The use of conservative threshold helps to reduce the number of false-positive associations arising from the vast number of statistical tests performed.

**2. It would be useful if the authors could add the number of genetic instruments in the Results for the four practical examples that they use.**

Answer) We provided the number of genetic instruments used in the analysis in the Table 1 and 2.

**Reviewer #3**

**Major comments**

**1. METHOD**

**1) To determine the pool of candidate exposures, why use IVW (and excluding the outlier SNP) and not directly multivariable MR with both the hypothesized exposure and the candidate exposure to select candidate exposures "causal" to the outcome conditional on the hypothesized exposure?**

**Answer)** Multivariable MR estimates the direct effect of each exposure included in the model on the outcome. Therefore, if any of the effect on the exposure of interest is mediated by other exposures included this mediated effect will not appear as part of the direct effect.MR TRYX however is a univariable MR and estimates the total effect of the exposure of interest on the outcome, this will include any effect that is mediated through other potential exposures (including those that are affected by pleiotropic SNPs). MR TRYX removes the effect of pleiotropy in the conventional MR estimates whilst still estimating the total effect of the exposure on the outcome. We used IVW to estimates the total effect of the exposure of interest on the outcome.

**2) LASSO provides the best estimator in terms of MSE, therefore in terms of prediction, thus it introduced a bias to decrease the variance. However, the aim in MR is to have the best estimator in terms of unbiased causal estimate, not in terms of prediction. How does the LASSO selection step influence the bias in the causal estimate?**

**Answer)** As we described above (#Reviewer 3; Q1-1), MR-TRYX calculate the total effect of the original exposure on the outcome. We suggested the adjustment method to correct the bias due to pleiotropy using MVMR. For MVMR analysis, we applied LASSO step to reduce the number of overlapped variables. Under the sufficient condition for identification that less than 50% of variables are redundant, the LASSO selection may select the valid traits. However, this step could influence the power of MVMR, which affects the validity of adjusted estimates. We described this potential limitation in the Discussion. Additionally, we provided additional simulations to show the validity of the LASSO step (Please see #Reviewer 3; Q2-4).

**3) I believe LASSO is very sensitive and the selection procedure is not very stable. For instance, I wouldn't be surprise to see the set of selected candidate exposures be affected by the clumping procedure. How stable is the LASSO selection procedure in MR-TRYX? Does the LASSO procedure favor some candidature exposures, such as exposures with many instruments? For instance, in example 1 of the real data, Height has been selected as a candidate exposure for CHD, wouldn't body mass index/weight/hip circumference be more relevant? Was BMI part of the original pool of candidate traits? If so, was Height selected by LASSO because Height has more instruments and hence provides a better predicted outcome (from the MSE standpoint) even if it is not the "causal" exposure?**

**Answer)** The identification of candidate traits is not influenced by the selection procedure of LASSO as the LASSO step is applied afterwards. This step however can influence adjusted estimates. We simulated whether the number of instruments for the candidate trait affect the ability of LASSO to select candidate traits.

**[Additional simulation 1]** In this additional simulation, we generated traits X1 and X2 which are instrumented by 100 SNPs and 20 SNPs, respectively. We set that 10 of them are pleiotropic and associated with both of X1 and X2. Four scenarios are considered:

1. X1 has an effect on Y (β=0.3), where X2 has no effect on Y (β=0.0)
2. X1 has no effect on Y (β=0.0), where X2 has an effect on Y (β=0.3)
3. Both of X1 and X2 have effects on Y (β=0.3)
4. Neither of X1 and X2 have effects on Y (β=0.0)

In each scenario, we performed the LASSO regression to see if the LASSO selects the traits in a manner that is determined by which trait is causal, rather than which trait has more instruments. The simulation was replicated 1000 times in each scenario. The results showed that the LASSO step worked well across the scenarios. The LASSO kept both X1 and X2 when both traits have effects on the outcome Y with a probability of 1.000 (1000 times per 1000 simulations), regardless of the number of instruments for the traits. Whilst, the LASSO removed both traits with a probability of 0.693 when both of X1 and X2 have no effects on Y. In this scenario, the trait X1 was removed with a probability of 0.775 where the trait X2 was removed with a probability of 0.783. In scenario 1, LASSO kept the trait X1 100% but failed to remove X2 sometimes, which is instrumented by a small number of variants and has no effect on Y, with a probability of 0.473. In scenario 2, X2 wasn’t removed by the LASSO step but X1 was remained with a probability of 0.467. In conclusion, the results suggested that the LASSO regression may not favour to the candidate traits with a larger number of instruments when a trait is weakly associated with the outcome.

**4) On a related note, would there be any benefit in having a step selection procedure where the instruments associated with unselected exposures at step i are removed at step i+1.**

**Answer)** In terms of MVMR, MR-TRYX keeps only the instruments associated with the exposures (or candidate traits) that are involved in the analysis. Unselected candidate traits and related instruments are automatically excluded from the analysis.

**5) To free MR-TRYX of the disadvantage of the outlier detection step, could the whole MR-TRYX framework be applied to all the instruments? Would the results differ much? My intuition would be that the results would be very similar in case the selection of candidate exposure is done using a conditional model on the hypothesized exposure (multivariable MR instead of IVW), is that a reasonable assumption?**

**Answer)** The aim of this paper is to utilise the outliers in MR model that has been treated as nuisance. We suppose that there would be a reason for the outlier to happen in MR analysis and we took a different approach. In present study, our preference is to provide an insight of pleiotropic effect and to show how to identify new pathways for disease outcomes exploiting outliers.

**6) Since MR-TRYX is fully dependent on MR-Base, how does MR-Base deal with ancestry? with sample overlap?**

**Answer)** As you pointed out, the association estimates can be biased when candidate traits and outcome studies are conducted in different populations. MR-Base provides meta-data on population characteristics [1], including ancestry, and geographic origin to guide the user in selecting the most appropriate design for their analysis. So far, the majority of studies are derived from European population, but we are expanding the scope of MR-Base by collecting summary statistics from various populations. Therefore, it is recommended to ensure that selected candidate traits and the outcome were obtained from the same population. We added this in the Discussion as follows:

“Fifth, since MR-TRYX uses the resource from MR-Base, it is recommended that the user acknowledge the limitation and restriction of MR-Base [1]. For example, the population should be the same for the exposure (or the candidate traits) and the outcome traits to avoid mis-estimation of the magnitude of the association. The users should consider modifying their analyses when the limitation indicated above are avoidable.”

**7) Instead of manually removed redundant traits, especially traits similar to the outcome, would it be a good idea to use genetic correlations using LD score?**

**Answer)** Thanks for your suggestion. MR-TRYX was designed to create a user-friendly environment. User can use an automated approach for identifying candidate traits, but as part of sensitivity analysis, they can curate the pool of candidate traits manually according to their study hypothesis. We have tried to employ LD score to select the traits, but there was a practical difficulty in configuration of the package.

**2. SIMULATIONS**

**1) I feel like the simulation design could be extended to span a larger range of scenarios.**

Answer) Thanks for the suggestion. The simulations in this study were designed to show the performance of MR-TRYX with respect to adjusting for pleiotropic bias. We performed simulations to evaluate how our new approach of adjusting outlier effects compares against standard analyses, and against outlier removal. Two scenarios of simulations are performed, the first using a null causal effect (gy = 0), and the second a positive causal effect (gy = 0.2). In each set, four methods are considered for handling outliers. We ran the latter three methods by detecting outliers empirically, but also, ran the hypothetical case in which we know the pleiotropic variants a priori for comparison. To provide a clear message, we would like to keep our simulation simple.

**2) Most of the simulation scenarios are at a disadvantage for outlier-based method since the proportion of SNPs with pleiotropic effects is mostly high (6/8 scenarios with at least 30% of pleiotropic SNPs).**

Answer) In the simulation, we asked if adjustment for that pathway improve the original exposure-outcome hypothesis where it is possible to identify the pathway through which an outlier SNP has horizontal pleiotropic effect. The graph presents the trend of proportion of estimates that are 1) detected at conventional threshold of statistical significance and 2) that are substantially different from the simulated effect due to pleiotropic variants. We have revised the figure legend to provide more clear information.

**3) I didn't find any mention of the direction of the pleiotropic effects which is of importance for outlier-based methods.**

Answer) We performed simulation to test if the performance of MR-TRYX is influenced by the direction of pleiotropy.

Thank you for pointing this out – we realise we haven’t been specific enough in the methods about the parameters used for simulations. That section is now updated with more detailed information about how data was simulated. To answer the question – we are simulating pleiotropic effects under a balanced pleiotropic model, in that the SNPs which have a pleiotropic effect have an influence on the exposure that could be positive or negative, and have an additional influence on the outcome mediated by another trait that could also be positive or negative. So, the SNP-exposure and pleiotropic effects are uncorrelated and could be in either direction.

**[Additional simulation 2]** We generated random phenotype X1 and X2, which have the opposite effects on the outcome Y (simulated effect size = -0.3 and 0.3, respectively), considering those phenotypes share some variants (20 variants among 80 variants; which indicates that 20 variants are pleiotropic). We then ran MVMR analysis with and without the LASSO regression step to see if the direction of pleiotropy influences the performance of LASSO MVMR. As a result, we found that the method is reasonably robust to the presence of directional pleiotropy. In term of the performance of the LASSO step, it didn’t remove neither of X1 and X2. Also, the result showed that MVMR, which is used in adjustment method, yielded more precise estimates. This simulation suggested that the adjustment step though the LASSO MVMR is valid even though the direction of the pleiotropic effects is different.

**Table R3-2. The result of the simulation where the direction of pleiotropic effects is opposite.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | NSNPs | IVW estimate (beta, se) | P value |
| MVMR with /without LASSO |  |  |  |
| X1 | 29 | -0.316 (0.013) | 2.880e-134 |
| X2 | 32 | 0.304 (0.013) | 1.028e-118 |

**4) It is not clear how big the pool of potential candidate exposures is and if which part of the MR-TRYX framework is run in the simulations. Is the LASSO step tested? Are the candidate exposures considered as known? If so it favors MR-TRYX.**

Answer) We used MR-Base database to search for candidate traits. MR-Base database includes 342 complex traits and diseases obtained from 123 GWAS studies, as well as GWAS results of phenotypes (n=605) from UK Biobank. As we described in the answer to Q. 2-1, the simulations in this study was performed to compare adjustment methods to standard analyses, and outlier removal. The use of LASSO in MR approach has been proved by the previous studies [1, 2]. We performed additional simulations to evaluate the validity of the LASSO-MVMR.

**[Additional simulation 3]** In this additional simulation we asked how accurately the LASSO MVMR predicts outlier effect to avoid the bias from over-adjustment. We generated traits X1 and X2 which are instrumented by 50 SNPs, respectively. We set that X1 and X2 have a 0.2 influence on the outcome Y. For each trait, we created three extra traits which are the same as the original trait (e.g. X1a, X1b, X1c and X2a, X2b, X1c). First, we hypothesised that there is no pleiotropy (X1 and X2 are predicted by independent 50 variants) to test how often LASSO removes extra traits correctly. We then repeated the analyses assuming 20 of the variants have pleiotropic effect (which means that 20 of them are associated with both X1 and X2). In each scenario, we performed normal MVMR and LASSO MVMR. The results showed that the LASSO removed redundant traits X1a, X2a, X2b as well as X1 and X2. Here we have X1b and X1c but not X1, but it is still valid as the LASSO captured X1 through its proxies. This would not cause the problem since the purpose of LASSO MVMR is to estimate how much of outlier effect (pleiotropic pathways) we should consider. In the LASSO MVMR, X1b and X1c have effects of 0.164 and 0.036, then the adjustment method will correct for both X1b and X1c (0.164+0.036 = 0.2), which is the same as the effect of X1 on Y. Therefore, it can be suggested that the LASSO MVMR corrects for effect of candidate traits appropriately. As MR-TRYX repeats this step 1000 times, the LASSO MVMR in MR-TRYX would perform better than MVMR on its own.

**Table R3-1. The result of the simulation.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | No pleiotropy | | |  | Pleiotropy | | |
|  | NSNPs | IVW estimate (beta, se) | P value |  | NSNPs | IVW estimate (beta, se) | P value |
| MVMR | | | | | | | |
| X1 | 31 | -0.076 (0.326) | 0.815 |  | 36 | 0.529 (0.397) | 0.183 |
| X2 | 31 | -0.07 (0.332) | 0.833 |  | 25 | 0.166 (0.363) | 0.647 |
| X1a | 29 | -0.133 (0.197) | 0.499 |  | 32 | -0.272 (0.198) | 0.171 |
| X1b | 29 | 0.261 (0.196) | 0.185 |  | 33 | 0.096 (0.222) | 0.666 |
| X1c | 29 | 0.15 (0.205) | 0.464 |  | 33 | -0.168 (0.222) | 0.451 |
| X2a | 25 | -0.035 (0.212) | 0.867 |  | 25 | 0.010 (0.208) | 0.96 |
| X2b | 27 | -0.059 (0.218) | 0.787 |  | 22 | -0.106 (0.270) | 0.694 |
| X2c | 25 | 0.345 (0.207) | 0.095 |  | 23 | 0.133 (0.207) | 0.519 |
| LASSO MVMR | | | | | | | |
| X1b | 29 | 0.164 (0.137) | 0.233 | X1 | 36 | 0.044 (0.215) | 0.839 |
| X1c | 29 | 0.036 (0.136) | 0.789 | X2 | 25 | 0.080 (0.201) | 0.691 |
| X2c | 25 | 0.181 (0.014) | 5.38E-38 | X1b | 33 | 0.138 (0.214) | 0.520 |
|  |  |  |  | X2c | 23 | 0.125 (0.201) | 0.535 |

**Reference**

[1] Windmeijer F, Farbmacher H, Davies N, Davey Smith G. On the use of the lasso for instrumental variables estimation with some invalid instruments. Journal of the American Statistical Association. 2018 Oct 22:1-2.

[2] Hemani G, Bowden J, Haycock PC, Zheng J, Davis O, Flach P, Gaunt TR, Davey Smith G. Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome. BioRxiv. 2017 Jan 1:173682.

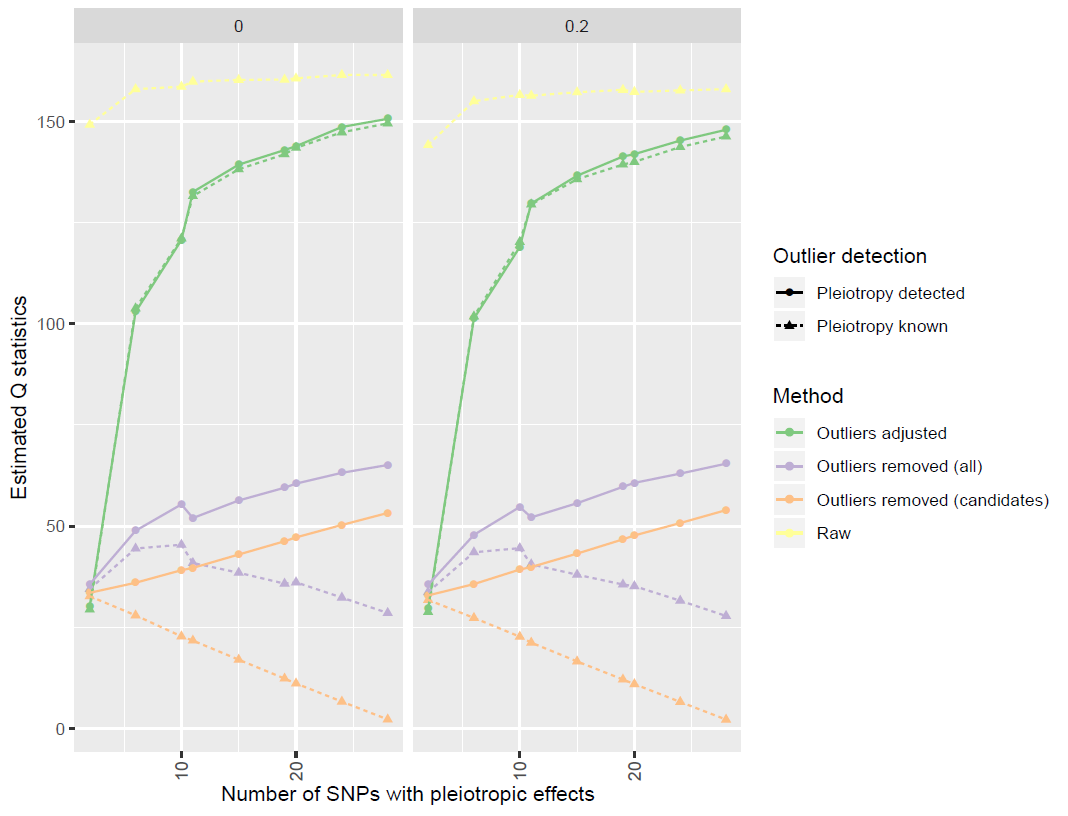
**5) It would be interesting to see how the multivariable MR with the hypothesized exposure + the candidate exposure(s) (meaning without adjusting for outliers) compares to the tested methods. Is it perfectly equivalent to adjusting for outliers in MR-TRYX?**

Answer)

**[Additional simulation 4]**

**6) In the real data, the authors use the reduction in Q statistic to asses how successful the MR-TRYX outlier adjustment is, but I have the intuition that any adjustment should reduce the Q statistic (by including an additional variable in a linear model, the residual variance decreases). How does the Q statistic compare between removing the outliers and adjusting for outliers? So, it would be nice to have such quantitative results in the simulations.**

Answer) Our adjustment model doesn’t include an extra term in the revised X-Y association. Instead of adding variables, we correct the G-Y association by subtracting estimated pleiotropic effect obtained from LASSO MVMR from the total effect. The Q statistics derived from our method is not necessarily decreased indeed when we adjust the pleiotropy in the wrong direction (Please see the example 4). We have added a supplementary figure to show the change in q stats over the different simulation scenarios.



**3. RESULTS**

**1) The 4 examples highlighted by the authors are relevant and of interest however, we are lacking more global results insofar as possible. For instance, among a certain set of exposure-outcome traits, how often are outliers detected / are candidate exposures detected? What is the average number of candidate exposures per outlier SNP / per hypothesized exposure-outcome pair? How many new exposures can be discovered using MR-TRYX?**

Answer) When a numerous SNPs (n > 1) are used in MR analysis, it is possible that some variants are valid instruments, but others are not. The number of invalid instruments (so called as outliers) and candidate exposures can vary depending on the research hypothesis. The primary goal of this study is to suggest the idea to overcome the problem of pleiotropy in the current MR model and to open the possibility of hypothesis free screen for potential exposure. In this paper, therefore, we didn’t examine all possible exposure-outcome associations to let the users explore their research hypothesis where horizontal pleiotropy exists. We believe that the 4 examples we provided would help potential users understand. Currently, 11 billion SNP-trait association from 1673 GWAS are available in MR-TRYX.

**2) Reduction in the Q statistic is a way for the authors to assess the success of their adjustment, however, I believe that when adjusting for an additional variable in a linear model setting, the Q statistic automatically decreases. How "significant" are these observed reductions in the Q statistic value? Is heterogeneity still found after MR-TRYX adjustment, is Q still significant?**

Answer)

**Minor comments**

**1. l149: did the authors mean figure 1b?**

Answer) Sorry for the confusion. It has been revised.

**2. typo? l222: "To perform multivariable MR, more SNPs were introduced into the analysis that instrument the candidate traits.", pardon my misunderstanding of the sentence, do the authors mean "more SNPs, that are instruments for the candidate traits, were introduced into the analysis"?**

Answer) Sorry for the confusion. To convey clear message, this has been revised.

**3. Figure 1: Number SNPs represented in b) and c) is different**

Answer) Revised.

**4. Figure 2: it would be nice to specify the total number of SNPs (30) somewhere in the x-axis label or having proportions.**

Answer) Revised.