Exploiting horizontal pleiotropy to infer new causal pathways

Yoonsu Cho1, Philip C Haycock1, Tom Gaunt1, George Davey Smith1, Gibran Hemani1

1 MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

## Abstract

**Background:** Violations in the assumptions of Mendelian randomization (MR) can introduce bias and heterogeneity in the causal estimate. A major source of heterogeneity is horizontal pleiotropy, where an instrumenting single nucleotide polymorphism (SNP) influences the outcome through pathways which bypass the exposure. Those SNPs that induce heterogeneity in MR are typically treated as a nuisance, but they could be a powerful gateway for learning novel pathways to the traits under investigation.

**Methods:** Following the advice of William Bateson to “TReasure Your eXceptions”, we developed the MR-TRYX framework (<https://github.com/explodecomputer/tryx>). Here, we begin with a single exposure-outcome hypothesis and perform radial inverse variance weighted 2-sample MR analysis. Outliers are then detected using heterogeneity statistics, and we search through the MR-Base database of GWAS summary statistics to identify other (“candidate”) traits that associate with the outliers. We then use multivariable MR analysis to test the extent to which horizontal pleiotropy with the candidate trait can explain the heterogeneity identified in the original exposure-outcome analysis. In doing so, MR-TRYX identifies novel traits influencing the outcome, and accounts for some of the heterogeneity in the original exposure-outcome analysis.

**Results:** Through simulations we showed that commonly used outlier removal methods can increase type 1 error rates, but SNP effect adjustment can improve power without the increase in type 1 error rates. We illustrated the use of MR-TRYX by estimating the causal effect of: i) systolic blood pressure (SBP) and coronary heart disease (CHD); ii) education level (year of schooling) and body mass index (BMI); iii) urate and CHD; and iii) sleep duration and schizophrenia.

**Conclusion:** We show that incorporating broad phenotypic information to model horizontal pleiotropy in MR analysis can improve power through reducing heterogeneity and build a more detailed impression of the causal influences on complex traits.

## Introduction

Mendelian randomization is now widely used to infer the causal influence of one trait (the exposure) on another (the outcome) (1, 2). It is generally performed by obtaining instruments for an exposure through genome-wide association studies (GWAS). If the instruments are valid, in that they influence the outcome only through the exposure (vertical pleiotropy), then they will each provide an independent, unbiased estimate of the causal effect of the exposure on the outcome (3). Meta-analysing these estimates can provide a more precise estimate of the causal relationship between the exposure and the outcome (4, 5). If, however, some of the instruments are invalid, particularly because they additionally influence the outcome through pathways that do not go through the exposure (horizontal pleiotropy) (3), then the causal effect estimate is liable to be biased. To-date, MR method development has viewed horizontal pleiotropy as a nuisance that needs to be factored out of the meta-analysis (6-9). Departing from this stance, here we exploit horizontal pleiotropy as an opportunity to identify new traits that putatively influence the outcome. We then use this knowledge to improve the original exposure-outcome estimates.

A crucial feature of MR is that it can be performed using only GWAS summary data, where the causal effect estimate can be obtained solely from the association results of the instrumenting SNPs on the exposure and on the outcome (5). This means that causal inference between two traits can be made even if they have never been measured together in the same samples. Complete GWAS summary results have now been collected from thousands of GWAS analyses (9), meaning that one can search the database of GWAS results for candidate traits that might be influenced by the outliers. In turn, the causal influence of each of those candidate traits on the outcome can be estimated using MR by identifying their instruments (and excluding the original outlier). Should any of these candidate traits putatively associate with the outcome then this goes some way towards explaining the horizontal pleiotropic effect that was exhibited by the outlier SNP in the initial exposure-outcome hypothesis.

Several methods exist for identifying outliers in MR, each likely to be sensitive to different patterns of horizontal pleiotropy. Cook’s distance can be used as to measure the influence of a particular SNP on the combined estimate from all SNPs (10), labelling SNPs with large influences as being outliers. Steiger filtering removes those SNPs that do not explain substantially more of the variance in the exposure trait than in the outcome, attempting to guard against using SNPs as instruments that is likely associated with the outcome through other pathway than the exposure (11). Finally, meta-analysis tools can be used to evaluate if a particular study contributes disproportionately to the heterogeneity between the estimates obtained from the set of instruments, and this has been adapted recently to detect outliers in MR analysis (12-14). A potential limitation of heterogeneity-based outlier removal is that this practice is an implicit form of cherry picking (9, 15). While outlier removal can certainly improve power by reducing noise in estimation, it could also potentially induce higher type 1 error rates, which we go on to explore through simulations.

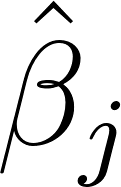
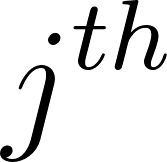
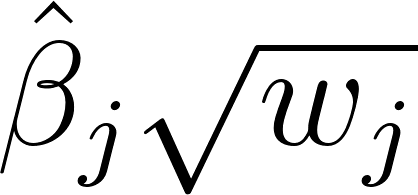
Recent large-scale MR scans have indicated that horizontal pleiotropy is widespread based on systematic analysis of heterogeneity (16, 17). This suggests that many SNPs used as instruments are likely to associate with other traits, which in turn might associate with the original outcome of interest – hence giving rise to heterogeneity. As such we have a tremendous opportunity to identify novel pathways through exploiting outliers. Equipped with automated MR analysis software, outlier detection methods and a database of complete GWAS summary datasets, we developed MR-TRYX (from the phrase coined by William Bateson, “Treasure your exceptions”) (18), a framework for identifying novel putative causal factors when performing a simple exposure-outcome analysis. In this paper we present simulations to show how knowledge of horizontal pathways can be used to discover novel putative causal factors for an outcome of interest, and to also improve the power and reliability of the original exposure-outcome association. We apply MR-TRYX to several example analyses to demonstrate its potential utility.

## Methods

### Overview of MR-TRYX

Figure 1 and 2 show an overview of the approach. MR-TRYX analysis is applied to a hypothesised exposure-outcome analysis and it has two objectives. The first is to use outliers in the original exposure-outcome analysis to identify novel putative factors that influence the exposure and/or the outcome (Figure 1). The second is to re-estimate the original exposure-outcome analysis by adjusting outlier SNPs for the horizontal pleiotropic pathways that might arise through the novel putative associations (Figure 2).

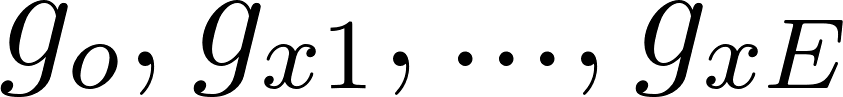
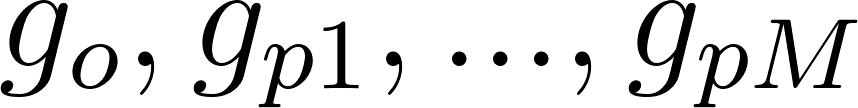
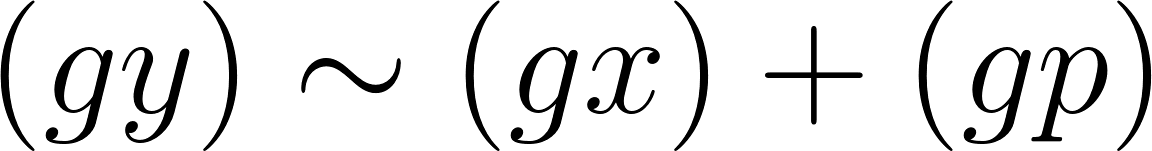
### Outlier detection

We use the RadialMR R package (https://github.com/WSpiller/RadialMR), which identifies outliers through radial IVW analysis by considering the contribution of each variant to overall heterogeneity within the set of individual ratio estimates. Details of Radial MR are described elsewhere (13). Let [](about:blank) represent the ratio estimate obtained from the [](about:blank) genetic variant in the analysis, and [](about:blank) represent the corresponding weighting. Radial MR regresses the set of ratio estimates multiplied by the positive square root of their corresponding weighting ([](about:blank)) upon the positive square root of their weighting ([](about:blank)). As a consequence, for each variant the residual is proportional to its contribution to heterogeneity using either Cochran’s Q statistic or Rucker’s Q statistic for IVW or MR Egger respectively. As a practical approach, we adopted a conservative threshold for identifying outliers, dividing 0.05 by the number of SNPs as a correction for multiple testing, though users can use other approaches through the software. We note that the use of arbitrary thresholds is problematic, but we use them here to make high dimensional investigations more manageable. We employed modified 2nd order weights throughout this paper to avoid problems arising due to the no measurement error in the exposure (NOME) assumption (19), assuming a multiplicative random effects model if any residual heterogeneity was detected.

### Candidate trait detection

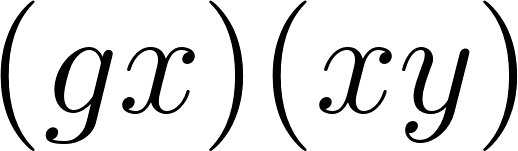
Traits associated with the detected outliers could causally influence the outcome. MR-TRYX searches the MR-Base database to identify the traits that have associations with the detected outliers. By default, we limit the search to traits for which the GWAS results have greater than 500,000 SNPs and sample sizes exceeding 5,000. Traits that have an association with outlier SNPs at genome-wide threshold (p < 5 x 10-8; in keeping with traditional GWAS thresholds used for instrument selection) are regarded as potential risk factors for the outcome and defined as “candidate traits”. Each candidate trait is tested for its influence on the original outcome (Figure 1) using the IVW random effects model. Further, the candidate traits that are putatively associated with the outcome (p <0.05) are used for MR analysis on the exposure to elucidate whether candidate traits explain the relationship between the exposure and the outcome independently.

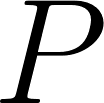
#### Obtaining candidate trait - outcome effect estimates

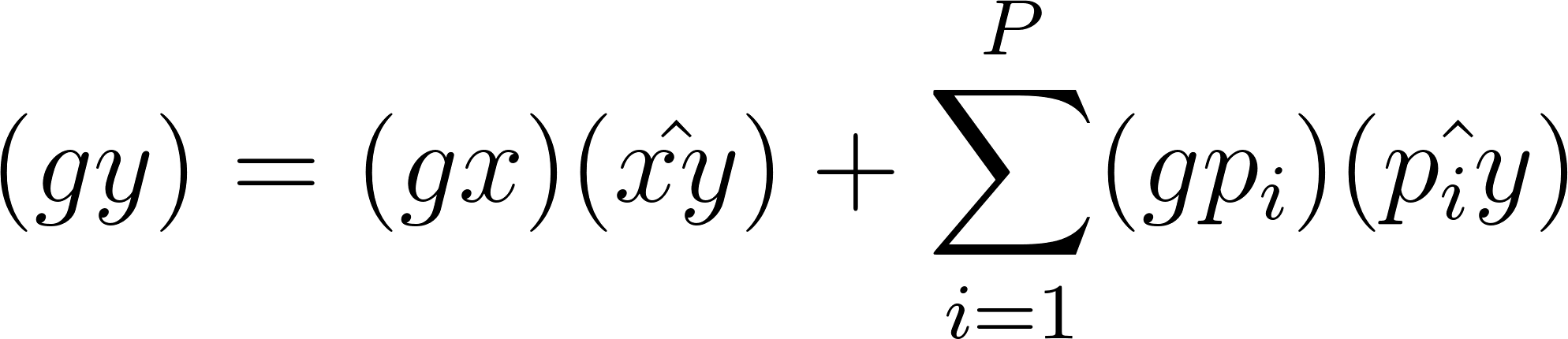
Suppose we have [](about:blank) instruments for the exposure [](about:blank) where [](about:blank) is an outlier in the x-y MR analysis due to an association with candidate trait [](about:blank). Also, [](about:blank) has [](about:blank) genetic instruments. To obtain the influence of [](about:blank) uncontaminated by shared genetic effects between [](about:blank) and [](about:blank), we obtain a unique list of [](about:blank) clumped instruments for both [](about:blank) and [](about:blank), and then obtain the genetic effects of each of these SNPs on the exposure [](about:blank), candidate trait [](about:blank), and outcome [](about:blank). Finally, we estimate the causal influence of [](about:blank) on [](about:blank) conditioning on [](about:blank) by regression [](about:blank) weighted by the inverse of the variance of the [](about:blank) estimates.

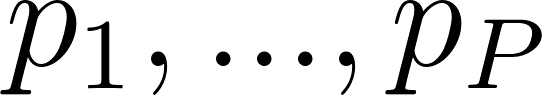
In the case of an outlier SNP associating with many candidate traits we first apply a LASSO regression of [](about:blank) and use cross validation to obtain the shrinkage parameter that minimises the residual squared error. We retain only the candidate traits that are putatively associated with the outcome and have non-zero effects after shrinkage. Then we apply remaining traits in a multivariable model with [](about:blank) against the outcome, as described above. We perform the LASSO step because many traits in the MR-Base database have considerable overlap and redundancy, and the statistical power of multivariable analysis depends on the heterogeneity between the genetic effects on the exposure variables (20). Using LASSO therefore automates the removal of redundant traits. With the remaining traits we then obtain estimates of [](about:blank) that are conditionally independent of *x* and amongst all *P* traits by combining them in a multivariable analysis on the outcome *y*.

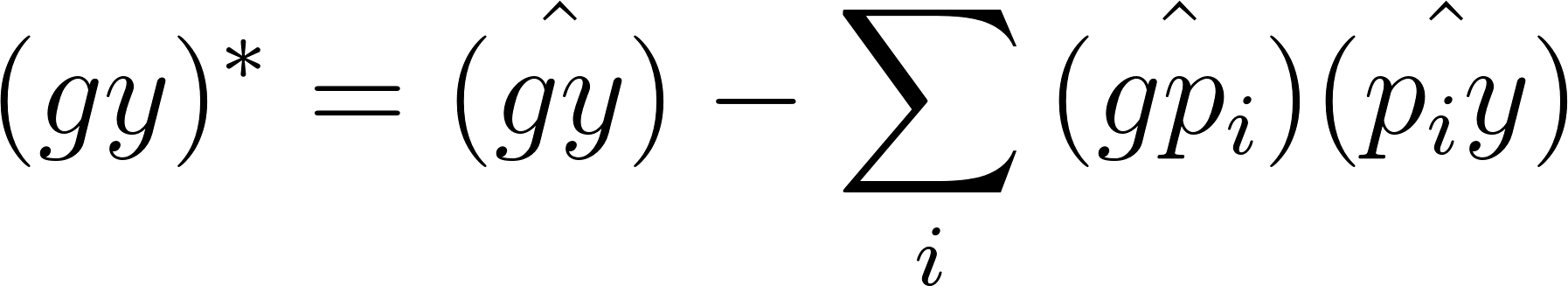
### Adjusting exposure-outcome associations for known candidate-trait associations

An illustration of how outliers arise in MR analyses is shown in Figure 2. If a SNP [](about:blank) has some influence on exposure [](about:blank), and [](about:blank) has some influence on outcome [](about:blank), the SNP effect on [](about:blank) is expected to be [](about:blank), where [](about:blank) is the SNP effect on [](about:blank) and [](about:blank) is the causal effect of [](about:blank) on [](about:blank). Any substantive difference between [](about:blank) and [](about:blank) could be due to an additional influence on [](about:blank) arising through the SNP’s effect through an alternative pathway.

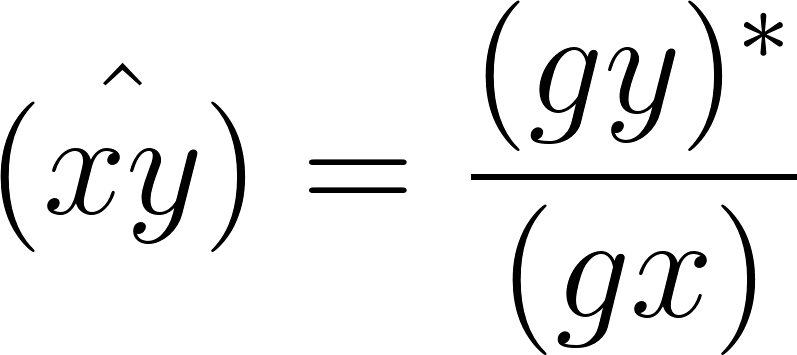
If a SNP influences a ‘candidate trait’, [](about:blank), which in turn influences the outcome (or the exposure and the outcome), then the SNP’s influence on the exposure and the outcome will be a combination of its direct effects and indirect effects through the candidate trait. If we have estimates of how the candidate trait influences the outcome, then we can adjust the original SNP-outcome estimate to the effect that it would have exhibited had it not been influencing the candidate trait. In other words, we can obtain an adjusted SNP-outcome effect conditional on the ‘candidate-trait – exposure’ and ‘candidate-trait – outcome’ effects. If the SNP influences [](about:blank) independent candidate traits (as selected from the LASSO step), then the expected effect of the SNP on [](about:blank) is

[](about:blank)

Hence, the effect of the SNP on the outcome adjusted for alternative pathways [](about:blank) is

[](about:blank)

We use parametric bootstraps to estimate the standard error of the [](about:blank) estimate, where 1000 resamples of [](about:blank), [](about:blank) and [](about:blank) are obtained based on their respective standard errors and the standard deviation of the resultant estimate, represents its standard error. Finally, an adjusted effect estimate of [](about:blank) due to SNP [](about:blank) is obtained through the Wald ratio

[](about:blank)

### Simulations

We assess the performance of MR-TRYX with respect to adjusting for pleiotropic bias due to outliers. In these simulations we ask: if we can identify the pathway through which an outlier SNP has horizontal pleiotropic effect, can adjustment for that pathway improve the original exposure-outcome hypothesis? Two scenarios of simulations are performed, the first using a null causal effect (), and the second a positive causal effect (). In each set, four methods are considered for handling outliers:

1. Raw, where all SNPs are used in a standard IVW analysis.
2. Outliers adjusted, where the outlier SNPs are adjusted for effect of the candidate trait on the outcome.
3. All outliers removed, where all detected outliers are removed.
4. Candidate outliers removed, where only outliers that are found to influence a candidate trait are removed.

We run the latter three methods by detecting outliers empirically, but also, run the hypothetical case in which we know the pleiotropic variants *a priori* for comparison. Individual level data is generated including 5000 samples. The results for each case represent the mean values for 1000 simulated datasets.

### Empirical analysis

As applied examples, we chose two robust findings and two controversial findings that are potentially biased due to pleiotropy: i) systolic blood pressure (SBP) and coronary heart disease (CHD); ii) education level (year of schooling) and body mass index (BMI); iii) urate and CHD; and iii) sleep duration and schizophrenia. Those examples were chosen based on previous findings (21-24) to illustrate how pleiotropic variants can be used to identify other pathways and be ruled out to estimate the causal effect of original exposure on the outcome independent of pleiotropic bias.

Summary statistics (beta coefficients and SEs) for the associations of the SNPs with each exposure were obtained from the publicly available GWAS database (Supplementary Table S1). Selected SNPs were harmonised for the analysis, excluding linkage disequilibrium and palindromic SNP. We primarily used the two-sample MR inverse-variance weighted (IVW) method to obtain causal estimate between exposures and outcomes. A number of sensitivity analyses were applied to ensure the robustness of the findings and validity of genetic instrument; the MR-Egger, weighted median and mode approach.

Outliers were detected among the instruments for each exposure. We searched MR-Base database to identify the candidate traits that are associated with outliers (p < 5 x 10-8). We then perform multivariable MR analysis to test which candidate trait can explain the heterogeneity in the original exposure-outcome analysis.

To illustrate how the pleiotropic association affects the result, we re-estimated the association of each exposure and each outcome using the same methods that used for simulations: a) using all SNPs, b) using SNPs excluding all outliers, c) using SNPs excluding “pleiotropic outliers” that are associated with candidate traits, and 4) adjusting effect of the candidate trait on the outcome. All analyses were performed assuming random effect model.

All analyses were conducted with the two-sample MR package of MR-Base (<https://github.com/MRCIEU/TwoSampleMR>) and MR-TRYX (https://github.com/explodecomputer/tryx) in R statistical software (ver 3.4.1).

## Results

### Adjusting SNP effects for pleiotropic pathways can improve exposure-outcome causal estimates

IVW effect estimates are liable to be biased when at least some of the instrumenting SNPs exhibit horizontal pleiotropy, and those SNPs tend to contribute disproportionately towards the heterogeneity in the effect estimate. We performed simulations to evaluate how outlier removal compares against standard analyses, and against our new approach of adjusting outlier effects given knowledge of horizontal pleiotropy pathways.

Our simulations show that balanced horizontal pleiotropy lead to elevated type 1 error rates for the ‘raw’ approach (Figure 2A). Statistical power is highest when outliers are removed, but at a cost of substantially higher type 1 error. Type 1 error rates are maintained at expected levels when adjusting for outliers, and the power improves over the ‘raw’ approach. A similar pattern of results among the three methods is seen for the likelihood of estimates being biased, with outlier removal and raw estimates performing worse than outlier adjustment. Removing outliers increased false discovery rates (FDR) and bias when compared with raw method and adjustment for outliers (Figure 2B). In both scenarios, adjustment of outliers was generally less biased than other methods, especially when the proportion of invalid instrument was high. Outlier removal and outlier adjustment performance are limited by the efficacy and power of outlier detection methods: we note that when we assume all outliers are detected correctly in our simulation scenarios the performance of outlier removal and outlier adjustment both improve in terms of FDR, power and bias. Outlier adjustment is also dependent on availability of the candidate trait, and power to detect the variant’s association with it.

### Empirical TRYX analysis using four exposure-outcome hypotheses

##### To examine the performance of TRYX analysis empirically, we analysed four separate exposure-outcome hypothesis. For each analysis we a) obtain MR estimates of the exposure-outcome causal relationship and detect outlier instruments; b) identify putative novel influences (candidate traits) on the outcome trait based on their associations with outlier variants (Table 1; Supplementary Table S2); c) adjust the original SNP-outcome estimates for the putative influences operating through the candidate traits (Table 2); and d) compare the changes in heterogeneity in the MR estimates of the adjusted SNP-outcome effects to standard outlier removal methods (Figure 4).

##### Example 1: Systolic blood pressure and coronary heart disease

Random effects IVW estimates indicated that higher SBP associates with higher risk of CHD (Beta: 0.57; 95% CI: 0.39, 0.74). While there was substantial heterogeneity in this estimate (Q=682.7 on 157 SNPs, p=5.74 x 10-67), the estimates from MR Egger, weighted median and weighted mode methods were fairly consistent (Table 2). Seven of the 157 SNPs were detected as outliers. We identified 69 candidate traits that were associated with these outliers (p < 5 x 10-8). We manually removed redundant traits and traits that are similar to the exposure and the outcome (e.g. high blood pressure). Among the candidate traits, 15 were putatively causally associated with the risk of CHD (Figure 3A). After applied LASSO regression, 6 traits remained (Table 1): Anthropometric measures (e.g. height), lipid levels (e.g. cholesterol level), and self-reported ibuprofen use were amongst the candidate traits that associated with CHD, uncovered due to two outliers (rs3184504 near SH2B3 and rs9349279 near PHACTR). Additionally, we found that the experience of headache and the presence of migraine were associated with lower risk of CHD (Beta= -1.11; 95% CI: -2.10, -0.12; and Beta= -4.08; 95% CI: -7.73, -0.43, respectively). However, those traits were excluded from further analysis as we set the traditional threshold p-value of 0.05.

We next adjusted the two outlier SNP-outcome effects for their effects through the detected pleiotropic pathways and obtained an adjusted IVW estimate. The total heterogeneity, based on adjusting only these two of 157 SNP effects, reduced by 17%. The effect estimate remained consistent with the original estimate, as did the IVW estimates when removing all outliers, or just outliers known to associate with the candidate traits that associated with the outcome. However, the width of the confidence interval was substantially larger (including the null) after removing outliers known to associate with candidate traits (beta: 0.59; 95% CI: -0.58, 1.76).

##### Example 2: Years of schooling and body mass index

We used 59 independent genetic instruments (REF) to estimate the influence of years of schooling on body mass index (REF) (Table 2). All MR methods indicated that years of schooling has a causal protective effect on the BMI ( (e.g IVW Beta: -0.27; 95% CI: -0.39, -0.16), except the estimate from MR Egger which had a very imprecise estimate (beta: 0.01; 95% CI: -0.67, 0.70), but the degree of heterogeneity was large (Q = 211.9 on 59 SNPs (p=2.17e-19). Three outliers (rs6882046 near *LINC00461*, rs4800490 near *NPC1*, rs8049439 near *ATXN2L*) were identified as contributors to heterogeneity, and they showed associations (p<5e-8) with 48 candidate traits. Among those candidate traits, two were causally associated with BMI (Figure 3B): alcohol intake frequency (which associated with all three outliers) and usual walking pace.

We next re-estimated the influence of years of schooling on body mass index by accounting for outliers. Adjusting the outliers for candidate trait pathways yielded a slightly reduced the heterogeneity (by 15%) and had a small reduction in the confidence intervals while the point estimate remained consistent (Table 1). By contrast there was a 48% of reduction in heterogeneity when removing outliers. Point estimates remained largely consistent across all outlier removal methods. However, we note that Figure 4B shows that one of the outliers (rs4800490, near gene *NPC1*) on the scatter plot moved away from the fitted line after adjusting for the pleiotropic pathway, indicating that if this outlier is due to a pleiotropic pathway we have estimated its indirect effect imprecisely.

##### Example 3: Urate and coronary heart disease

The influence of circulating urate levels on risk of coronary heart disease is not clear (REF). We re-estimated the associations here using a range of methods. As has been previously reported the estimate from IVW suggested a weak association between urate and the risk of CHD using all variants (Beta: 0.08; 95% CI: -0.00, 0.16), while there was a large intercept in the MR Egger analysis (intercept = 0.02; 95% CI: 0.003, 0.03) with a much attenuated causal effect estimate (Table 2). The median and mode based estimates were also consistent with the MR Egger estimate, indicating weak support for urate having a causal influence on coronary heart disease. Three variants were detected as outliers, which associated with 61 candidate traits (p < 5e-8). Among those outliers, rs653178 (*ATNX2*), and rs642803 (*OVOL1*) associated with 14 traits that had conditionally independent influences on the outcome (Figure 3C), including anthropometric measures (e.g. hip circumference), cholesterol levels, diagnosis of thyroid disease, and smoking status.

Removing the outliers in the IVW analysis led to a more precise (though slightly attenuated) estimate of the influence of higher urate levels on CHD risk (Beta: 0.05; 95% CI: 0.01, 0.10 and Beta: 0.06, 95% CIs: 0.06, 0.12, respectively, Table 2). The adjustment model indicated an attenuated IVW estimate, with confidence intervals spanning the null (Beta: 0.068, 95% CI: -0.008,0.144) whilst the degree of heterogeneity was halved by accounting for the pleiotropic pathways through two outlier SNPs. The adjusted scatter plot showed that outliers moved towards the fitted line after controlling for the SNP effect on the candidate traits (Figure 4C). The results in this analysis suggest that it is unlikely that urate has a strong causal influence on CHD. Here, outlier removal appears to strengthen evidence for the wrong answer.

##### Example 4: Sleep duration and schizophrenia

Previous studies have shown … We observed weak evidence for any association between sleep duration and schizophrenia (Beta: 0.17; 95% CIs: -0.56, 0.89), but there was substantial heterogeneity when all SNPs were used (Q= 204.8, p=6.9e-26). Six outlier instruments were detected which associated with 46 candidate traits (p < 5e-8). Among those outliers, the rs7764984 variant near *HIST1H2BJ* and the rs13107325 variant near *SLC39A8* were associated with three traits that putatively influenced the outcome: self-reported coeliac disease, body composition (impedance of leg) and memory function (Figure 4D).

We re-estimated the original association accounting for the detected outliers. The degree of heterogeneity was reduced when removing all 6 outliers (by 73.5%) and also from adjusting for the two SNP effect putative pleiotropic pathways to the outcome (by 46.0%). Both methods of outlier removal and adjustment provide similar estimates in terms of direction, whilst magnitude of estimates were differed. After removing outliers, MR Egger causal estimates were substantially larger (Beta= 0.25; 95% CI: -0.19, 0.70 and Beta= 0.20; 95% CI: -0.40, 0.79, respectively) than those from the method using all variants. IVW causal estimates from adjustment method was virtually identical with the original estimates, with narrower CIs (Beta= 0.17; 95% CI: -0.46, 0.79). While all methods indicate that the sleep duration is unlikely to be a major risk factor for schizophrenia, pursuing outliers in the analysis provided putative indications that coeliac disease and memory function may be risk factors for schizophrenia (Figure 4D).

## Discussion

The problem of instrumental variables being invalid due to horizontal pleiotropy has received much attention in MR analysis. Detecting and excluding such invalid instruments, based on whether they appear to be outliers in the analysis, is now a common strategy that exists in various forms. We have shown here that outlier removal could in some circumstances compound rather than reduce bias, and also misses an opportunity to better understand the traits under study. We developed the MR-TRYX framework, which taps into the MR-Base database of GWAS summary data to identify potential explanations for outlying variants, and to improve estimates by accounting for the pleiotropic pathways that give rise to them. We have also demonstrated the use and interpretation of MR-TRYX in four sets of empirical analyses.

There are three components upon which MR-TRYX depends in order to perform accurately. First, detecting SNPs that are being used as instruments which exhibit horizontal pleiotropy; second, identifying the candidate traits on the path from the variant to the outcome; and third, correctly estimating the effects of the the candidate traits on the outcome. Each of these components are difficult problems in themselves, but they are modular and build upon existing methods and resources, and the TRYX framework will naturally improve as those methods and resources themselves improve. We will now discuss the consequences of underperformance of each of these components on the TRYX analysis.

The classification of an outlier in MR analysis can be based on the likelihood of a SNP being included as an instrument due to being reverse causal (Steiger filtering), the extent to which a single SNP disproportionately influences the overall result (Cook’s distance), but most commonly the extent to which a SNP contributes to heterogeneity (Cochran’s Q statistic, MR-PRESSO, and implicitly in median- and mode-based estimators). The philosophy of the latter two approaches is that proving horizontal pleiotropy is impossible, but that it should lead to outliers. While a useful approximation, this strategy has two main limitations. First, determining whether a SNP is an outlier depends on the use of arbitrary thresholds, and this entails a tradeoff between specificity and sensitivity. Second, if the majority of variants are pleiotropic then it is possible that the outlier SNPs are the only valid instruments. Such a scenario can arise for complex traits that have a few large effects and many small effects such as gene expression or protein levels. For example C-reactive protein (CRP) levels, where the SNP in the *CRP* gene region is likely the only valid instrument in some analyses. MR-TRYX should in principle avoid both of these problems because instead of removing outliers in their entirety, it attempts to eliminate the component of the SNP-outcome effect that is due to horizontal pleiotropy. Hence, we avoid implicitly cherry picking from amongst the SNPs to be used in the analysis, and if we have low sensitivity (i.e. a more relaxed threshold for outlier detection) it doesn’t mean that there will be an unnecessary loss of power in the overall analysis. One solution is to dispense with the outlier detection step entirely, and search for potential horizontal pleiotropy pathways occurring through all instruments regardless of their contributions to heterogeneity.

Upon identification of potentially pleiotropic SNPs, MR-TRYX can only account for these if the pathways through which pleiotropy is acting can be identified. Detecting the pathways depends on the density and coverage of the human phenome available for the analysis. We use the MR-Base database of GWAS summary results, which comprises several hundred independent traits. While a valuable resource, it is certainly not covering the whole human phenome, and therefore even if a pleiotropic variant is detected correctly, we may not be able to adjust for them. In the empirical analyses that we detected, often fewer than half of the outliers associated traits which we inferred to be associating with the outcome. Broadening phenotype coverage is an on-going pursuit that will continually improve MR-TRYX analysis.

Finally, it is necessary for the effects through the identified pleiotropic pathways to be accurately estimated. This is a recursive problem – MR-TRYX adjusts the SNP-outcome effects based on the pleiotropic effect through the outlier SNP, but it does this by introducing more SNPs into the analysis that instrument the candidate traits. These new SNPs may themselves exhibit pleiotropic effects which could lead to bias in the estimates of the candidate traits on the outcome, requiring a second round of TRYX-style candidate trait searches; and so on. Further developments in MR-TRYX could involve recursively analysing alternative pathways in this manner. In the current format, to some extent the TRYX framework does limit bias in the estimates of effects through the candidate pathways through the automated use of multivariable MR. Here we account for horizontal pleiotropy by including traits that are suspected to be on the causal pathway from SNP to outcome. We have adapted the method to include a shrinkage step in order to reduce redundancy among traits being jointly estimated, which will increase the SNP-exposure effect heterogeneity which is an important contributor to power of the analysis.

The fact that the same genetic variant affects multiple outcomes through different biological pathways, known as horizontal pleiotropy (25), makes it difficult to use genetic variants as instrument variables in MR. A couple of MR approaches has been developed to deal with violation of the critical assumption of MR (12, 26, 27): the genetic variant should not directly affect the outcome (No pleiotropy). To avoid bias due to pleiotropy, the genetic variants that are likely to be violated the assumption have been considered to be excluded from the analysis (28). However, removal of those variants (outliers) can cause severe bias in MR since the biological working of genes has not been fully understood. In this regard, adjustment for pleiotropic effect can increase the precision of the estimates and can prevent loss of information that explains the variability of the exposure. Therefore, we focused on the way to exploit pleiotropic variants and to adjust for pleiotropic effect rather than excluding them from the analysis.

Since pleiotropic SNPs have independent effects via different pathways (29), we can exploit them to identify putative risk factors for the outcome. In this scenario, a pleiotropic variant need not necessarily be treated as nuisance. Using MR-TRYX package, we demonstrated that pleiotropic variants can be used to explore alternative pathways in disease pathogenesis. For example, in our applied examples, the 2 sample MR results showed that elevated levels of SBP lead to increased risk of CHD, but also suggested that headache related traits (e.g. experience of pain due to headache and self-reported status of ibuprofen intake) may influence the original association. This result confirmed the previous study that showed shared genetic risk between headache (migraine) and CHD (30), suggesting potential role of migraine in vascular mechanisms. Also, we found that MR-TRYX detected well established risk factors to CHD (e.g. blood levels of cholesterol) as well as alternative pathway. This result demonstrated the validity of MR-TRYX to detect the majority of traits that possibly influence the disease outcome. Furthermore, our results found the putative risk factors for schizophrenia such as coeliac disease, and body composition which have not been reported before using MR. Our example illustrated how outliers can be used to identify alternative pathway, opening the door for hypothesis free MR approaches to establish disease network. MR-TRYX may help us understand and represent all known traits and disease associations in a single framework.

In our simulation, adjustment method improved FDR, power and bias whilst removing outliers increased FDR and bias. Even if we assume all outliers are detected correctly, as our simulation presents, there is a practical difficulty of setting a threshold for excluding a genetic variant as pleiotropic. A previous study supported this, reporting that bias due to weak IVs cannot be avoided by selection of instruments (31). This study also reported that the measured strength of instruments in data that relies on providing a threshold is flawed and may generate more bias. Additionally, our simulation showed that adjusting for pleiotropic pathway clearly outperforms removal of the outlying variants if pleiotropic variants are more than 10. In the applied analyses, we found that adjusting for effect of candidate traits on the original association yielded unbiased estimate. Also, this adjustment method reduced the heterogeneity due to pleiotropic outliers. The example of urate and CHD showed that the noise due to pleiotropy was substantially decreased after correcting for the effect of candidate traits. However, we acknowledge that MR-TRYX does not always correct for pleiotropy in the expected direction. In the example of education level and BMI, adjustment for the pleiotropic pathway failed to reduce the degree of heterogeneity. Since MR-TRYX relies on the publicly available database, it may fail to correct for pleiotropy if the role of variants was not fully understood by previous GWAS. Yet, as we illustrated in other examples, MR-TRYX allow for an informative analysis that could routinely be applied in the MR analyses.

There are several limitations to this approach. First, investigation of pleiotropy and candidate traits are based on statistical evidence. For example, experience of headache and migraine were excluded from the multivariable analysis as we only included the traits with P-value <0.05, even though confidence interval did not include null value. Whilst MR-TRYX package allows to choose the threshold desired by the users, biological evidence of possible pleiotropy should be considered. Second, as we described above, although MR-TRYX allows us to deal with genetic variants with measured pleiotropic associations, it is unable to address unmeasured or unknown pleiotropy. Third, there were cases where adjustment could not completely correct horizontal pleiotropy. Also, it should be considered that MR-TRYX does not correct other sources that influence heterogeneity in the causal effect rather than pleiotropy. Fourth, in the case of the binary outcome, we used odds ratio, a non-collapsibility measure of association (32). Therefore, there may be parametric restriction on the conditional causal odds ratio in our adjustment model where the exposure effect is in linear in the exposure on log odd ratio scale (33). However, the two-stage estimator with a logistic second-stage model still yield a valid test of the causal null hypothesis (33).

Strength of this study include the use of data from large GWA studies of exposure and outcome, and the MR design. This design allows us to avoid bias from confounding and reverse causation. Potential bias from population stratification was reduced in our study since we restricted the analyses to individuals of European ancestry. MR-TRYX can be used within the framework of the IVW meta-analysis and therefore retains statistical power of the approach. More speculatively, MR-TRYX enables us to obtain hypothesis-free causal inference. The potential of identifying putative risk factors using outliers can be growing along with the increasing availability of genetic variants from large-scale GWAS.

In conclusion, we have shown a new method to deal with the bias from horizontal pleiotropy, and to identify putative risk factors for outcomes, exploiting association between outliers and other traits. MR-TRYX may be beneficial where genetic variants are associated with several related risk factors, and where it is required to searching for the cause of a disease that has not been fully identified.

## References

1. Holmes, M.V., Ala-Korpela, M. and Smith, G.D. (2017) Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nat Rev Cardiol*, **14**, 577-590.

2. Smith, G.D. and Ebrahim, S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, **32**, 1-22.

3. Davey Smith, G. and Hemani, G. (2014) Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet*, **23**, R89-98.

4. Johnson, T. and Uk, S. (2012) Efficient calculation for multi-SNP genetic risk scores. in press.

5. Pierce, B.L. and Burgess, S. (2013) Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am J Epidemiol*, **178**, 1177-1184.

6. Bowden, J., Smith, G.D. and Burgess, S. (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, **44**, 512-525.

7. Bowden, J., Smith, G.D., Haycock, P.C. and Burgess, S. (2016) Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol*, **40**, 304-314.

8. Hartwig, F.P., Davey Smith, G. and Bowden, J. (2017) Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol*, **46**, 1985-1998.

9. Hemani, G., Bowden, J. and Davey Smith, G. (2018) Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum Mol Genet*, in press.

10. Corbin, L.J., Richmond, R.C., Wade, K.H., Burgess, S., Bowden, J., Smith, G.D. and Timpson, N.J. (2016) BMI as a Modifiable Risk Factor for Type 2 Diabetes: Refining and Understanding Causal Estimates Using Mendelian Randomization. *Diabetes*, **65**, 3002-3007.

11. Hemani, G., Tilling, K. and Smith, G.D. (2017) Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *Plos Genet*, **13**.

12. Bowden, J., Del Greco, M.F., Minelli, C., Davey Smith, G., Sheehan, N. and Thompson, J. (2017) A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat Med*, **36**, 1783-1802.

13. Bowden, J., Spiller, W., Del Greco, M.F., Sheehan, N., Thompson, J., Minelli, C. and Davey Smith, G. (2018) Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. *Int J Epidemiol*, in press.

14. Verbanck, M., Chen, C.Y., Neale, B. and Do, R. (2018) Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet*, **50**, 693-698.

15. Bakker, M. and Wicherts, J.M. (2014) Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples t tests: the power of alternatives and recommendations. *Psychol Methods*, **19**, 409-427.

16. Verbanck, M., Chen, C.-Y., Neale, B. and Do, R. (2017) Widespread pleiotropy confounds causal relationships between complex traits and diseases inferred from Mendelian randomization. *BioRxiv*, in press., 157552.

17. Hemani, G., Bowden, J., Haycock, P.C., Zheng, J., Davis, O., Flach, P., Gaunt, T.R. and Smith, G.D. (2017) Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome. *bioRxiv*, in press., 173682.

18. Thomson, J. (1909) The methods and scope of genetics. An inaugural lecture delivered October 23rd, 1908. *The Eugenics Review*, **1**, 59.

19. Bowden, J., Fabiola Del Greco, M., Minelli, C., Lawlor, D., Zhao, Q., Sheehan, N., Thompson, J. and Smith, G.D. (2018) Improving the accuracy of two-sample summary data Mendelian randomization: moving beyond the NOME assumption. *bioRxiv*, in press., 159442.

20. Sanderson, E., Davey Smith, G., Windmeijer, F. and Bowden, J. (2018) An examination of multivariable Mendelian randomization in the single sample and two-sample summary data settings. *bioRxiv*, in press.

21. Bennett, D.A. and Holmes, M.V. (2017) Mendelian randomisation in cardiovascular research: an introduction for clinicians. *Heart*, **103**, 1400-1407.

22. White, J., Sofat, R., Hemani, G., Shah, T., Engmann, J., Dale, C., Shah, S., Kruger, F.A., Giambartolomei, C., Swerdlow, D.I. *et al.* (2016) Plasma urate concentration and risk of coronary heart disease: a Mendelian randomisation analysis. *Lancet Diabetes Endocrinol*, **4**, 327-336.

23. Tyrrell, J., Jones, S.E., Beaumont, R., Astley, C.M., Lovell, R., Yaghootkar, H., Tuke, M., Ruth, K.S., Freathy, R.M., Hirschhorn, J.N. *et al.* (2016) Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank. *BMJ*, **352**, i582.

24. Kaskie, R.E., Graziano, B. and Ferrarelli, F. (2017) Schizophrenia and sleep disorders: links, risks, and management challenges. *Nat Sci Sleep*, **9**, 227-239.

25. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. and Smoller, J.W. (2013) Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet*, **14**, 483-495.

26. Schmidt, A.F. and Dudbridge, F. (2017) Mendelian randomization with Egger pleiotropy correction and weakly informative Bayesian priors. *Int J Epidemiol*, in press.

27. van Kippersluis, H. and Rietveld, C.A. (2017) Pleiotropy-robust Mendelian randomization. *Int J Epidemiol*, in press.

28. Burgess, S., Bowden, J., Fall, T., Ingelsson, E. and Thompson, S.G. (2017) Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology*, **28**, 30-42.

29. Pickrell, J.K., Berisa, T., Liu, J.Z., Segurel, L., Tung, J.Y. and Hinds, D.A. (2016) Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet*, **48**, 709-717.

30. Winsvold, B.S., Bettella, F., Witoelar, A., Anttila, V., Gormley, P., Kurth, T., Terwindt, G.M., Freilinger, T.M., Frei, O., Shadrin, A. *et al.* (2017) Shared genetic risk between migraine and coronary artery disease: A genome-wide analysis of common variants. *Plos One*, **12**.

31. Burgess, S., Thompson, S.G. and Collaboration, C.C.G. (2011) Avoiding bias from weak instruments in Mendelian randomization studies. *Int J Epidemiol*, **40**, 755-764.

32. Hernan, M.A., Clayton, D. and Keiding, N. (2011) The Simpson's paradox unraveled. *Int J Epidemiol*, **40**, 780-785.

33. Vansteelandt, S., Bowden, J., Babanezhad, M. and Goetghebeur, E. (2011) On Instrumental Variables Estimation of Causal Odds Ratios. *Stat Sci*, **26**, 403-422.