Inferring the existence and direction of causal associations in the face of measurement error

29 December 2016

Gibran Hemani*, Kate Tilling and George Davey Smith

MRC Integrative Epidemiology Unit (IEU) at the University of Bristol, School of Social and Community Medicine, Bristol, UK

* Correspondence to: g.hemani@bristol.ac.uk

Abstract

With our ability to characterize the human phenome ever improving it is becoming increasingly common to use statistical tests to infer the causal relationships between correlated variables. A simple and commonly used method for inferring if an exposure is on the causal pathway to an outcome is through using genetic instruments in mediation analysis, where the effect of an instrument on the outcome is tested before and after adjusting for the exposure. We show that in the face of measurement error this can lead to erroneous inference of both existence of causality and the causal direction, and that increasing sample size, rather than reducing bias, can often have the effect of increasing certainty in the wrong answer. Commonly used tools that are predicated on this methodology such as the causal inference test (CIT) are shown to be susceptible. We demonstrate that Mendelian randomization (MR) is a method for causal inference that is robust to measurement error, and introduce a simple extension to estimate the direction of causality between correlated variables even when the biology of the instrument is unknown. This extension is less susceptible to measurement error than mediation-based analyses. Using these methods we demonstrate empirically that the inferred direction of causality between correleted DNA methylation and gene expression levels is highly susceptible to differences in measurement error between the two assaying platforms, demanding caution when making such claims about causality.

The last decade has seem a rapid acceleration in technological advances that enable us to assess the human phenome in ever increasing detail, furnishing biologists with the tantalising opportunity to map the molecular basis of complex traits. This does, however, require overcoming the statistical challenge of correctly defining the causal architectures that underlie these high dimensional data. For example, identifying metabolic perturbations that lead to disease[1], or understanding the influence of epigenomic influences on gene expression[2] all depend on understanding the causal nature of any putative observational associations.

Introduction

Observational study data are now widely available but notoriously susceptible to pitfalls in making causal inference, with basic regression-based techniques unable to distinguish a true causal association from reverse causation or confounding[3–5]. In response to this, genetic instrumentation has emerged as a valuable technique for improving the reliability of causal inference in observational data, and with the coincident rise in genome-wide association studies it is now a prominent tool that is applied in several different guises, and has been applied many forms of high throughput ’omic variables. However, an often neglected pitfall to these approaches is the influence of non-differential measurement error on the reliability of causal inference.

Non-differential measurement error (or observational error, henceforth referred to simply as measurement error) is the difference between the measured value of a quantity and its true value. Such variability can arise through a whole plethora of mechanisms, which are often unique to the study design and difficult to avoid[6,7]. Array technology is now commonly used to obtain high throughput phenotyping at low cost, but they come with the problem of having imperfect resolution, for instance methylation levels as measured by the Illumina450k chip are prone to have some amount of noise around the true value[8,9]. Relatedly, if the measurement of biological interest is the methylation level in a T cell, then measurement error of this value can be introduced by using methylation levels from whole blood samples because the measured value will be an assay of many cell types[10].

Measurement error will of course arise in more low-tech data too, for example when measuring BMI one is typically interested in using this as a proxy for adiposity, but it is clear that the correlation between BMI and underlying adiposity is not perfect[11]. A similar problem of biological misspecification is unavoidable in disease diagnosis, and measuring behaviour such as smoking or diet is notoriously difficult to do accurately. Measurement error can also be introduced after the data has been collected, for example the transformation of non-normal data for the purpose of statistical analysis will lead to a new variable that will typically incur both changes in scale and imprecision (noise) compared to the original variable. The sources of measurement error are not limited to this list[7], and its impact has been explored in the epidemiological literature extensively[12–14].

Given the near-ubiquitous presence of measurement error in phenomic data it is vital to understand its impact on the tools we use for causal inference. An established study design that can provide information about causality is randomisation. Given the hypothesis that trait A (henceforth referred to as the exposure) is causally related to trait B (henceforth referred to as the outcome), randomisation can be employed to assess the causal nature of the association by randomly splitting the sample into two groups, subjecting one group to the exposure and leaving the other as a control. The association between the exposure and the outcome in this setting provides a robust estimate of the causal relationship. This provides the theoretical basis behind randomised control trials, but in practice randomisation is often impossible to implement in an experimental context due to cost, scale or inability to manipulate the exposure. The principle, however, can be employed in extant observational data through the use of genetic variants associated with the exposure (instruments), where the inheritance of an allele serves as a random lifetime allocation of differential exposure levels[3]. Two statistical approaches to exploiting the properties of genetic instruments are widely used: mediation-based approaches and Mendelian randomisation (MR).

Mediation-based approaches employ genetic instruments (typically single nucleotide polymorphisms, SNPs) to orient the causal direction between the exposure and the outcome. If a SNP is associated with an exposure, and the exposure is associated with some outcome, then it logically follows that the indirect influence of the SNP on the outcome will disappear when conditioning on the exposure. Here, the exposure 'mediates' the association between the SNP and the outcome, providing information about the causal influence of the exposure on the outcome. This forms the basis of a number of methods such as genetical genomics[15], the regression-based causal inference test (CIT)[16,17], a structural equation modelling (SEM) implementation in the NEO software[18], and various other methods including Bayesian approaches[19]. They have been employed by a number of recent publications that make causal inferences in large scale 'omics datasets[19–24].

MR can operate on the same data - phenotypic measures of the exposure and the outcome variables and a genetic instrument for the exposure - but the genetic instrument is employed in a subtly different manner. Here the SNP, which is assumed to have a direct association with the exposure, is used as a surrogate for the exposure. The causal effect of the exposure and the outcome can be estimated by scaling the association between the SNP and the outcome by the association between the SNP and the exposure.

By utilising genetic instruments in different ways, mediation-based analysis and MR models have properties that confer some advantages and some disadvantages for reliable causal inference. The main advantage of mediation over MR is that the true underlying biology of the genetic instrument doesn't necessarily need to be known as long as it associates with the putative exposure. In the CIT framework (described fully in the Methods) for example, the test statistic is different if you test for the exposure causing the outcome or the outcome causing the exposure, allowing the researcher to infer the direction of causality between two variables by performing the test in both directions and choosing the model with the strongest evidence.

Such is not the case for MR, where in order to infer the direction of causality the instrument must be known to influence the exposure directly, associating with the outcome only through the exposure. This requirement can be problematic because if there exists a putative association between two variables, with the instrumental SNP being robustly associated with each, it can be difficult to determine which of the two variables is subject to the direct effect of the SNP (i.e., for which of the two variables is the SNP a valid instrument? Figure 1).

By definition, we expect that if the association is causal then a SNP for the exposure will be associated with the outcome, so if the researcher erroneously uses the SNP as an instrument for the outcome then they are likely to see an apparently robust causal association of outcome on exposure. Such a situation can arise in many scenarios. Genome-wide association studies (GWASs) that identify genetic associations for complex traits are, by design, hypothesis free and agnostic of genomic function, and it often takes years of follow up studies to understand the biological nature of a putative GWAS hit[25]. For example, the same FTO variant that is associated with body mass index (BMI) is also associated with C-reactive protein (CRP). It is known that this is because BMI causes CRP[26], but supposing we only had *a priori* knowledge that FTO associated with CRP, and we used it to instrument CRP in an MR analysis of CRP against BMI, we would mistakenly infer that CRP causes BMI. Another situation is where the causal directions between 'omic measures need to be determined, for example if a DNA methylation probe is associated with expression of an adjacent gene, then is a cis-acting SNP an instrument for the DNA methylation level, or the gene expression level[27] (Figure 1)?

MR does however have some important advantages over mediation based approaches. First, MR is used as a tool to not just deduce the existance of causality, but to also estimate the causal effect size. Second, using mediation requires that the exposure, outcome and instrument variables are all measured in the same data, whereas recent extensions to MR circumvent this requirement, allowing causal inference to be drawn when exposure variables and outcome variables are measured in different samples[28]. This has the crucial advantage of improving statistical power by allowing analysis in much larger sample sizes, and dramatically expands the breadth of possible phenotypic relationships that can be evaluated. Third, MR is substantially more robust to there being measurement error in the exposure and outcome variables. Indeed instrumental variable (IV) analysis was in part initially introduced as a correction for measurement error[29], whereas it has been noted that both classic mediation-based analyses[13,14,30,31] and mediation-based methods that use instrumental variables[32,33] are prone to be unreliable in its presence. We argue that this is a serious limitation.

Here we show using theory and simulations how non-differential measurement error and measurement imprecision can lead to unreliable causal inference in the mediation-based CIT method. We then present an extension to MR that allows researchers to ascertain the causal direction of an association even when the biology of the instruments are not fully understood, and to evaluate the sensitivity of the result to measurement error. This improves the utility of MR in cases where mediation based methods might have otherwise been used preferentially. Finally, we apply this new method to infer the direction of causation between DNA methylation levels and gene expression levels.

Methods

The analysis proceeds by simulating variables that have known causal effects and valid instruments, and then applying measurement error in different configurations. We then assess the performance of the CIT test and MR analysis for a) inference of a causal relationship and b) inferring the correct direction of causality. Included in the MR analysis is a simple extension that is used to make inference of the causal direction when it is unknown which of the correlated variables the instrument is directly acting upon. All analysis was performed using the R programming language[34] and code is made available at github url to go here.

CIT test

The CIT method[16] is implemented in the R package *R/cit*[17]. The methodology of the CIT is as follows. Assume an exposure x is instrumented by a SNP g, and the exposure x is causally related to an outcome y. Thus,

$$\begin{aligned}
g &\sim Binom(2, 0.5) \\
x &= \alpha_g + \beta_g g + \epsilon_g \\
y &= \alpha_x + \beta_x x + \epsilon_x
\end{aligned}$$

The following tests are then performed:

1. $H_0 : cov(g, x) = 0; H_1 : cov(g, x) \neq 0$; *the SNP associates with the exposure*
2. $H_0 : cov(g, y) = 0; H_1 : cov(g, y) \neq 0$; *the SNP associates with the outcome*
3. $H_0 : cov(x, y) = 0; H_1 : cov(x, y) \neq 0$; *the exposure associates with the outcome*
4. $H_0 : cov(g, y - \hat{y}) \neq 0; cov(g, y - \hat{y}) = 0$; *the SNP is independent of the outcome when the outcome is adjusted for the exposure*

where $y - \hat{y} = y - \hat{\alpha}_g + \hat{\beta}_g x$ is the residual of y after adjusting for the x, where x is assumed to mediate the association between the SNP and the outcome. If all four tests reject the null hypothesis then it is inferred that x is causally related to y. The CIT measures the strength of causality by generating an omnibus p-value, p_{CIT}, which is simply the largest (least extreme) p-value of the four tests, the intuition being that causal inference is only as strong as the weakest link in the chain of tests.

In these analyses the *cit.cp* function was used to obtain an omnibus p-value. To infer the direction of causality using the CIT method, an omnibus p-value generated by CIT, p_{CIT}, was estimated for the correct model of x causing y, and for incorrect model of y causing x. For some significance threshold α, the existence of causality and its direction was inferred based on the following scenarios:

- If $p_{CIT,correct} < \alpha$ and $p_{CIT,incorrect} > \alpha$ then the correct model is accepted
- If $p_{CIT,correct} > \alpha$ and $p_{CIT,incorrect} < \alpha$ then the incorrect model is accepted
- If $p_{CIT,correct} > \alpha$ and $p_{CIT,incorrect} > \alpha$ then an alternative causal model is accepted
- If $p_{CIT,correct} < \alpha$ and $p_{CIT,incorrect} < \alpha$ then no inference is made

For the purposes of compiling simulation results we use an arbitrary $\alpha = 0.05$ value, though we should stress that for real analyses it is not good practice to use p-values for making causal inference, nor is it reliable to depend on arbitrary significance thresholds.

MR causal test

Two stage least squares (2SLS) is a commonly used technique for performing MR when the exposure, outcome and instrument data are all available in the same sample. Assuming the following model

$$y = \alpha_{MR} + \beta_{MR}\hat{x} + \epsilon_{MR}$$

where $\hat{x} = \hat{\alpha}_g + \hat{\beta}_g g$, a causal relationship is inferred if the null hypothesis that $\beta_{MR} = 0$ is rejected. A p-value for this test, p_{MR}, was obtained using the R package $R/systemfit$[35].

The direction of the causal association was inferred as follows. Assuming no horizontal pleiotropy (i.e. g only has an effect on one of the variables through the other variable) it is desirable to know which of the variables, x or y, it has a direct influence on. This can be achieved by assessing which of the two variables has the biggest absolute correlation with g (Appendix 2). This test can be formalised by testing for a difference in the correlations ρ_{gx} and ρ_{gy} using Steiger's Z-test for correlated correlations within a population[36]. It is calculated as

$$Z = (Z_{gx} - Z_{gy})\frac{\sqrt{N-3}}{\sqrt{2(1-\rho_{xy})h}}$$

where Fisher's z-transformation is used to obtain $Z_{g*} = \frac{1}{2}\ln\left(\frac{1+\rho_{g*}}{1-\rho_{g*}}\right)$,

$$h = \frac{1-(frm^2)}{1-rm^2}$$

where

$$f = \frac{1 - \rho_{xy}}{2(1 - rm^2)}$$

and

$$rm^2 = \frac{1}{2}(\rho_{gx}^2 + \rho_{gy}^2).$$

The Z value is interpreted such that

$$Z \begin{cases} > 0, & x \rightarrow y \\ < 0, & y \rightarrow x \\ = 0, & x \perp\!\!\!\perp y \end{cases}$$

and a p-value is generated from the Z value to indicate the probability of obtaining the observed difference in correlations ρ_{gx} and ρ_{gy} under the null hypothesis that both correlations are identical. The existence of causality and its direction is inferred based on the following scenarios:

- If $p_{Steiger} < \alpha$ and $p_{MR} < \alpha$ and $Z > 0$ then a causal association for the correct model is accepted
- If $p_{Steiger} < \alpha$ and $p_{MR} < \alpha$ and $Z < 0$ then a causal association for the incorrect model is accepted
- Otherwise no inference is made

For the purposes of compiling simulation results, we use an arbitrary $\alpha = 0.05$ value.

Note that the same correlation test approach can be applied to a two-sample MR[28] setting. Two-sample MR refers to the case where the SNP-exposure association and SNP-outcome association are calculated in different samples (or from publicly available summary statistics). Here the Steiger test of two independent correlations can be applied where.

$$Z = \frac{Z_{gx} - Z_{gy}}{\sqrt{1/(N_1 - 3) + 1/(N_2 - 3)}}$$

An advantage of using the Steiger test in the two sample context is that it can compare correlations in independent samples where sample sizes are different. Steiger test statistics were calculated using the *r.test* function in the R package *R/psych*[37].

Simulations

Simulations were conducting by creating variables of sample size n for the exposure x, the measured values of the exposure x_o, the outcome y, the measured values of the outcome y_o and the instrument g. In all models x causes y and g is an instrument for x. Each variable was simulated such that:

$$\begin{aligned} g &\sim Binom(2, 0.5) \\ x &= \alpha_g + \beta_g g + \epsilon_g \\ x_o &= \alpha_{mx} + \beta_{mx} x + \epsilon_{mx} \\ y &= \alpha_x + \beta_x x + \epsilon_x \\ y_o &= \alpha_{my} + \beta_{my} y + \epsilon_{my} \end{aligned}$$

where $\epsilon_{m*} \sim N(0, \sigma^2_{m*})$, α_{mx} and β_{mx} are parameters that represent non-differential measurement error into the exposure variable x, and α_{my} and β_{my} are parameters for non-differential measurement error in the outcome y.

All α values were set to 0, and β values set to 1. Normally distribted values of ϵ_* were generated such that

$$
\begin{aligned}
cor(g,x)^2 &= 0.1 \\
cor(x,y)^2 &= \{0.2, 0.4, 0.6, 0.8\} \\
\sigma^2_{mx} &= \{0, 0.2, 0.4, 0.6, 0.8, 1\} \\
\sigma^2_{my} &= \{0, 0.2, 0.4, 0.6, 0.8, 1\} \\
n &= \{100, 1000, 10000\}
\end{aligned}
$$

giving a total of 432 combinations of parameters. Simulations using each of these sets of variables were performed 100 times, and the CIT and MR methods were applied to each in order to evaluate the causal association of the simulated variables. Similar patterns of results were obtained for different values of $cor(g, x)$.

Two sample MR

Two sample MR[28] was performed using the summary statistics presented in[27] for each of 4122 associations between DNA methylation and gene expression levels. Because only summary statistics were available (effect, standard error, effect allele, sample size, p-values) for the instrumental SNP on the methylation and gene expression levels, the Steiger test of two independent correlations was used to infer the direction of causality for each of the associations. The Wald ratio test was then used to estimate the causal effect size for the estimated direction for each association.

Results

The influence of measurement error on mediation

Measurement error of an exposure can be modeled as some transformation of the true value that leads to the observed value, $x_o = f(x)$. For example, we can define $f(x) = \alpha_{mx} + \beta_{mx}x + \epsilon_{mx}$, where α_{mx} and β_{mx} influence the error in the measurement of x, and ϵ_{mx} represents the imprecision in the measurement of x. Here the true value of the exposure is partially explained by the genetic instrument, g, such that

$$x = \alpha_g + \beta_g g + \epsilon_g$$

where β_g is the effect of the SNP on the exposure, and ϵ_g is the normally distributed residual value of x; and the outcome is partially explained by the exposure

$$y = \alpha_x + \beta_x x + \epsilon_x$$

where β_x is the true effect of the exposure on the outcome. In the causal inference test (CIT), an omnibus p-value is generated from four hypothesis tests: 1) g is associated with x; 2) g is associated with y; 3) x is associated with y; and 4) g is independent of $y|x$. The 4th condition employs mediation for causal inference, and can be expressed as $cov(g, y - \hat{y}) = 0$, where $\hat{y} = \hat{\alpha}_x + \hat{\beta}_x x_o$. When measurement error is introduced it can be shown using basic covariance properties (Appendix 1) that

$$
\begin{aligned}
cov(g, y - \hat{y}) &= cov(g, y_o) - cov(g, \hat{y}_o) \\
&= \beta_{my}\beta_g\beta_x var(g) - D\beta_{my}\beta_g\beta_x var(g)
\end{aligned}
$$

where

$$D = \frac{\beta^2_{mx} var(x)}{\beta^2_{mx} var(x) + var(\epsilon_{mx})}$$

Thus an observational study will find $cov(g, y_o - \hat{y_o}) = 0$ when the true model is causal only when $D = 1$. Therefore, if there is any measurement error that incurs imprecision (i.e. $var(\epsilon_{mx}) \neq 0$) then there will remain an association between g and $y_o|x$, which is in violation of the the 4th condition of the CIT. Note that linear transformation of x or y without any incurred imprecision is insufficient to lead to a violation of the test statistic assumptions.

We performed simulations to verify that this problem does arise using the CIT method. Figure 2 shows that when there is no measurement error in the exposure or outcome variables ($\rho_{x,x_o} = 1$) the CIT is reliable in identifying the correct causal direction. However, as measurement error increases in the exposure variable, eventually the CIT is more likely to infer a robust causal association in the wrong direction. Also of concern here is that increasing sample size does not solve the issue, indeed it only strengthens the apparent evidence for the incorrect inference.

Using MR to infer the existence of causality

When selecting an instrument g that has a direct influence on x where x is causally related to y, one can reason that the influence of g on y is the proportion of the effect of x on y that is explained by the effect of g on x. Hence the effect estimate β_{MR} of x on y is estimated as

$$\begin{aligned}\beta_{MR} = \frac{\beta_{y_o}}{\beta_g} &= \frac{cov(y_o, g)}{cov(x_o, g)} \\ &= \frac{\beta_{my}\beta_x cov(x, g)}{\beta_{mx} cov(x, g)} \\ &= \frac{\beta_{my}}{\beta_{mx}}\beta_x\end{aligned}$$

Thus we obtain an estimate of the effect of x_o on y that is scaled by the measurement error in x and y, but is unrelated to measurement imprecision. Measurement error and imprecision are likely to reduce the power of MR inferring a causal association (increased false negative rate), but importantly the evidence for causal association, based on rejecting the null hypothesis that $\beta_{MR} = 0$, is not inflated when the null hypothesis is true.

We performed simulations to compare the performance of MR against CIT in detecting a causal association between simulated variables under different levels of imprecision simulated in the exposure. Figure 4 shows the true positive rates between the CIT and MR for detecting a causal association. We observe that the CIT has lower power in all cases, with performance declining as measurement imprecision increases in the exposure. The performance of MR in detecting an association is unrelated to measurement error in. Measurement error in the outcome does not induce a substantive difference in performance between CIT and MR.

Using MR to infer the direction of causality

If we do not know whether the SNP g has a direct influence on x or y then further efforts are required. If x causes y and g is a valid instrument for x then we expect that $\rho_{g,x}$ to be greater than $\rho_{g,y}$ because the association of g with y is the proportion of the true causal association between x and y scaled by the association between g and x. We can use the Steiger test of correlated correlations to formally test the difference in magnitude of $\rho_{g,x}$ and $\rho_{g,y}$, where the variable with the largest correlation on g is deemed to be the variable which is imposing the causal effect.

This is a simple method to orient the direction of causality in an MR analysis when the underlying biology of the SNP is not fully understood. But it can be shown that in the presence of measurement imprecision, $d = \rho_{x,x_o} - \rho_{x,y}\rho_{y,y_o}$ (Appendix 2), where d is a metric of the degree to which the Steiger test is liable to provide the wrong estimate (*i.e.* if $d > 0$ then the Steiger test is likely to be correct about the causal direction). Thus it is important to note that under certain conditions when measurement imprecision is present the Steiger test could make erroneous inference about the causal direction because the value of d is not restricted to being greater than 0. However Figure 7 shows that when there is no measurement error in y, the Steiger test is unlikely to infer the wrong direction of causality even if there is measurement error in x. It also shows that in most cases where y is measured with error, especially when the causal effect between x and y is not very large, the condition of $d > 0$ is satisfied.

We performed simulations to explore the performance of this approach in comparison to CIT in terms of the rate at which a strong causal relationship is obtained for the correct direction of causality, and the rate at which a strong causal relationship is obtained where the reported direction of causality is incorrect. Figure 5 shows that when $d < 0$ the MR analysis is indeed liable to infer the wrong direction of causality, and that this erroneous result is more likely to occur with increasing sample size. However, the CIT is in general much more fallable to reporting a robust causal association for the wrong direction of causality. When $d > 0$ is satisfied we observe that in most cases the MR method has greater power to obtain evidence for causality than CIT, and always obtains the correct direction of causality.

MR analysis of gene expression and DNA methylation

We used the Steiger test to infer the direction of causality between DNA methylation levels and gene expression levels. To do this we obtained a list of 458 gene expression - DNA methylation associations as reported in Shakhbazov et al[27]. These were filtered to be located on the same chromosome, have robost correlations after correcting for multiple testing, and to share a SNP that had a robust cis-acting effect on both the DNA methylation probe and the gene expression probe. Using the reported summary data we used the Steiger test of two independent correlations to identify which direction of causality was most likely for each association. We found that the causal direction commonly goes in both directions (Table 1) but methylation levels were more likely to influence expression levels than vice versa (probability of expression causing methylation = 0.0163873 (95% CI 0.3738727 - 0.4874269) amongst NA associations with $p_{Steiger} < 0.05$).

We performed two sample MR[28] for each association in the direction of causality that we estimated to be the correct direction from the Steiger test. We observed that the sign of the MR estimate was always in the same direction as the Pearson correlation coefficient reported by Shakhbazov et al[27], and that there was a moderate correlation between the magnitude of the causal correlation and the observational Pearson correlation (Table 1).

We also observed that for associations where methylation caused gene expression the causal effect was more likely to be negative than for the associations where gene expression caused methylation (OR = 0.5326663 (95% CI 0.3172203 - 0.8851868)), suggesting that reducing gene expression levels at a controlling CpG typically leads to increased gene expression levels, consistent with expectation[2].

Discussion

Researchers are often confronted with the problem of making causal inferences using a statistical framework on observational data, and using genetic instruments as surrogates for randomisation is a useful approach. MR and the CIT employ genetic instruments in different ways, and unfortunately neither offer solutions that work perfectly in all scenarios. In the face of measurement error it is evident that causal inference drawn through mediation based methods such as the CIT analysis typically fare worse than MR and will be difficult to interpret, and this is a serious problem considering that it is often impossible to estimate the extent of measurement error present in observational data.

In the epidemiological literature issues of measurement error in mediation is relatively well explored[12]. Our analysis extends this to related methods such as CIT that are widely used in predominantly 'omic data. These methods are indeed liable to the same issues as standard mediation based analysis, and specifically we show that as measurement error in the exposure variable increases, CIT is likely have reduced statistical power, and liable to infer the wrong direction of causality. We also demonstrate that, though seemingly unintuitive, increasing sample size does not resolve the issue, rather it leads to more extreme p-values for the model that predicts the wrong direction of causality.

Under many circumstances a practical solution to this problem is to use Mendelian randomisation instead of methods such as the CIT or similar that are based on mediation. Inferring the existence of causality using Mendelian randomisation is robust in the face of measurement error and, if the researcher has knowledge about the biology of the instrument being used in the analysis, can offer a direct solution to the issues that the CIT faces. This assumption is often reasonable, for example SNPs are commonly used as instruments when they are found in genes with known biological relevance for the trait of interest. But on many occasions this is not the case, and it may be tempting to resort to using methods based on mediation in order to be able to both ascertain if there is a causal association and to infer the direction of causality. Here we have described a simple extension to MR which can be used as an alternative to mediation based methods. We show that this method is still liable to measurement error, but because it has different properties to the CIT it offers two main advantages. First, it uses a formal statistical framework to test for the robostness of the assumed direction of causality. Second, after testing in a comprehensive range of scenarios the MR based approach is less likely to infer the wrong direction of causality compared to CIT, while substantially improving power over CIT in the cases where $d > 0$.

Mediation based network approaches are very well established[33] and have a number of extensions that make them valuable tools, including for example network construction. But because they are predicated on the basic underlying principles of mediation they are liable to suffer from the same issues of measurement error. Recent advances in MR methodology, for example applying MR to genetical genomics[38], multivariate MR[39], network MR[40] and mediation through MR[41] may offer more robust alternatives for these more complicated problems.

In our simulations we focused on the simple case of a single instrument in a single sample setting, and assumed that pleiotropy (the influence of the instrument on the outcome through a mechanism other than the exposure) was not present. Recent method developments in MR[42] have focused on accounting for the issues that horizontal pleiotropy can introduce, but how they perform in the presence of measurement error remains to be explored. An important advantage that MR confers over most mediation based analysis is that it can be performed in two samples, which can considerably improve power and expand the scope of analysis. However, whether there is a substantive difference in two sample MR versus one sample MR in how measurement error has an effect is not yet fully understood. We have also assumed no measurement error in the genetic instrument, which is not unreasonable given the strict QC protocols that ensure high quality genotype data is available to most studies. We have restricted the scope to only exploring non-differential measurement error and avoided the complications incurred if measurement error in the exposure and outcome is correlated, however our analysis goes beyond many previous explorations of measurement error by assessing the impacts of both imprecision (noise) and linear transformations of the true variable on causal inference.

Our analysis of DNA methylation and gene expression summary data demonstates the simplicity of inferring causal existence, direction and magnitude in the two sample MR framework. We note that verifying these results with regards to the extent to which measurement error has influenced them empirically is a challenging task and beyond the scope of this work.

The overarching result from our simulations is that, regardless of the method used, inferring causal direction using an instrument of unknown biology is highly sensitive to measurement error. With the presence of measurement error near ubiquitous in most observational data, and our ability to measure it limited, we argue that it needs to be central to any consideration of approaches which are used in attempt to strengthen causal inference, and any putative results should be accompanied with appropriate sensitivity analysis that assesses their robustness under varying levels of measurement error.

Figures

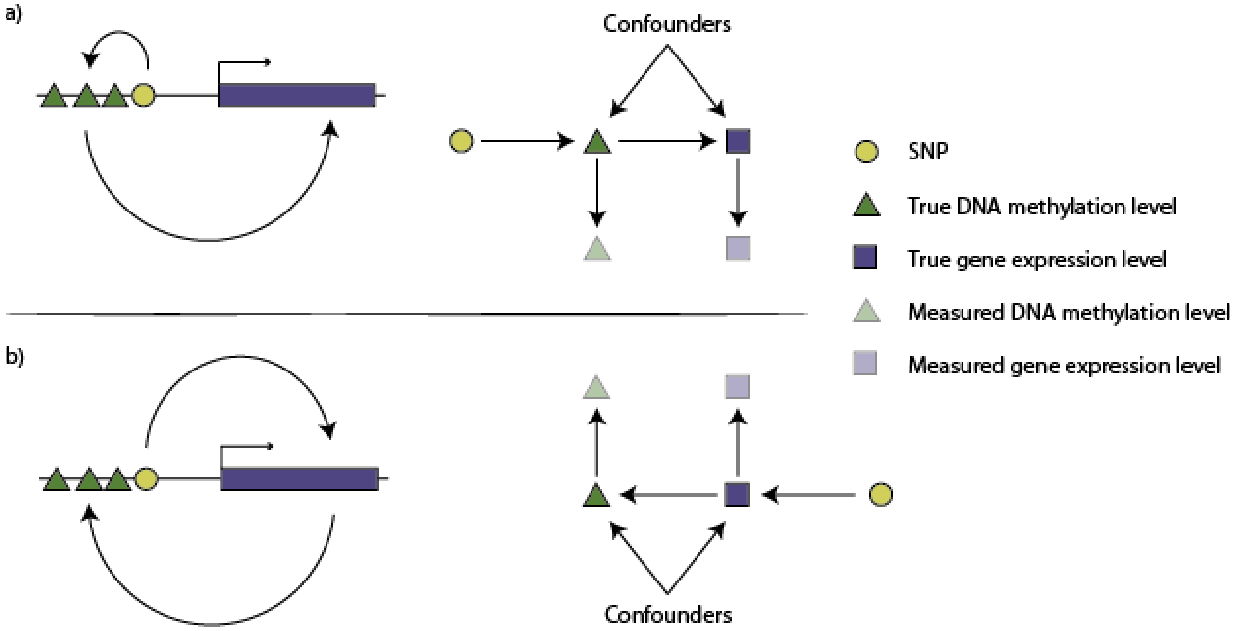
a)
Confounders
SNP
True DNA methylation level
True gene expression level
Measured DNA methylation level
Measured gene expression level
b)
Confounders

Figure 1: Gene expression levels (blue blocks) and DNA methylation levels (green triangles) may be correlated but the causal direction is unknown. If a SNP (yellow circle) is associated with both DNA methylation and gene expression levels then it can be used as an instrument, but there are two competing models for these variables. a) Methylation causes gene expression. The left figure shows that the SNP influences methylation levels that in turn influence gene expression levels. The right figure shows the directed acyclic graph that represents this model. Faded symbols represent the measured values whereas solid symbols represent the true values. b) The same as in A, except the causal direction is from gene expression to DNA methylation.

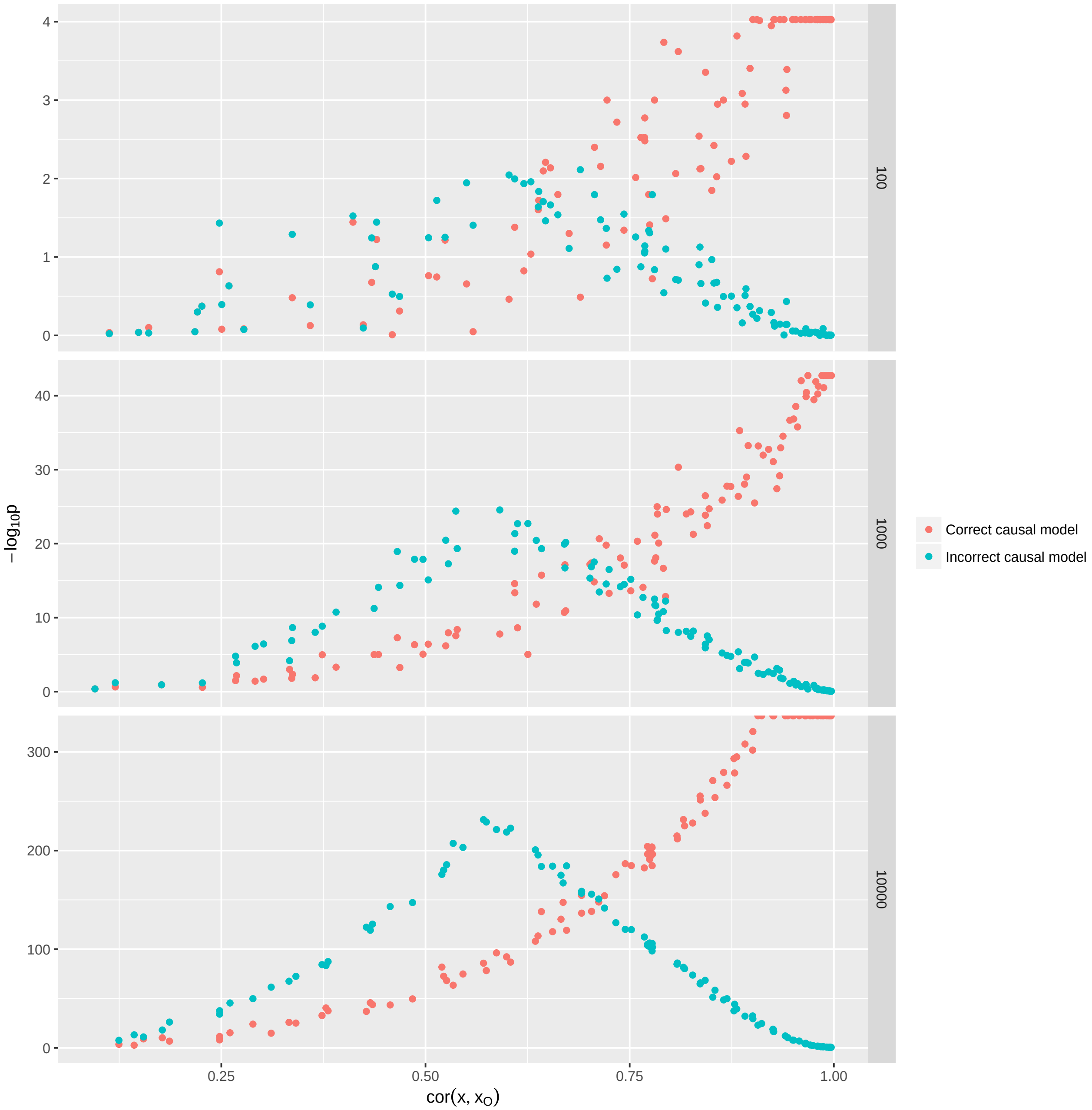

Figure 2: The CIT was performed on simulated variables where the exposure influenced the outcome and the exposure was instrumented by a SNP. The test statistic from CIT when testing if the exposure caused the outcome (the true model) is in red, and the test for the outcome causing the exposure (false model) is in green. Rows of plots represent the sample sizes used for the simulations. As measurement error increases (decreasing values on x-axis) the test statistic for the incorrect model gets stronger and the test statistic for the correct model gets weaker.

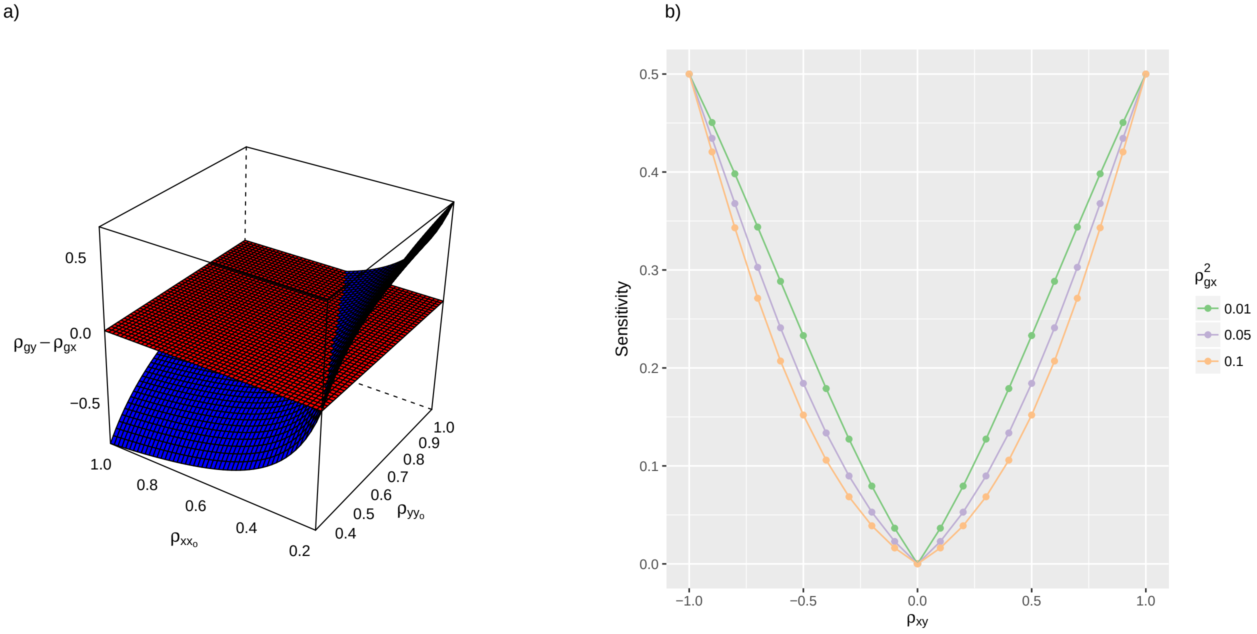

Figure 3: a) We can predict the values the Steiger test would take (z-axis) for different potential values of measurement error (x and y axes), drawn here as the blue surface. When $\rho_{g,y} > \rho_{g,x}$, as denoted by range of values where the blue surface is above the red plane, those values of measurement error lead to our observed Steiger test inferring the wrong causal direction. Where the blue surface lies below the red plane, these measurement error values support the inferred causal direction of X to Y. A measure of sensitivity, therefore, is the volume of the positive space bound by the blue surface as a proportion of the total space bound by the blue surface. Sensitivity of 0 denotes no measurement error values that could lead to inferrence of incorrect direction, whereas the maximum value of 0.5 denotes measurement error equal chance of either direction across all measurement error ranges. In this case, where $\rho^2_{g,x} = 0.01$ and $\rho^2_{x,y} = 0.1$, the sensitivity is 0.18. b) The influence of $\rho_{x,y}$ (x-axis) on the Steiger sensitivity values (y-axis) for different values of $\rho^2_{g,x}$. Inferring the direction of causality is more susceptible to measurement error when there are stronger causal associations between X and Y.

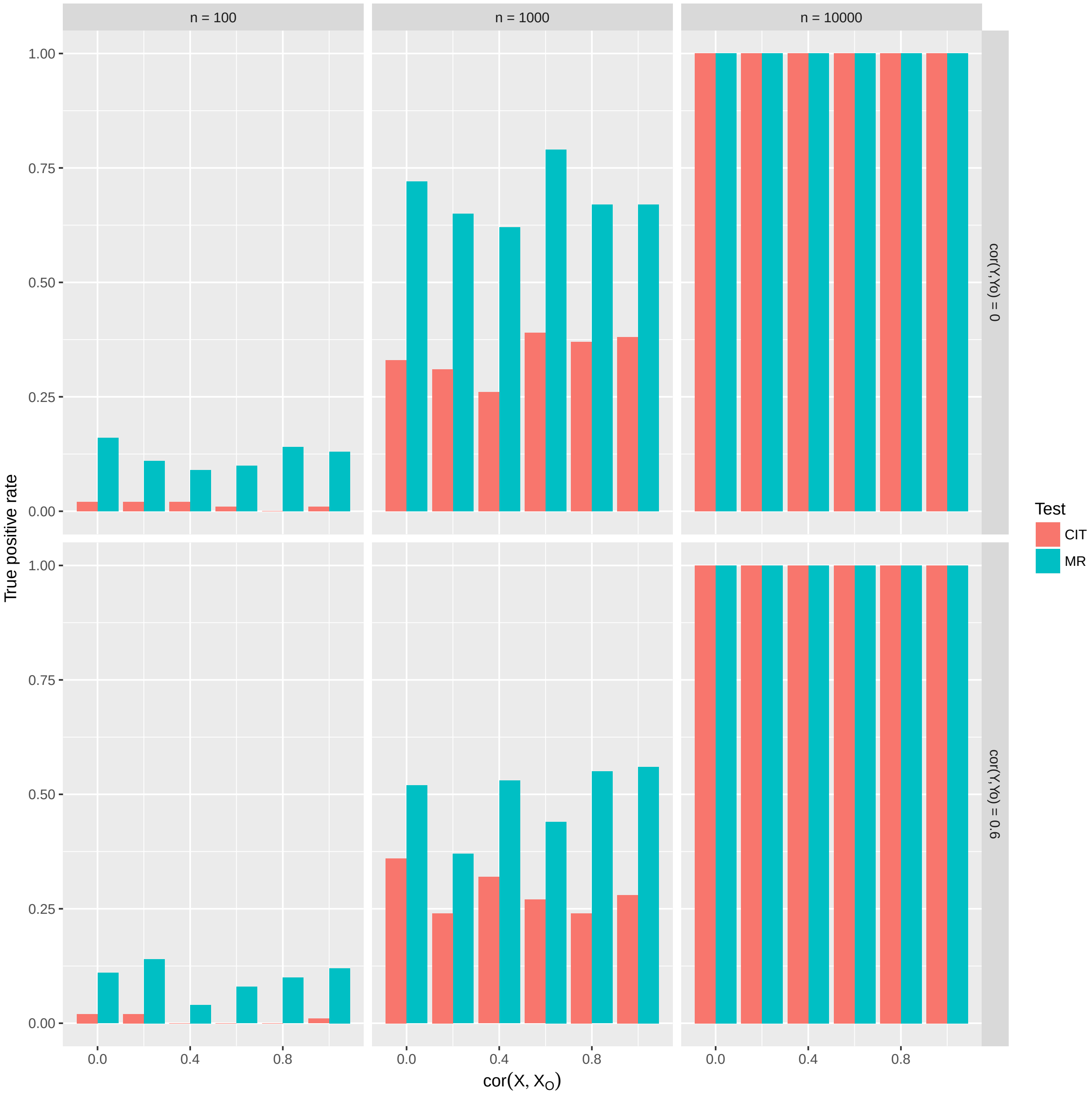

Figure 4: Outcomes were simulated to be causally influenced by exposures with varying degrees of measurement imprecision applied to the exposure variable (x axis). True positive rates (y axis) for MR and CIT were compared for varying levels of measurement imprecision in the outcome variable (rows of boxes) and sample sizes (columns of boxes).

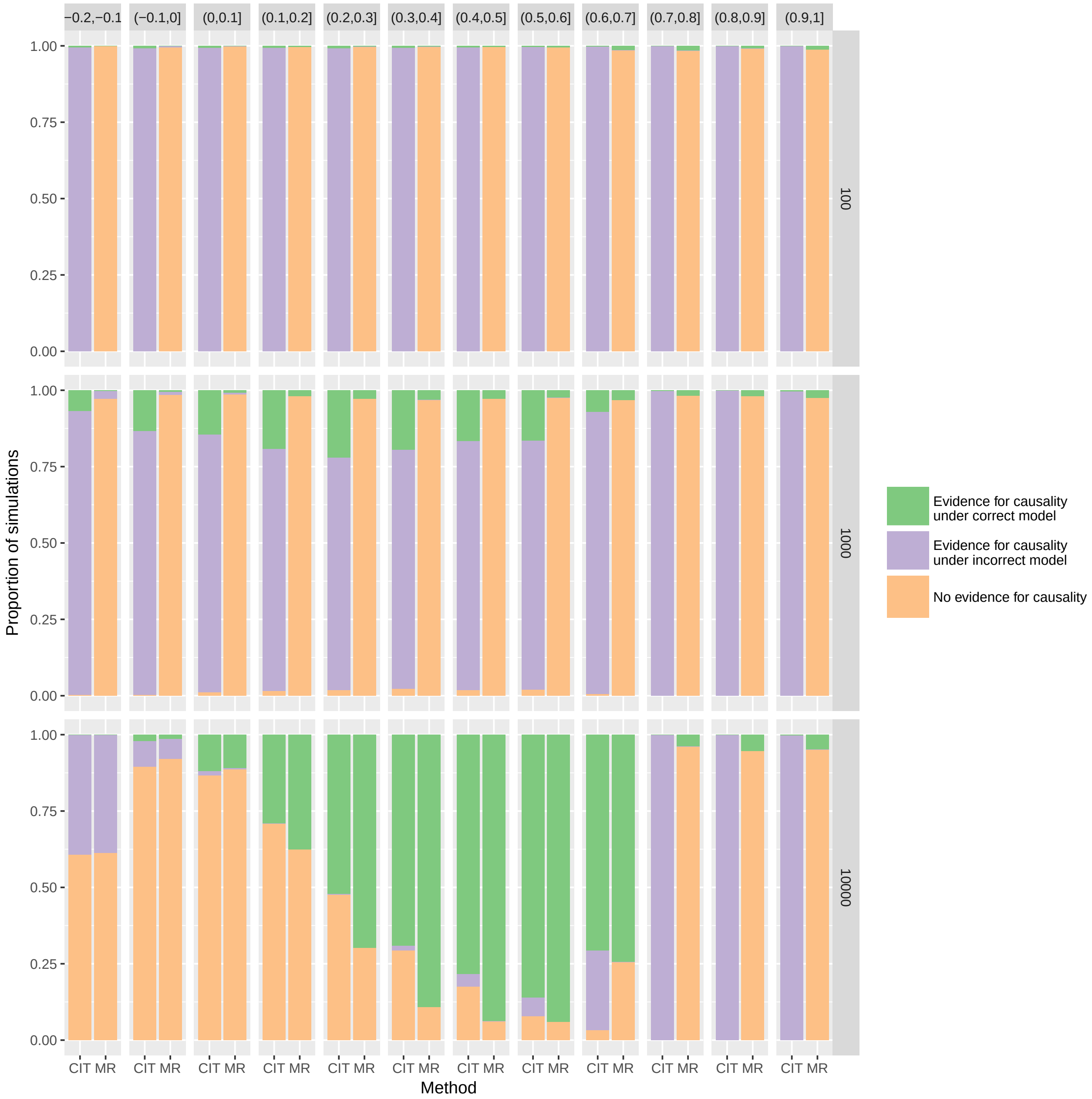

Figure 5: Outcomes were simulated to be causally influenced by exposures, with varying degrees of measurement error applied to both. CIT and MR were used to infer evidence for causality between the exposure and outcome, and to infer the direction of causality. The value of $d = \rho_{x,x_o} - \rho_{x,y}\rho_{y,y_o}$, such that when d is negative we expect the Steiger test to be more likely to be wrong about the direction of causality. Rows of graphs represent the sample size used in the simulations. For the CIT method, outcome 1 denoted evidence for causality with correct model, outcomes 2 or 3 denoted evidence for causality with incorrect model, and outcome 4 denoted no evidence for causality.

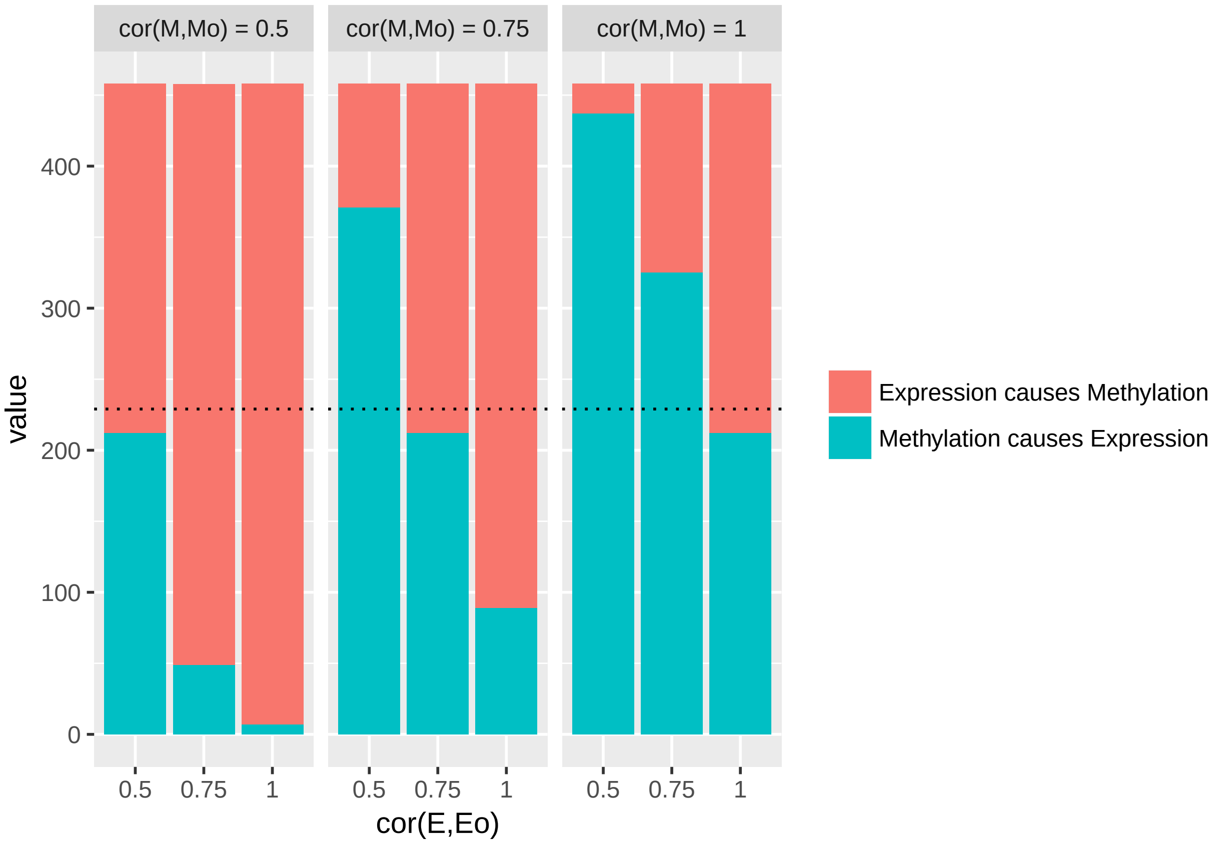

Figure 6: Using 458 putative associations between DNA methylation and gene expression we used the Steiger test to infer the direction of causality between them. The rightmost bar shows the proportion of associations for each of the two possible causal directions (colour key) assuming no measurement error in either gene expression or DNA methylation levels. The proportions change when we assume different levels of measurement error in gene expression levels (x-axis) or DNA methylation levels (columns of boxes). If there is systematically higher measurement error in one platform it will appear to be less likely to be the causal factor.

Appendix 1

We assume the following model

$$\begin{aligned}
x &= \alpha_g + \beta_g g + \epsilon_g \\
x_o &= \alpha_{mx} + \beta_{mx} x + \epsilon_{mx} \\
y &= \alpha_x + \beta_x x + \epsilon_x \\
y_o &= \alpha_{my} + \beta_{my} y + \epsilon_{my}
\end{aligned}$$

where x is the exposure on the outcome y, g is an instrument that has a direct effect on x, x_o is the measured quantity of x, where measurement error is incurred from linear transformation in α_{mx} and β_{mx} and imprecision from ϵ_{mx}, and y_o is the measured quantity of y, where measurement error is incurred from linear transformation in α_{my} and β_{my} and imprecision from ϵ_{my}. Our objective is to estimate the expected magnitude of association between g and y after conditioning on x. Under the CIT, this is expected to be $cov(g, y_o - \hat{y}_o) = 0$ when x causes y, where $\hat{y}_o = \hat{a}_{x_o} + \hat{\beta}_{x_o} x_o$ is the predicted value of y_o using the measured value of x_o.

We can split $cov(g, y_o - \hat{y}_o)$ into two parts, $cov(g, y_o)$ and $cov(g, \hat{y}_o)$.

Part 1

$$\begin{aligned}
cov(g, y_o) &= cov(g, \beta_{my} y) \\
&= cov(g, \beta_{my} \beta_x x) \\
&= cov(g, \beta_{my} \beta_x \beta_g g) \\
&= \beta_{my} \beta_x \beta_g var(g)
\end{aligned}$$

Part 2

$$\begin{aligned}
cov(g, \hat{y}_o) &= cov(g, \hat{\beta}_{x_o} x_o) \\
&= cov(g, \hat{\beta}_{x_o} \beta_{mx} x) \\
&= cov(g, \hat{\beta}_{x_o} \beta_{mx} \beta_g g) \\
&= \hat{\beta}_{x_o} \beta_{mx} \beta_g var(g)
\end{aligned}$$

Simpifying further

$$\begin{aligned}
\hat{\beta}_{x_o} &= \frac{cov(y_o, x_o)}{var(x_o)} \\
&= \frac{cov(\beta_{my} y, \beta_{mx} x)}{\beta_{mx}^2 var(x) + var(\epsilon_{mx})} \\
&= \frac{\beta_{mx} \beta_{my} cov(y, x)}{\beta_{mx}^2 var(x) + var(\epsilon_{mx})} \\
&= \frac{\beta_{mx} \beta_{my} \beta_x var(x)}{\beta_{mx}^2 var(x) + var(\epsilon_{mx})}
\end{aligned}$$

which can be substituted back to give

$$\begin{aligned}
cov(g, \hat{y}_o) &= \frac{\beta_{my} \beta_x \beta_g var(g) \beta_{mx}^2 var(x)}{\beta_{mx}^2 var(x) + var(\epsilon_{mx})} \\
&= \frac{\beta_{mx}^2 var(x)}{\beta_{mx}^2 var(x) + var(\epsilon_{mx})} \times \beta_{my} \beta_x \beta_g var(g)
\end{aligned}$$

Finally

$$cov(g, y_o - \hat{y}_o) = \beta_{my}\beta_x\beta_g var(g) - \frac{\beta_{mx}^2 var(x)}{\beta_{mx}^2 var(x) + var(\epsilon_m)} \times \beta_{my}\beta_x\beta_g var(g)$$

thus $cov(g, y_o - \hat{y}_o) = 0$ if the measurement imprecision in x_o is $var(\epsilon_m) = 0$. However if there is any imprecision then the condition $cov(g, y_o - \hat{y}_o) = 0$ will not hold.

Appendix 2

The Steiger test is used to infer if the variant g has a direct influence on x or y when it is known that it associates with both, but the direction of causality between x and y is unknown. Assuming the causal direction is $x \rightarrow y$, two stage MR is formulated using the following regression models:

$$x = \alpha_1 + \beta_1 g + e_1$$

for the first stage and

$$y = \alpha_2 + \beta_2 \hat{x} + e_2$$

where $\hat{x} = \hat{\alpha}_1 + \hat{\beta}_1 g$. Writing in scale free terms, $\rho_{g,x}$ denotes the correlation between g and the exposure variable x, and it is expected that $\rho_{g,x} > \rho_{g,y}$ because $\rho_{g,y} = \rho_{g,x}\rho_{x,y}$, where $\rho_{x,y}$ is the causal association between x and y (which is likely to be less than 1).

In the presence of measurement error in x and y, however, the Steiger test will instead be assessing the inequality $\rho_{g,x_o} > \rho_{g,y_o}$. In order to assess how sensitive the Steiger test is to measurement error, we can evaluate the range of potential values of measurement error in x and y over which the empirical result from the Steiger test would return the wrong value.

For different values of ρ_{x,x_o}, $\rho_{g,x} = \frac{\rho_{g,x_o}}{\rho_{x,x_o}}$ and $\rho_{g,x_o} \leq \rho_{x,x_o} \leq 1$. For different values of ρ_{y,y_o}, $\rho_{g,y} = \frac{\rho_{g,y_o}}{\rho_{y,y_o}}$ and $\rho_{g,y_o} \leq \rho_{y,x_o} \leq 1$.

Where $z = \rho_{g,y} - \rho_{g,x} < 0$ the measurement error range in x_o and y_o will not change the qualitative inference of the direction of causality. To evaluate the sensitivity, S, of the Steiger test to measurement error, the objective is to calculate the proportion of the total absolute volume of this function that is above the $z = 0$ plane,

$$S = \frac{V_{z>0}}{2V_{z>0} + |V_z|}$$

V_z, the total volume of the function (Figure 3), can be obtained analytically by solving:

$$\begin{aligned} V_z &= \int_{\rho_{g,x_o}}^{1} \int_{\rho_{g,y_o}}^{1} \frac{\rho_{g,y_o}}{\rho_{y,y_o}} - \frac{\rho_{g,x_o}}{\rho_{x,x_o}} \;\; d\rho_{y,y_o} d\rho_{x,x_o} \\ &= \rho_{g,x_o} log(\rho_{g,x_o}) - \rho_{g,y_o} log(\rho_{g,y_o}) + \rho_{g,x_o}\rho_{g,y_o}(log(\rho_{g,y_o}) - log(\rho_{g,x_o})) \end{aligned}$$

$V_{z>0}$, the proportion of the volume that lies above the $z = 0$ plane, can also be obtained analytically. The region of this volume is bound by the values of ρ_{x,x_o} and ρ_{y,y_o} where $0 = \rho_{g,y} - \rho_{g,x}$, which can be expanded to $\rho_{y,y_o} = \rho_{g,y_o}\rho_{x,x_o}/\rho_{g,x_o}$. Hence,

$$\begin{aligned} V_{z>0} &= \int_{\rho_{g,x_o}}^{1} \int_{\rho_{g,y_o}}^{\frac{\rho_{g,y_o}\rho_{x,x_o}}{\rho_{g,x_o}}} \frac{\rho_{g,y_o}}{\rho_{y,y_o}} - \frac{\rho_{g,x_o}}{\rho_{x,x_o}} \;\; d\rho_{y,y_o} d\rho_{x,x_o} \\ &= 2\rho_{g,x_o}\rho_{g,y_o} - 2\rho_{g,y_o} - \rho_{g,y_o} log(\rho_{g,x_o}) - \rho_{g,x_o}\rho_{g,y_o} log(\rho_{g,x_o}) \end{aligned}$$

The sensitivity, S, can take values from 0 to 0.5, where low values suggest there are

References

1. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature genetics* **44,** 269–76 (2012).

2. Bird, A. DNA methylation patterns and epigenetic memory. *Genes & development* **16,** 6–21 (2002).

3. Davey Smith, G. & Ebrahim, S. Mendelian randomization: prospects, potentials, and limitations. *International journal of epidemiology* **33,** 30–42 (2004).

4. Phillips, A. N. & Davey Smith, G. How independent are 'independent' effects? relative risk estimation when correlated exposures are measured imprecisely. *Journal of Clinical Epidemiology* **44,** 1223–1231 (1991).

5. Davey Smith, G. & Ebrahim, S. Data dredging, bias, or confounding. *BMJ* **325,** (2002).

6. Houle, D., Pélabon, C., Wagner, G. & Hansen, T. Measurement and meaning in biology. *The Quarterly Review of Biology* **86,** 3–34 (2011).

7. Hernán, M. a & Cole, S. R. Invited Commentary: Causal diagrams and measurement bias. *American journal of epidemiology* **170,** 959–62; discussion 963–4 (2009).

8. Harper, K. N., Peters, B. a & Gamble, M. V. Batch effects and pathway analysis: two potential perils in cancer studies involving DNA methylation array analysis. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **22,** 1052–60 (2013).

9. Chen, Y.-a. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics : official journal of the DNA Methylation Society* **8,** 203–9 (2013).

10. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* **13,** 86 (2012).

11. Ahima, R. S. & Lazar, M. A. Physiology. The health risk of obesity–better metrics imperative. *Science (New York, N.Y.)* **341,** 856–8 (2013).

12. Fewell, Z., Davey Smith, G. & Sterne, J. A. C. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *American journal of epidemiology* **166,** 646–55 (2007).

13. Cessie, S. le, Debeij, J., Rosendaal, F. R., Cannegieter, S. C. & Vandenbroucke, J. P. Quantification of bias in direct effects estimates due to different types of measurement error in the mediator. *Epidemiology (Cambridge, Mass.)* **23,** 551–60 (2012).

14. Blakely, T., McKenzie, S. & Carter, K. Misclassification of the mediator matters when estimating indirect effects. *Journal of epidemiology and community health* **67,** 458–66 (2013).

15. Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37,** 710–717 (2005).

16. Millstein, J., Zhang, B., Zhu, J. & Schadt, E. E. Disentangling molecular relationships with a causal inference test. *BMC genetics* **10,** 23 (2009).

17. Millstein, J. cit: Causal Inference Test. R package version 1.9. (2016). at <http://cran.r-project.org/package=cit>

18. Aten, J. E., Fuller, T. F., Lusis, A. J. & Horvath, S. Using genetic markers to orient the edges in quantitative trait networks: the NEO software. *BMC systems biology* **2,** 34 (2008).

19. Waszak, S. M. *et al.* Variation and genetic control of chromatin architecture in humans. *Cell* **162,** 1039–1050 (2015).

20. Koestler, D. C. *et al.* Integrative genomic analysis identifies epigenetic marks that mediate genetic risk

for epithelial ovarian cancer. *BMC medical genomics* **7,** 8 (2014).

21. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology* **31,** 142–7 (2013).

22. Yuan, W. *et al.* An integrated epigenomic analysis for type 2 diabetes susceptibility loci in monozygotic twins. *Nature communications* **5,** 5719 (2014).

23. Tang, Y. *et al.* Genotype-based treatment of type 2 diabetes with an α2A-adrenergic receptor antagonist. *Science translational medicine* **6,** 257ra139 (2014).

24. Hong, X. *et al.* Genome-wide association study identifies peanut allergy-specific loci and evidence of epigenetic mediation in US children. *Nature communications* **6,** 6304 (2015).

25. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *The New England journal of medicine* **373,** 895–907 (2015).

26. Timpson, N. J. *et al.* C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. *International journal of obesity (2005)* **35,** 300–8 (2011).

27. Shakhbazov, K. *et al.* Shared genetic control of expression and methylation in peripheral blood. *BMC genomics* **17,** 278 (2016).

28. Pierce, B. L. & Burgess, S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *American journal of epidemiology* **178,** 1177–84 (2013).

29. Ashenfelter, O. & Krueger, A. B. Estimates of the Economic Return to Schooling from a New Sample of Twins. *The American Economic Review* **84,** 1157–1173 (1994).

30. Nagarajan, R. & Scutari, M. Impact of noise on molecular network inference. *PloS one* **8,** e80735 (2013).

31. Shpitser, I., VanderWeele, T. & Robins, J. On the validity of covariate adjustment for estimating causal effects. *Proceedings of the Twenty Sixth Conference on Uncertainty in Artificial Intelligence (UAI-10)* 527–536 (2010).

32. Wang, L. & Michoel, T. Detection of regulator genes and eQTLs in gene networks. *arXiv* **arXiv:1512,** (2015).

33. Lagani, V., Triantafillou, S., Ball, G., Tegner, J. & Tsamardinos, I. in *Uncertainty in biology: A computational modeling approach* 47 (Springer, 2015). at <https://books.google.com/books?id=8SLUCgAAQBAJ{\&}pgis=1>

34. R Core Team. R: A Language and Environment for Statistical Computing. (2015). at <https://www.r-project.org/>

35. Henningsen, A. & Hamann, J. D. systemfit : A Package for Estimating Systems of Simultaneous Equations in R. *Journal of Statistical Software* **23,** 1–40 (2007).

36. Steiger, J. H. Tests for comparing elements of a correlation matrix. *Psychological Bulletin* **87,** 245–251 (1980).

37. Revelle, W. psych: Procedures for Psychological, Psychometric, and Personality Research. (2015). at <http://cran.r-project.org/package=psych>

38. Relton, C. L. & Davey Smith, G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *International journal of epidemiology* **41,** 161–76 (2012).

39. Burgess, S., Freitag, D. F., Khan, H., Gorman, D. N. & Thompson, S. G. Using multivariable Mendelian randomization to disentangle the causal effects of lipid fractions. *PloS one* **9,** e108891 (2014).

40. Burgess, S., Daniel, R. M., Butterworth, A. S. & Thompson, S. G. Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *International*

journal of epidemiology **44,** 484–95 (2015).

41. Varbo, A. *et al.* Remnant cholesterol, low-density lipoprotein cholesterol, and blood pressure as mediators from obesity to ischemic heart disease. *Circulation research* **116,** 665–73 (2015).

42. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **In press,** (2015).