# Improving the accuracy of two-sample summary data Mendelian randomization: moving beyond the NOME assumption

Jack Bowden[1,2*], Fabiola Del Greco M[3], Cosetta Minelli[4], Qingyuan Zhao[5], Debbie A Lawlor[1,2], Nuala A Sheehan[6] John Thompson[6] George Davey Smith[1,2]

[1] *MRC Integrative Epidemiology Unit at the University of Bristol, U.K*
[2] *Population Health Sciences, University of Bristol, U.K*
[3] *Institute for Biomedicine, Eurac Research, Bolzano, Italy*
[4] *Population Health and Occupational Disease, NHLI, Imperial College, London, U.K*
[5] *Department of Statistics, The Wharton School, University of Pennsylvania, U.S.A.*
[6] *Department of Health Sciences, University of Leicester, Leicester, U.K*

*Address for correspondence:
Jack Bowden
MRC Integrative Epidemiology Unit
Oakfield House, Bristol, BS8 2BN, U.K
jack.bowden@bristol.ac.uk.

## Abstract

**Background:** Two-sample summary data Mendelian randomization (MR) incorporating multiple genetic variants within a meta-analysis framework is a popular technique for assessing causality in epidemiology. If all genetic variants satisfy the instrumental variable (IV) and necessary modelling assumptions, then their individual ratio estimates of causal effect should be homogeneous. Observed heterogeneity signals that one or more of these assumptions could have been violated.

**Methods:** Causal estimation and heterogeneity assessment in MR requires an approximation for the variance of each ratio estimate. We show that the most popular (1st order) approximation can lead to an inflation in the chances of detecting heterogeneity when in fact it is not present. Conversely, an ostensibly more accurate (2nd order) approximation can dramatically increase the chances of failing to detect heterogeneity, when it is truly present. Here we derive a modified 2nd order approximation to the variance that makes use of the derived causal estimate to mitigate both of these adverse effects.

**Results:** Using Monte Carlo simulations, we show that the modified 2nd order approximation outperforms both its 1st and 2nd order counterparts in terms of heterogeneity quantification and causal estimation. The added benefit is most noticeable when the genetic instruments are weak, or the causal effect is large. We illustrate the utility of the new method using data from a recent two-sample summary data MR analysis to assess the causal role of systolic blood pressure on coronary heart disease risk.

**Conclusions:** We propose the use of modified 2nd order weighting within two-sample summary data MR studies for model fitting, quantifying heterogeneity and outlier detection.

**Key words**: Two-sample summary data Mendelian randomization, Inverse variance weighted estimate; Cochran's $Q$ statistic; Outlier detection.

# Introduction

Mendelian randomization (MR) [1] is an instrumental variable approach that uses genetic data, typically in the form of single nucleotide polymorphisms (SNPs), to assess whether a modifiable exposure exerts a causal effect on a health outcome in the presence of unmeasured confounding. Traditionally, researchers have assumed that SNPs used for MR studies are valid instrumental variables (IVs) for the purposes of inferring the causal effect of an exposure, $X$, on an outcome, $Y$. Specifically, the SNP is: associated with $X$ (IV assumption 1 (IV1)); not associated with any confounders of $X$ and $Y$ (IV2); and can only be associated with $Y$ through $X$ (IV3). The IV assumptions are represented by the solid lines in Figure 1 for a SNP $G_j$, with unobserved confounding represented by $U$. Dotted lines represent dependencies between $G$ and $U$, and $G$ and $Y$ that are prohibited by the IV assumptions.

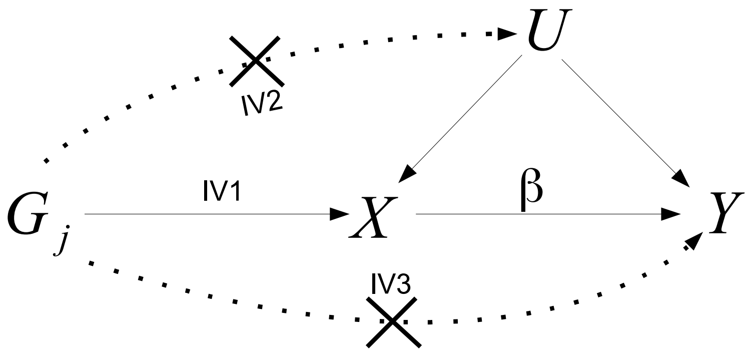

Figure 1: *Causal diagram representing the IV assumptions (and violations thereof) for a SNP $G_j$, an exposure $X$ and an outcome $Y$. The causal effect of $X$ on $Y$, denoted by $\beta$, is the quantity we wish to estimate.*

Suppose initially that a SNP, $G_j$, is a valid IV. Further assume that the association of $G_j$ with $X$ and $X$ with $Y$ are linear with no effect modification. The underlying SNP-outcome association $\Gamma_j$ - the increase in $Y$ for a unit increase in $G_j$ - can then be expressed as a scalar multiple of the underlying SNP-exposure association estimate, $\gamma_j$ - the increase in $X$ for a unit increase in $G_j$. That is: $\Gamma_j = \beta\gamma_j$, where $\beta$ denotes the causal effect of a unit increase in $X$ on the outcome $Y$.

Figure 1 encodes the assumptions that are traditionally required for a single sample of individuals for whom $G$, $X$ and $Y$ are measured. A particular MR study design gaining in popularity instead combines publically available summary data on SNP-exposure and SNP-outcome associations from two separate studies for large numbers of uncorrelated variants $G_1$,...,$G_L$ within the framework of a meta-analysis. These studies should ideally contain no overlapping individuals (to ensure independence) but should also originate from the same source population. This is referred to as two-sample summary data MR [2]. Providing the aforementioned modelling assumptions are met and each SNP is a valid IV, when SNP-exposure and SNP-outcome associations are estimated from their respective samples the ratio $\hat{\beta}_j = \hat{\Gamma}_j/\hat{\gamma}_j$ for any single variant should also provide a consistent estimate for $\beta$. Combining the set of $L$ ratio estimates obtained across all variants into an overall inverse variance weighted (IVW) estimate using the standard meta-analytic formula:

$$\hat{\beta}_{IVW} = \frac{\sum_{j=1}^{L} w_j \hat{\beta}_j}{\sum_{j=1}^{L} w_j} \quad \text{where} \quad w_j = var(\hat{\beta}_j)^{-1}, \tag{1}$$

then provides an efficient and consistent estimate for $\beta$. For a more detailed description of the assumptions required by two-sample summary data MR, see Bowden et al. [3] and Zhao et al. [4].

## Heterogeneity assessment

If the aforementioned modelling and IV assumptions hold, then Cochran's $Q$ statistic:

$$Q = \sum_{j=1}^{L} Q_j = \sum_{j=1}^{L} w_j(\hat{\beta}_j - \hat{\beta}_{IVW})^2, \tag{2}$$

should follow, asymptotically, a $\chi^2$ distribution on $L$-1 degrees of freedom. Excessive heterogeneity could therefore indicate the modelling assumptions have been violated, or that some of the genetic variants violate the IV assumptions.

For example, the modelling assumptions are very often knowingly violated in applied MR studies when the outcome is binary and the SNP-outcome association is measured an odds ratio [5]. Indeed, the MR study we analyse in this paper relates to a binary outcome. In this case, the causal effect identified by each SNP will in general depend on its strength of association with the outcome, rather than being equal to a constant value $\beta$. This will induce some theoretical heterogeneity amongst the causal estimates, even when all variants are valid instruments. In practice, the magnitude of this effect will be negligible because each SNP explains a very small amount of variation in the outcome (see [6] for further details). It could instead be the case that a SNP actually increases the exposure for one group of individuals and decreases it for another, which would be a violation of the monotonicity assumption [7].

Another potential source of heterogeneity that has received a lot of attention in the literature is that some or all of the SNPs may exert a direct effect on the outcome not through the exposure [8] by violating assumptions IV2, IV3 or both, which is termed 'horizontal pleiotropy' [9, 10]. 'Vertical pleiotropy' - in which the effect of a SNP on the exposure of interest is actually mediated through other, earlier exposures, does not pose a problem. For brevity we will refer to problematic horizontal pleiotropy simply as pleiotropy from now on.

The presence of heterogeneity does not necessarily invalidate an MR study. For example if the underlying cause of the heterogeneity is pleiotropy but, across all variants: (i) the amount of pleiotropy is independent of instrument strength (the InSIDE assumption [11]) and (ii) it has a zero mean , then a standard random effects meta-analysis will still yield reliable inferences [3, 11]. If the InSIDE assumption is satisfied but the pleiotropy is instead 'directional' (i.e. it has a non-zero mean) then a random effects meta-analysis will be biased, but MR-Egger regression [11] can still yield reliable inferences. MR-Egger regression has been used extensively as a sensitivity analysis tool since its proposal for this reason (see [12, 13, 14, 15, 16] for some recent examples), although its estimate can often suffer from a lack of precision. In this paper we choose to focus solely on the IVW estimate.

### Choice of weights in two-sample summary data MR

Two popular choices for the inverse variance weights used to calculate both the IVW estimate in (1) and Cochran's $Q$ in (2) are:

$$\text{1st order weights:} \quad w_j = \frac{\hat{\gamma}_j^2}{\sigma_{Yj}^2} \tag{3}$$

$$\text{2nd order weights:} \quad w_j = \left(\frac{\sigma_{Yj}^2}{\hat{\gamma}_j^2} + \frac{\hat{\Gamma}_j^2\sigma_{Xj}^2}{\hat{\gamma}_j^4}\right)^{-1} \tag{4}$$

where $\sigma_{Yj}^2$ represents the variance of $\hat{\Gamma}_j^2$ and $\sigma_{Xj}^2$ represents the variance of $\hat{\gamma}_j^2$. Provided that the two samples used in the analysis are homogeneous, and the SNPs used as IVs are mutually independent, the IVW estimate obtained using 1st order weights is asymptotically equivalent to the two-stage least squares (TSLS) estimate for the causal effect obtained using individual level data on $G$, $X$ and $Y$ from either sample, if such data were available (see for example Section 2.2 in [17]). 2nd order weights (see for example Thomas et al. [18]), which are derived via a Taylor series expansion, attempt to acknowledge uncertainty in both the numerator and denominator of the ratio estimate. In the two-sample setting, the Taylor series expansion is simplified because it is not necessary to include terms involving the covariance of $\hat{\gamma}_j$ and $\hat{\Gamma}_j$, since they are obtained from independent samples.

### Choice of weights and the NOME assumption

1st order weights ignore uncertainty in the denominator of the ratio estimate, which is equivalent to making the 'NO Measurement Error' (NOME) assumption, as defined by Bowden et al. [19] within the context of a two-sample MR analysis. The NOME assumption reminds the practitioner that the SNP-exposure association estimates, $\hat{\gamma}_j$, which play the role of the explanatory variable in both the IVW and MR-Egger regression models, are only equal to the true associations, $\gamma_j$, when measured with infinite precision. In practice, therefore, NOME is always violated, and so $\hat{\gamma}_j$ can be viewed as the association, $\gamma_j$, plus some uncertainty or error, with mean zero and variance $\sigma_{Xj}^2$. It is helpful to conceptualize this uncertainty as measurement error because: (a) it induces classical regression dilution bias in the IVW estimate towards the null, and (b); it can be detected (and corrected) using established methods from the measurement error literature, such as Simulation Extrapolation [3, 19, 20].

In practice when SNP-exposure association estimates are known very precisely, for example when they are derived from a study with a large sample size, then the NOME assumption is only very weakly violated. In this case, 1st and 2nd order weighting will give near identical results. Unfortunately, this is not always the case, for example if no SNPs can be identified at 'genome-wide significance' for the trait of interest and a less stringent threshold is instead used.

Given that 2nd order weights provide an ostensibly more accurate reflection of the variance of each ratio estimate, it would seem obvious that they should be used as standard within an MR study to calculate the IVW estimate and Cochran's $Q$. However, Thompson et al. [21] showed that 2nd order weights produce causal estimates which are generally more biased than using 1st order weights. The reason for this apparent paradox is that 2nd order weights can be highly correlated with the ratio estimates themselves. Strict independence is required between the $w_j$ and $\hat{\beta}_j$ terms in (1) in order for the IVW estimate to function as intended.

## Further remarks on Cochran's $Q$

Following recent methodological work by Windmeijer [22] on the TSLS estimator in the single sample setting, it is possible to view Cochran's $Q$ statistic not just as a method for quantifying heterogeneity, but as a tool for directly estimating the causal effect. That is, the IVW estimate $\hat{\beta}_{IVW}$ in (1) actually minimises Cochran's $Q$, so that:

$$\frac{\partial Q}{\partial \beta}(\beta = \hat{\beta}_{IVW}) = 0.$$

This expression presents Cochran's $Q$ as an estimating equation for the causal parameter $\beta$. We build on this idea by deriving an alternative estimating equation based on an extended version of Cochran's $Q$ statistic, which uses (what we term) 'modified 2nd order' weights. Our new $Q$ statistic is shown to yield IVW estimates that outperform those obtained from either 1st or 2nd order weights, and also enables heterogeneity to be more reliably detected. We conclude by applying our improved $Q$ statistic to a recent two-sample summary data MR study to determine the causal effect of systolic blood pressure on the binary outcome of coronary heart disease originally published by Ference et al. [23].

# Methods

We start by motivating the derivation of the IVW estimate using 1st and 2nd order weights. We assume the basic underlying model generating the observed SNP-outcome association estimates:

$$\text{True model:} \quad \hat{\Gamma}_j = \beta\gamma_j + \sigma_{Yj}\epsilon_j, \tag{5}$$

Here $\sigma_{Yj}$ is the known standard error of the $j$th SNP-outcome association and $\epsilon_j$ is a standard normal random error variable with mean zero and variance 1. Note that model (5) is a function of the true underlying SNP-exposure association $\gamma_j$. It is also idealised, in the sense that it assumes a constant causal effect (see our earlier discussion of binary outcomes). This simplification means that data generated under model (5) contains no underlying heterogeneity (e.g. due to pleiotropy), so that we can transparently assess the effect of different weighting schemes on the amount of apparent heterogeneity actually detected.

In practice, when fitting this model we must work with the SNP-exposure association estimate $\hat{\gamma}_j$ (with variance $\sigma^2_{Xj}$) instead. Substituting $\hat{\gamma}_j$ into (5) in place of $\gamma_j$ therefore yields the fitted model:

$$\text{Fitted model:} \quad \hat{\Gamma}_j = \beta\hat{\gamma}_j + \sqrt{\beta^2\sigma^2_{Xj} + \sigma^2_{Yj}}\epsilon^{'}_j, \tag{6}$$

where $\epsilon'_j$ again represents a standard normal random error. Note that the random variation around the fitted model is inflated compared to the true model by the additional factor $\beta^2\sigma^2_{Xj}$. We can derive an expression for the ratio estimate $\hat{\beta}_j$ and its variance that is consistent with 2nd order weighting, by replacing $\beta$ with $\hat{\Gamma}_j/\hat{\gamma}_j$ in equation (6), and by dividing through by $\hat{\gamma}_j$ to give:

$$\hat{\beta}_j = \beta + \sqrt{\frac{\hat{\Gamma}^2_j}{\hat{\gamma}^4_j}\sigma^2_{Xj} + \frac{\sigma^2_{Yj}}{\hat{\gamma}^2_j}}\epsilon^{'}_j. \tag{7}$$

Setting $\sigma^2_{Xj}$ in formula (7) to zero (the NOME assumption) yields an expression for the ratio estimate $\hat{\beta}_j$ and its variance that is consistent with 1st order weighting.

## Modified 2nd order weights

### An iterative approach

Replacing $\beta$ with $\hat{\Gamma}_j/\hat{\gamma}_j$ in equation (6), as suggested by 2nd order weighting, means that the variance of each ratio estimate will be a function of the ratio estimate itself. It is easy to see that this will induce a negative bias in the IVW estimate because whenever $\hat{\beta}_j$ is randomly large, its contribution to (1) will be down-weighted (likewise its contribution to (1) will be up-weighted when $\hat{\beta}_j$ is randomly small). This negative bias will also effect Cochran's $Q$ statistic. This problem is crudely avoided when using 1st order weights by artificially setting $\sigma^2_{Xj}$ to zero, but the obvious downside is that the variance of each $\hat{\beta}_j$ is then under-estimated. We therefore suggest the following scheme to address both shortcomings, by plugging in an overall estimate for $\beta$ in model (6) instead. The procedure for calculating the weights is as follows:

1. Use 1st order weights and formula (1) to derive the IVW estimate, $\hat{\beta}_{IVW}$;

2. Calculate 'modified 2nd order weights' via the formula:

$$w_j(\hat{\beta}_{IVW}) = \left( \frac{\sigma^2_{Yj} + \hat{\beta}^2_{IVW}\sigma^2_{Xj}}{\hat{\gamma}^2_j} \right)^{-1} \tag{8}$$

   where $\hat{\beta}_{IVW}$ is obtained from step 1;

3. Use the weights in step 2 to iteratively re-calculate $\hat{\beta}_{IVW}$ and Cochran's $Q$ statistic.

Completing steps 1-3 constitutes one full iteration of the re-weighting scheme.

Procedures like this have an established pedigree in econometrics within the generalized-method-of-moments (GMM) framework, and is referred to as 'two-step GMM' [24]. Our contribution has been to describe how this procedure can be implemented using Cochran's $Q$ statistic in the two-sample summary data MR setting.

### An exact approach

It will be shown that iterative re-calculation of our modified 2nd order weights can dramatically improve the statistical properties of Cochran's $Q$ statistic and its associated IVW estimate. However, regardless of the number of iterations performed, this procedure will not in general yield the same $Q$ statistic or IVW estimate as that obtained from directly minimising the generalised $Q$ statistic, $Q_m(\beta)$, where:

$$Q_m(\beta) = \sum_{j=1}^{L} w_j(\beta)(\hat{\beta}_j - \beta)^2, \tag{9}$$

and where $w_j(\beta)$ is taken from formula (8). That is, finding the value of $\beta$ such that $\frac{\partial Q_m(\beta)}{\partial \beta} = 0$. We refer to this approach as the 'exact' application of modified 2nd order weights. It can be viewed as a procedure to obtain a limited information maximum likelihood (LIML) estimate in the two-sample summary data MR setting [6]. We will subsequently highlight the role that both iterative and exact weights can play in improving the IVW analysis.

## Performance of Cochran's $Q$ under no pleiotropy

We would like Cochran's $Q$ to follow a $\chi^2_{L-1}$ distribution as closely as possible when no heterogeneity is present, to guard against the erroneous detection of pleiotropy. To assess the performance of all weighting schemes in this regard, two-sample summary data MR studies comprising $L$=25 SNP-exposure and SNP outcome association estimates $(\hat{\Gamma}_j, \hat{\gamma}_j)$ were generated from the following normal models:

$$\hat{\gamma}_j \sim N(\gamma_j, \sigma^2_{Xj}), \quad \hat{\Gamma}_j \sim N(\beta\gamma_j, \sigma^2_{Yj}) \tag{10}$$

given parameter vector values for $(\gamma_j, \sigma^2_{Xj}, \sigma^2_{Yj})$ and the causal parameter $\beta$. Under these models, the $F$-statistic for SNP $j$ can be approximated by $\hat{\gamma}^2_j/\sigma^2_{Xj}$. Data generated under model (10) furnishes a set of ratio estimates between which no additional variation should exist as their $F$-statistics grows large (because NOME is satisfied), or if the causal effect ($\beta$) equals zero. To highlight this the $\gamma_j$ parameters were simulated from a Uniform(0.34,1.1) distribution and $\sigma_{Xj}$ was simulated from a Uniform(0.06,UB) distribution. By varying UB between 0.095 and 1 we were able to mimic MR studies with weak instruments (a mean $F$-statistic of 10) and strong instruments (a mean $F$-statistic of 100). Data were simulated for a range of causal effects and, across all scenarios, $\sigma_{Yj}$ was simulated from a uniform(0.015,0.11) distribution.

Table 1 (columns 2-9) show the mean $Q$ statistic and the probability of the $Q$ statistic detecting heterogeneity at the 5% significance level (the type I error rate), when using 1st order, 2nd order and modified 2nd order weights. The results are the average of 10,000 simulations.

| Mean | 1st order $w_j$ | | 2nd order $w_j$ | | Modified 2nd order $w_j$ | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Iterative | | Exact | |
| $F$ | $Q$ | T1E($Q$) | $Q$ | T1E($Q$) | $Q$ | T1E($Q$) | $Q$ | T1E($Q$) |
| **No heterogeneity, $\beta$=0** | | | | | | | | |
| 100 | 23.9 | 0.044 | 22.8 | 0.022 | 23.9 | 0.044 | 23.9 | 0.044 |
| 61 | 24.1 | 0.052 | 21.9 | 0.016 | 24.1 | 0.051 | 24.1 | 0.051 |
| 40 | 23.9 | 0.049 | 20.3 | 0.006 | 23.9 | 0.048 | 23.9 | 0.048 |
| 25 | 24.0 | 0.052 | 17.7 | 0.002 | 23.9 | 0.051 | 23.9 | 0.051 |
| 10 | 24.0 | 0.052 | 12.3 | 0.000 | 23.6 | 0.047 | 23.4 | 0.042 |
| **No heterogeneity, $\beta$=0.05** | | | | | | | | |
| 100 | 24.2 | 0.053 | 22.9 | 0.028 | 24.0 | 0.049 | 24.0 | 0.049 |
| 61 | 24.4 | 0.058 | 21.9 | 0.017 | 24.0 | 0.051 | 24.0 | 0.051 |
| 40 | 24.7 | 0.064 | 20.3 | 0.007 | 23.9 | 0.050 | 23.9 | 0.049 |
| 25 | 25.9 | 0.092 | 17.8 | 0.002 | 24.1 | 0.052 | 23.9 | 0.048 |
| 10 | 31.4 | 0.272 | 13.4 | 0.000 | 25.6 | 0.095 | 23.7 | 0.043 |
| **No heterogeneity, $\beta$=0.1** | | | | | | | | |
| 100 | 24.7 | 0.065 | 22.8 | 0.027 | 23.9 | 0.052 | 23.9 | 0.051 |
| 61 | 25.6 | 0.084 | 21.8 | 0.017 | 23.9 | 0.048 | 23.9 | 0.047 |
| 40 | 27.3 | 0.132 | 20.5 | 0.009 | 24.1 | 0.053 | 24.0 | 0.050 |
| 25 | 31.7 | 0.282 | 18.2 | 0.003 | 24.4 | 0.060 | 23.9 | 0.048 |
| 10 | 53.9 | 0.792 | 15.8 | 0.004 | 27.8 | 0.166 | 23.9 | 0.051 |

Table 1: *Mean Q statistic and type I error rate (T1E) of 1st order, 2nd order and modified 2nd order weights (implemented using the iterative approach and exact approaches). Results calculated over 10,000 simulated data sets. Type I error rate (T1E(Q)) refers to the proportion of times Q is greater than the upper 95th percentile of a $\chi^2_{24}$ distribution.*

Our modified weights were implemented using the simple iterative approach (four iterations were performed) and using the 'exact' approach previously described. Five different mean $F$-statistic values were considered for $\beta$=0 (no causal effect), $\beta$=0.05 and $\beta$=0.1, giving 15 scenarios in total. We note that, in the absence of a causal effect

($\beta$=0), 1st order weights are exactly correct. In the presence of a causal effect, when the mean $F$-statistic is 100 all weighting methods are near-exact. Under the causal null, all weighting schemes control the type I error rate for detecting heterogeneity. 2nd order weighting is extremely conservative in this respect with weak instruments, however (e.g. a type I error rate near zero when $F$=10).

In the presence of a causal effect, 1st order weights under-estimate the true variability amongst the ratio estimates as the mean $F$-statistic reduces. The associated $Q$ statistics are then too large on average (i.e. positively biased beyond their expected value of 24). This inflates the type I error rate for detecting pleiotropy beyond nominal levels (e.g. a type I error rate of $\approx$ 80% when $F$=10 and $\beta$=0.1). 2nd order weighting continues to over-correct the $Q$ statistic so that it is negatively biased, thereby removing *any* ability to detect heterogeneity at all. In contrast, modified 2nd order weights (applied iteratively) are much more effective at preserving the type I error rate of the $Q$ statistic at its nominal level, unless the mean $F$-statistic is very low (indicating weak instruments). When used within an exact analysis, modified 2nd order weighting perfectly controls the type I error rate of Cochran's $Q$ across all the scenarios considered.

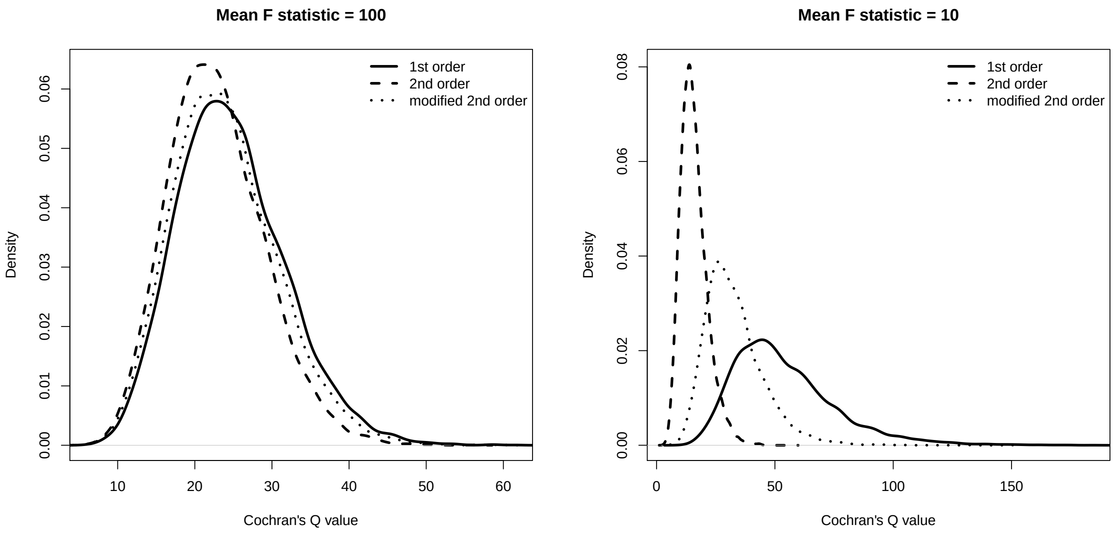

Figure 2: *Distribution of $Q$ statistics using 1st order, 2nd order and modified 2nd order weights (exact implementation) when $\beta$=0.1, and $F$ equals 100 (left) and 10 (right) respectively.*

Figure 2 (left and right) shows the distribution of $Q$ statistics using 1st order, 2nd order and modified 2nd order weights (exact implementation) when $\beta$=0.1 and when the mean $F$-statistic is 100 and 10. This illustrates how modified 2nd order weighting ensures Cochran's $Q$ statistic is faithful to its correct null distribution.

### Power to detect pleiotropy

In Table 1 the type I error rate of Cochran's $Q$ statistic for detecting heterogeneity using 2nd order weights was below its nominal level. This is detrimental if it translates into a low statistical power to detect heterogeneity when it *is* truly present. In order to investigate this, let $\alpha_j$ represent the pleiotropic effect of SNP $j$ on the outcome not via the exposure and let $\mu_\alpha$ and $\sigma^2_\alpha$ denote the sample mean and variance, respectively, of all $L$ pleiotropic effects. Suppose that the pleiotropic effects collectively satisfy the InSIDE assumption, and that the mean pleiotropic effect $\mu_\alpha = 0$. This is referred to as 'balanced' pleiotropy, and will induce heterogeneity amongst the ratio estimates. If heterogeneity is detected, inferences about the causal effect need to be adjusted to take this additional uncertainty into account, by assuming either by an additive random effects model [25] or a multiplicative random effects model [26]:

$$\begin{aligned} \text{Additive pleiotropy model: } \hat{\Gamma}_j &= \beta\gamma_j + \sqrt{\sigma^2_\alpha + \sigma^2_{Yj}}\epsilon_j & (11) \\ \text{Multiplicative pleiotropy model: } \hat{\Gamma}_j &= \beta\gamma_j + \sqrt{1 + \sigma^2_\alpha}\sigma_{Yj}\epsilon_j. & (12) \end{aligned}$$

In applied MR analyses, the multiplicative approach is much more common, because it is automatically implemented when the IVW estimate is obtained from fitting a regression model. Figure 3 (left) shows the power of Cochran's $Q$ to detect heterogeneity at the 5% significance level as a function of 1st order, 2nd order and modified 2nd order weights when data are simulated under a multiplicative random effects model (11) with balanced pleiotropy for increasing values of $\sqrt{1+\sigma^2_\alpha}$ between 1 and 2 (so that the value '1' indicates no heterogeneity). The simulation is repeated for MR analyses with $L = 10$, 25 and 100 SNPs. For all simulations, the causal effect equalled 0.05 and the mean $F$-statistic equalled 61. We see that the power of Cochran's $Q$ to detect heterogeneity approaches 100% for all weighting schemes as $\sigma_\alpha$ increases. Power also increases with the number of SNPs. The power of modified 2nd order weights is near identical using either the iterative or exact approach, therefore we only show results for the exact implementation for clarity. The most striking result in this plot is that the power of 2nd order weighting always lags considerably behind that of 1st order or modified 2nd order weights. Results for data simulated under an additive pleiotropy model are shown in Supplementary Online Material, and are highly similar.

Figure 3 (right) shows the results of a near identical simulation for the case $L$=25, except that the causal effect is set to 0.1 and the mean $F$-statistic is equal to 25. We see that the power to detect heterogeneity is always greatest when using 1st order weights, but only because its power curve starts at a baseline level of 28% when there is no pleiotropy. This corresponds to the type I error rate observed in row 14 of Table 1. The power of modified 2nd order weights starts at the correct 5% level, and rapidly increases to 100% as the pleiotropy variance increases. The two implementations of our modified weights can be differentiated in this simulation, with the iterative approach being slightly more powerful. The power of 2nd order weighting, unsurprisingly, lags considerably behind the rest.

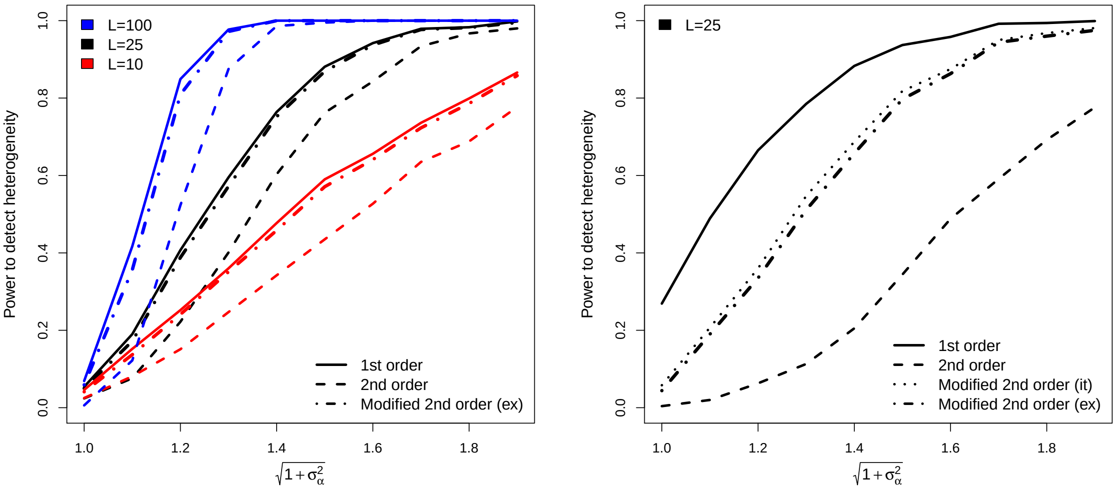

Figure 3: *Left: Power of Cochran's $Q$ statistic to detect heterogeneity as a function of the pleiotropy variance and number of SNPs ($L$) using 1st order, 2nd order and modified 2nd weights (ex=exact). Pleiotropy is simulated under a multiplicative random effects model. The causal effect is equal to 0.05 and the mean $F$-statistic is 61. Right: Equivalent power plot except the causal effect is equal to 0.1 and the mean $F$-statistic is 25 (ex=exact, it=iterative).*

## Detecting outliers using individual components of $Q$

When heterogeneity is detected by the IVW model, it is interesting to investigate whether this is contributed to by all SNPs, or if instead a small number of SNPs are responsible. Under the null hypothesis of no heterogeneity, $Q$ should follow a $\chi^2_{L-1}$ distribution. Likewise, each component of $Q$, $Q_j$, can be approximated by a $\chi^2_1$ distribution. If an individual SNP's $Q_j$ is extreme (for example above the 5% threshold of 3.84), then it may be desirable to exclude the SNP in a sensitivity analysis. Although we do not want to advocate a rigid, blanket policy of outlier removal, we illustrate how the reliability of such a procedure depends on the choice of weights. Motivated by the real data example in the following section, 10,000 summary associations are simulated for 25 SNPs for a range of mean $F$-statistics, a causal effect of 0.05 and under the assumption of no heterogeneity due to pleiotropy. That is, just as for rows 6-10 in Table 1. Each data set of 25 SNPs is then augmented with a single outlying SNP (with a fixed pleiotropic effect) which almost triples the magnitude of the observed heterogeneity across all 26 SNPs, as measured by Cochran's $Q$. Table 2 shows, for each weighting scheme: the mean $Q$ statistic, the median and mean number of 'outliers' detected at the 5% level and the proportion of times that the true outlier is detected ($P^*$) as $F$ is varied from 100 to 10. Figure 4 shows equivalent box plots of the outlier data, to highlight further summary quantities such as the inter quartile range.

We would expect approximately 25×0.05 = 1.25 of the normal, non-heterogeneous SNPs to be declared outliers by chance at the 5% significance level, and hope that the true outlier is detected as often as possible, giving an ideal mean total of 2.25. As the mean $F$-statistic decreases, the total number of outliers detected using 1st order weights steadily increases beyond this value (although the probability of decting the true outlier stays constant at $\approx$ 95%). By contrast, the total number of outliers detected using 2nd order weights substantially decreases, as well as the ability to detect the true outlier. For example, when $F$ is 10, the true outlier is detected in less than 30% of cases. The performance of modified 2nd order weights is much more stable across the range of instrument strengths, with the median and mean number of outliers never increasing beyond 2 and 3 respectively. However, in this case it is the iterative rather than the exact weights that appear to perform best. For example, when the mean $F$-statistic is 10 the power to detect the true outlier drops to only 87% using the exact approach, but stays at 94% for the iterative approach. Moreover, the box plots in Figure 4 show that the number of outliers detected across the simulations is much more variable for the exact, compared to the iterative implementation.

| Mean | | 1st order $w_j$ | | 2nd order $w_j$ | | Modified 2nd order $w_j$ | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Iterative | | Exact |
| | | 'Outliers' detected | | 'Outliers' detected | | 'Outliers' detected | | 'Outliers' detected |
| $F$ | Q | (Median,Mean,$P^*$) | Q | (Median,Mean,$P^*$) | Q | (Median,Mean,$P^*$) | Q | (Median,Mean,$P^*$) |
| | | | | **No heterogeneity for 25 SNPs + 1 outlier, $\beta$=0.05** | | | | |
| 100 | 67.6 | (2,2.54,0.94) | 42.3 | (2,1.98,0.93) | 64.1 | (2,2.44,0.94) | 63.4 | (2,2.69,0.94) |
| 61 | 68.5 | (2,2.55,0.94) | 37.0 | (2,1.78,0.90) | 62.8 | (2,2.38,0.94) | 61.4 | (2,2.77,0.94) |
| 40 | 69.6 | (2,2.58,0.94) | 31.6 | (1,1.50,0.81) | 60.4 | (2,2.32,0.94) | 57.9 | (2,2.82,0.94) |
| 24 | 71.2 | (2,2.71,0.95) | 25.8 | (1,1.10,0.62) | 56.8 | (2,2.25,0.94) | 52.6 | (2,2.82,0.94) |
| 10 | 80.0 | (3,3.27,0.95) | 17.5 | (0,0.53,0.28) | 53.2 | (2,2.23,0.94) | 41.1 | (2,2.58,0.87) |

Table 2: *The number of outliers detected at the 5% level by Cochran's Q statistic when using 1st order, 2nd order and modified 2nd orders weights for MR summary data containing 25 non-heterogeneous SNPs and 1 outlier.*

## Estimator performance with and without pleiotropy

Table 3 shows the performance of the 1st order, 2nd order and iterative modified 2nd order weighting in providing accurate point estimates, standard errors and confidence intervals for the causal effect. Only the mean causal estimate obtained via exact modified 2nd order weighting is shown in the last column of Table 3. This is because no simple, general and reliable formula for the variance of this estimate could be found. Rows 1-15 are for data simulated without heterogeneity due to pleiotropy, identical to that described in Table 1. In this case, all three of methods return unbiased estimates when $\beta$ = 0. Correct coverages are also observed at the causal null, with the exception of 2nd order weighting when $F$=10. In the presence of a non-zero causal effect, a decreasing mean $F$-statistic leads to increased bias in the IVW estimate and a drop in confidence interval coverage for all approaches, which can be equated with their statistical power.

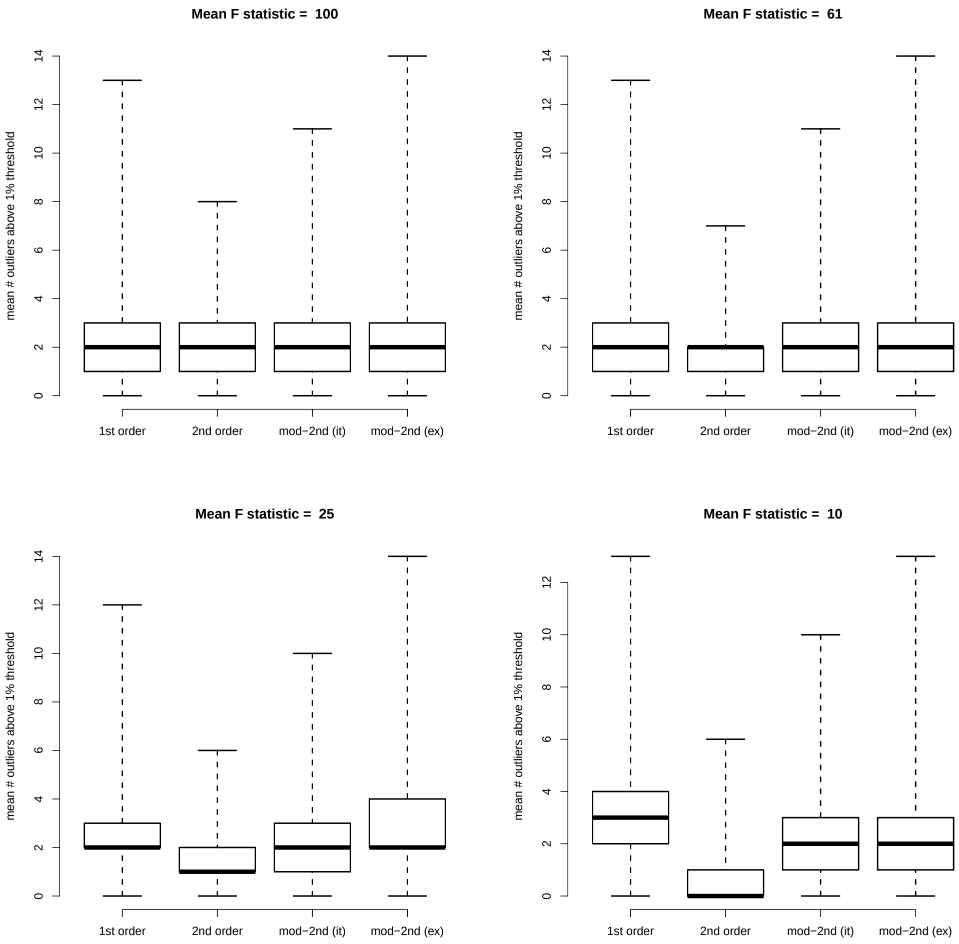

Figure 4: *Box plots summarising the total number of outliers detected by Cochran's Q statistic using 1st order, 2nd order and modified 2nd order weights (iterative 'it' and exact 'ex' implementations, respectively) when the mean $F$-statistic is varied betwen 100 (top-left) and 10 (bottom-right). Each box shows the 1st quartile, median line and 3rd quartile, so that its height represent the inter quartile range. Box 'whiskers' representing the full outlier range are also shown.*

Iterative modified 2nd order weighting is least affected, however. The IVW estimate is known to suffer from regression dilution bias towards zero by an amount approximately proportional to the inverse of the mean $F$-statistic. This dilution can be mitigated by applying bias adjustment techniques from the measurement error literature, such as simulation extrapolation [3, 19, 20]. For brevity, we do not additionally assess SIMEX correction here.

Rows 16-25 of Table 3 show results for data simulated under multiplicative pleiotropy model (12) with $\sqrt{1+\sigma_\alpha^2} = 1.4$. This gives rise to data with associated $Q$ statistics of 50 - double the expected magnitude under the null hypothesis of no pleiotropy.

| Mean | 1st order $w_j$ | 2nd order $w_j$ | Modified 2nd order $w_j$ | |
|---|---|---|---|---|
| | | | Iterative | Exact |
| $F$ | $\hat{\beta}_{IVW}$(SE); CF | $\hat{\beta}_{IVW}$(SE); CF | $\hat{\beta}_{IVW}$(SE); CF | $\hat{\beta}_{IVW}$ |
| **No heterogeneity, $\beta$=0** | | | | |
| 100 | 0.000 (0.011) 0.950 | 0.000 (0.011) 0.948 | 0.000 (0.011) 0.950 | 0.000 |
| 61 | 0.000 (0.011) 0.951 | 0.000 (0.011) 0.951 | 0.000 (0.011) 0.952 | 0.000 |
| 40 | 0.000 (0.011) 0.948 | 0.000 (0.010) 0.946 | 0.000 (0.011) 0.950 | 0.000 |
| 25 | 0.000 (0.011) 0.949 | 0.000 (0.009) 0.942 | 0.000 (0.011) 0.951 | 0.000 |
| 10 | 0.000 (0.009) 0.948 | 0.000 (0.007) 0.928 | 0.000 (0.009) 0.955 | 0.000 |
| **No heterogeneity, $\beta$=0.05** | | | | |
| 100 | 0.050 (0.011) 0.949 | 0.049 (0.011) 0.951 | 0.050 (0.011) 0.951 | 0.050 |
| 61 | 0.049 (0.011) 0.946 | 0.047 (0.011) 0.943 | 0.049 (0.011) 0.949 | 0.050 |
| 40 | 0.048 (0.011) 0.945 | 0.045 (0.011) 0.924 | 0.048 (0.011) 0.950 | 0.050 |
| 25 | 0.046 (0.011) 0.914 | 0.041 (0.010) 0.825 | 0.046 (0.012) 0.926 | 0.051 |
| 10 | 0.032 (0.010) 0.580 | 0.027 (0.008) 0.275 | 0.034 (0.011) 0.668 | 0.051 |
| **No heterogeneity, $\beta$=0.1** | | | | |
| 100 | 0.099 (0.011) 0.944 | 0.098 (0.011) 0.944 | 0.099 (0.012) 0.950 | 0.100 |
| 62 | 0.098 (0.011) 0.938 | 0.095 (0.011) 0.922 | 0.098 (0.012) 0.948 | 0.100 |
| 40 | 0.096 (0.012) 0.912 | 0.091 (0.011) 0.858 | 0.096 (0.012) 0.935 | 0.100 |
| 25 | 0.091 (0.012) 0.842 | 0.082 (0.011) 0.643 | 0.092 (0.013) 0.897 | 0.100 |
| 10 | 0.065 (0.013) 0.341 | 0.055 (0.010) 0.093 | 0.072 (0.015) 0.516 | 0.102 |
| **Heterogeneity, $\beta$=0** | | | | |
| 100 | 0.000 (0.015) 0.950 | 0.000 (0.015) 0.948 | 0.000 (0.015) 0.950 | 0.000 |
| 61 | 0.000 (0.015) 0.951 | 0.000 (0.015) 0.952 | 0.000 (0.015) 0.952 | 0.000 |
| 40 | 0.000 (0.015) 0.948 | 0.000 (0.014) 0.946 | 0.000 (0.015) 0.950 | 0.000 |
| 25 | 0.000 (0.015) 0.949 | 0.000 (0.013) 0.942 | 0.000 (0.015) 0.953 | 0.000 |
| 10 | 0.000 (0.013) 0.948 | 0.000 (0.009) 0.927 | 0.000 (0.013) 0.959 | 0.000 |
| **Heterogeneity, $\beta$=0.1** | | | | |
| 100 | 0.100 (0.016) 0.947 | 0.096 (0.016) 0.944 | 0.100 (0.016) 0.951 | 0.101 |
| 62 | 0.098 (0.016) 0.944 | 0.092 (0.015) 0.922 | 0.098 (0.016) 0.954 | 0.100 |
| 40 | 0.096 (0.016) 0.930 | 0.087 (0.015) 0.858 | 0.097 (0.017) 0.947 | 0.100 |
| 25 | 0.091 (0.016) 0.885 | 0.078 (0.015) 0.662 | 0.092 (0.018) 0.923 | 0.101 |
| 10 | 0.065 (0.016) 0.445 | 0.051 (0.012) 0.124 | 0.072 (0.018) 0.631 | 0.099 |

Table 3: *Mean causal estimate $\hat{\beta}_{IVW}$, standard error (SE) and coverage frequency (CF) of the 95% confidence interval when using 1st order, 2nd order and modified 2nd order weights.*

All three methods perform essentially the same as with non-heterogeneous data, except that their confidence interval coverages are improved when the mean $F$-statistic is low. The final column of Table 3 shows the mean IVW estimate obtained from the exact implementation of modified 2nd order weights. When calculating this estimate for heterogeneous data, a slight modification of the weighting procedure is necessary. Specifically we use the following scheme:

- Calculate Cochran's $Q$ statistic using 1st-iteration modified 2nd order weights under no heterogeneity. That is, use equation (8) to define $Q$, where $\hat{\beta}_{IVW}$ is the estimate obtained using 1st order weights;

- Calculate the quantity $\hat{\phi} = Q/(L\text{-}1)$, $\hat{\phi}$ being an estimate for $1 + \sigma_\alpha^2$;
- Replace the original weights in equation (9) with

$$w_j(\hat{\beta}_{IVW}) = \left(\frac{\hat{\phi}\sigma_{Yj}^2 + \hat{\beta}_{IVW}^2\sigma_{Xj}^2}{\hat{\gamma}_j^2}\right)^{-1} \tag{13}$$

and minimise to obtain a new heterogeneity-adjusted IVW estimate.

Table 3 demonstrates that the IVW estimates obtained from the exact implementation of modified 2nd order weighting are essentially free from regression dilution bias. Promising preliminary work to obtain a reliable variance formula for this estimator is underway [6], which we describe further in the discussion.

# Applied example

Figure 5 (top) shows a scatter plot of summary data estimates for the associations of 26 genetic variants with systolic blood pressure (SBP, the exposure) and coronary heart disease (CHD, the outcome). SNP-exposure association estimates were obtained from the International Consortium for Blood Pressure consortium (ICBP) [27]. SNP-CHD association odds ratios were collected from Coronary ARtery Disease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) consortium [28], which are plotted (and subsequently modelled) on the log odds ratio scale by making a normal approximation, as discussed in the introduction. These data have previously been used in a two-sample summary data MR analysis by Ference et al. [23] and Lawlor et al. [15], but we extend their original analysis here by applying modified 2nd order weights and conducting a more in depth inspection of each variant's contribution to the overall heterogeneity. The mean $F$-statistic for these data is 61. Using 1st order weights the IVW estimate, which represents the causal effect of a 1mmHg increase in SBP on the log-odds ratio of CHD, is 0.053. This is shown as the slope of a solid black line passing through the origin (note: the origin is not visible because of a truncated x-axis). Cochran's $Q$ statistic based on 1st order weights is equal to 67.1, indicating the presence of substantial heterogeneity.

Table 4 shows the results of further IVW analyses using all three weighting schemes. All three schemes detect significant heterogeneity. As expected, the observed heterogeneity is largest when using 1st order weights, smallest when using 2nd order weights, and in between the two when using modified 2nd order weights. Point estimates and standard errors are in good agreement across the different weights, because the mean instrument strength is high. Modified 2nd order weighting gives the largest point estimate of 0.054, followed by 1st order and then 2nd order weights respectively. This ordering is as expected, given their relative susceptibility to regression dilution bias. For comparison, we apply the Simulation Extrapolation (SIMEX) to adjust the IVW estimate for regression dilution. Reassuringly, its estimate is identical to that obtained using the exact implementation of modified 2nd order weights.

Finally, we calculate the Weighted Median MR [29], $\hat{\beta}_{WM}$, that can identify the causal effect when up to (but not including) half of the information in the analysis stems from genetic variants that are invalid IVs. Its estimate, which is calculated using 1st order weights, is 0.063.

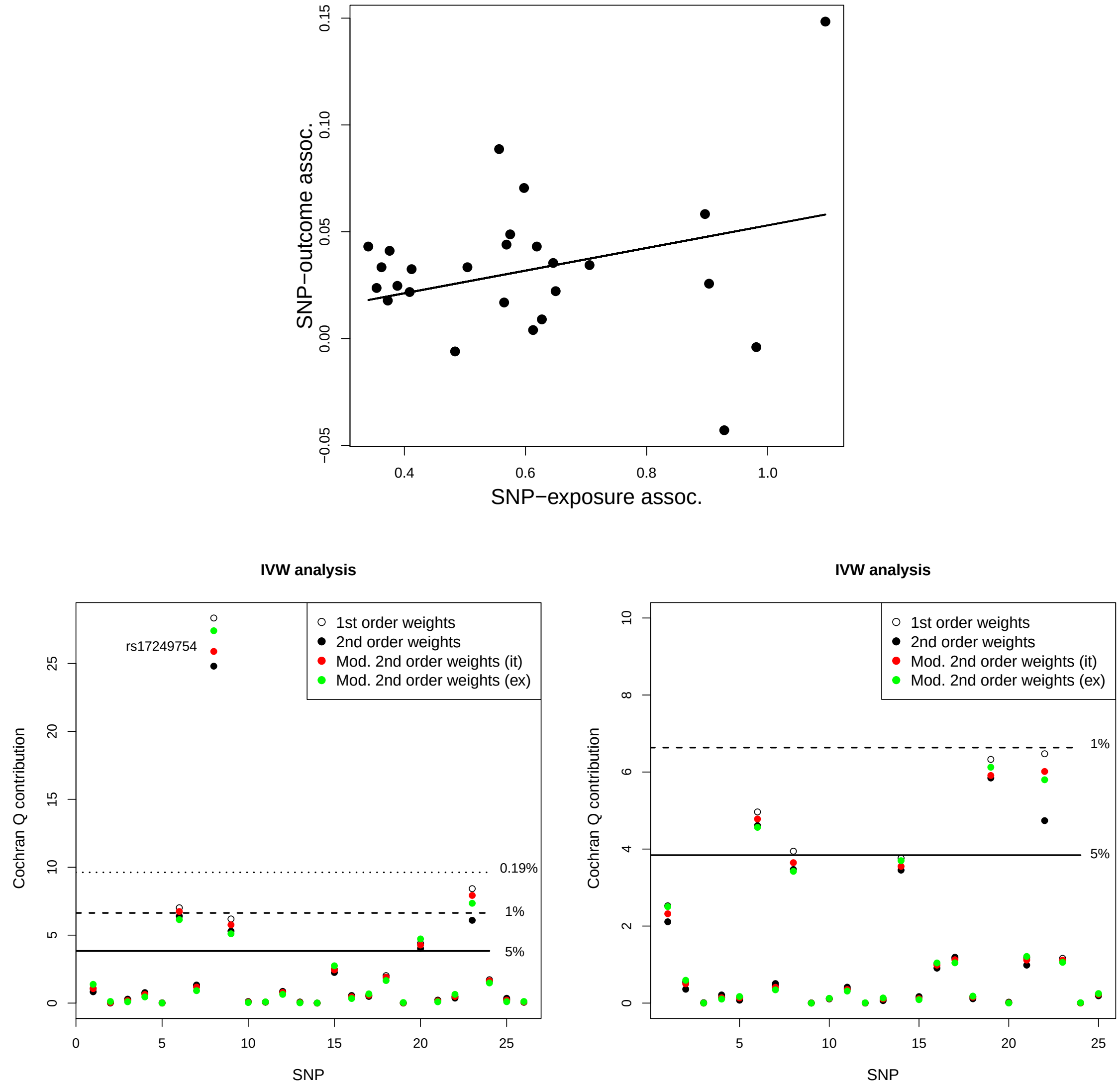

Figure 5: *Top: Scatter plot of SNP-outcome associations* $\hat{\Gamma}_j$ *versus SNP-exposure associations* $\hat{\gamma}_j$. *IVW estimate shown as a black slope. Bottom-left: Q contribution plots for the same data. Bottom-right: Q contributions after removal of rs17249754.*

Figure 5 (bottom-left) shows the individual contribution to $Q_j$ under each weighting scheme. Horizontal lines have been drawn to indicate the location of the 5th, 1st and 0.19th percentile of a $\chi^2_1$ in order to help assess the magnitude of the contributions. The 0.19th percentile is derived as a 0.05 threshold adjusted for multiple testing using the Bonferroni correction. We see that the eighth SNP in our list (rs17249754) is responsible for the vast majority of the excess heterogeneity. Its contribution, $Q_8$, ranges from approximately 24.5 to 28 depending on weighting.

| Method (weights) | Estimate | S.E. | P-value | Het. Stat (p) |
|---|---|---|---|---|
| **All 26 SNPs** | | | | |
| **Causal estimate** | | | | |
| IVW (1st) | $\hat{\beta}_{IVW}$: 0.053 | 0.010 | $3.01\times10^{-5}$ | $Q = 67.1$ ($1.03\times10^{-5}$) |
| IVW (2nd) | $\hat{\beta}_{IVW}$: 0.050 | 0.010 | $4.54\times10^{-5}$ | $Q = 58.8$ ($1.54\times10^{-4}$) |
| IVW (Mod 2nd, iterative) | $\hat{\beta}_{IVW}$: 0.054 | 0.010 | $2.40\times10^{-5}$ | $Q = 62.7$ ($4.43\times10^{-5}$) |
| IVW (Mod 2nd, exact) | $\hat{\beta}_{IVW}$: 0.054 | | | $Q = 62.4$ ($4.84\times10^{-5}$) |
| **Weighted median (1st order weights)** | | | | |
| Weighted Median | $\hat{\beta}_{WM}$: 0.063 | 0.011 | $4.90\times10^{-6}$ | - |
| **SIMEX adjusted IVW estimate (1st order weights)** | | | | |
| IVW (1st) | $\hat{\beta}_{IVW}$: 0.054 | 0.011 | $3.9\times10^{-5}$ | - |
| **SNP rs17249754 removed** | | | | |
| **Causal estimate** | | | | |
| IVW (1st) | $\hat{\beta}_{IVW}$: 0.066 | 0.008 | $2.63\times10^{-8}$ | $Q = 35.0$ (0.068) |
| IVW (2nd) | $\hat{\beta}_{IVW}$: 0.063 | 0.008 | $4.06\times10^{-8}$ | $Q = 30.6$ (0.164) |
| IVW (Mod 2nd, iterative) | $\hat{\beta}_{IVW}$: 0.066 | 0.008 | $2.90\times10^{-8}$ | $Q = 32.8$ (0.108) |
| IVW (Mod 2nd, exact) | $\hat{\beta}_{IVW}$: 0.067 | | | $Q = 32.8$ (0.108) |
| **Weighted median (1st order weights)** | | | | |
| Weighted Median | $\hat{\beta}_{WM}$: 0.065 | 0.010 | $1.81\times10^{-6}$ | - |
| **SIMEX adjusted IVW estimate (1st order weights)** | | | | |
| IVW (1st) | $\hat{\beta}_{IVW}$: 0.067 | 0.008 | $2.35\times10^{-8}$ | - |

Table 4: *IVW and Weighted Median analyses of the causal effect of SBP on CHD risk using 1st order, 2nd order and modified 2nd order weights for the complete data (top) and with SNP rs17249754 removed (bottom). $\hat{\beta}_{IVW}$ is the IVW estimate. $\hat{\beta}_{WM}$ is the Weighted Median estimate. SIMEX refers to estimates obtained by the method of simulation extrapolation.*

Variant rs17249754 sits in the ATPase plasma membrane Ca2+ transporting 1 (*ATP2B1*) gene, which is involved in intracellular calcium homeostasis, and is strongly associated with higher SBP. However, in the CARDIoGRAM consortium it is associated with reduced CHD. It could be that rs17249754 truly increases SBP in the ICBP population but decreases it in CARDIoGRAM, which would be a violation of the monotonicity assumption. Alternatively, rs17249754 could be exerting a pleiotropic effect on CHD not through SBP in a consistent manner for both the ICBP and CARDIoGRAM study populations, which is then reflected in the CARDIoGRAM estimate. As previously discussed, incorporating odds ratios into an MR analysis can lead to heterogeneity amongst causal estimates. However, this could only ever do so by shrinking estimates towards zero, not changing their sign [5]. We can therefore rule out this explanation here.

Since rs17249754 is also a strong instrument, and is potentially pleiotropic, its presence in the data could lead to the InSIDE assumption being violated. We therefore opt to remove it in a further sensitivity analysis, and Table 4 show the results. All IVW estimates increase by around 20% (lying between 0.063 and 0.067), but are ordered as before. We apply SIMEX to the IVW estimate using 1st order weights, and again observe that this agrees with the estimate obtained using exact modified 2nd order weights (0.067). The weighted median estimate without rs17249754 is 0.065 (compared to 0.063 with). This highlights its inherent robustness to outliers, which is a major strength.

Figure 5 (bottom-right) shows the updated contributions of each SNP to the $Q$ statistic after removing rs17249754. If only 1st order weighting were available, it might be tempting to exclude further variants from the analysis, but this signal is appropriately tempered when using modified 2nd order weights.

# Discussion

In this paper we have demonstrated the limitations of 1st and 2nd order weighting when used for IVW analysis in two-sample summary data Mendelian randomization. Most importantly, we highlight the potential for serious type I error inflation of Cochran's $Q$ statistic when using standard 1st order weights with weak instruments. In recent work, Verbank et al. [30] also noted this same tendency and proposed a simulation-based alternative to 1st order weighting named 'MR-PRESSO'. Our simulations show that modified 2nd order weights can deliver much more accurate causal estimates and reliable tests for heterogeneity than either 1st or 2nd order weighting, and suggests that the computationally intensive MR-PRESSO approach is unnecessary.

Modified 2nd order weights should also prove a more reliable tool for the detection and removal of outliers in a given data set, as apposed to 1st order weights (which may detect too many outliers) and 2nd order weights (that may detect too few). Our simulations suggest that the exact implementation of modified 2nd order weights should be used when testing for the overall presence of heterogeneity (referred to as the 'global' test by Verbank et al. [30]). However, they also suggest that the iterative implementation is preferable if looking at the individual contribution of each SNP to the $Q$ statistic to decide on its status as an outlier. We suspect this is because exact weighting makes a more aggressive correction for regression dilution than iterative weighting. Its resulting estimate then makes more variants appear as outliers, because their ratio estimates are further away from it. In this paper we used heterogeneity statistics for outlier detection, but many other test statistics such as Cook's distance have been used for this purpose in MR (see for example [31]). Modified 2nd order weights are likely to improve their performance too, but again this requires further investigation.

When implementing the Weighted Median estimate, 1st order weights were used. Another pleiotropy robust MR method - the Mode-based estimate [32] - also makes use of 1st order weights. As future work we will investigate whether modified 2nd order weights can improve the performance of both approaches, modifying their precise form in each case if necessary.

An exciting finding of this paper is that the exact implementation of modified 2nd order weights yields causal estimates that are remarkably robust to regression dilution bias. This is perhaps not surprising, given its connection to LIML. In the example analysis we showed that this approach, (a simple analytic formula) gave estimates in very close agreement to those obtained from applying SIMEX (a more complicated simulation-based method). Unfortunately, we were not able to derive a general expression for the variance of our analytic formula. In preliminary work, Zhao et al. [6] conduct a thorough theoretical investigation of the estimate obtained from exact modified 2nd order weights, and show that it can also be viewed as a profile likelihood maximisation problem. Their work suggests a variance formula can be derived under an additive random effects model. In the future we hope that this approach will be properly validated and extended to the multiplicative random effects model used predominantly in applied MR studies, and by ourselves in this paper.

A further interesting consequence of the apparent robustness of modified 2nd order weighting to weak instruments is that it opens up the potential for the significance threshold used to select SNPs as instruments to be substantially dropped, thus yielding many more instruments for a given trait, whilst at the same time improving the accuracy of MR estimates. This issue is also explored in further detail by Zhao et al. [6].

## Limitations

Our conclusions regarding the use of modified 2nd order weights are limited to the two-sample summary setting where SNP-outcome and SNP-exposure associations are estimated in independent but homogeneous samples. Further research would be required to decide if modified 2nd order weights should be used in MR analyses of summary data estimates when there is partial overlap between samples, or in the single sample (total overlap) setting.

When Cochran's $Q$ statistic detects significant amounts of heterogeneity, it is prudent to test whether it is meaningfully biasing the analysis. This would indeed be the case if the heterogeneity were caused in part by directional pleiotropy with a non-zero mean. This would lead to bias in the IVW estimte, unless of course it was caused by a small number of SNPs that could be identified and removed from the analysis. MR-Egger regression [3, 11] could instead be used to address this. This approach simply regresses SNP-outcome associations on the SNP-exposure associations, tests for bias via its intercept, and estimates a bias-adjusted causal effect via its slope. Observed heterogeneity around the MR-Egger fit can then be quantified using Rücker's

$Q'$ statistic [3, 33] and each variant's contribution to Rücker's $Q'$ statistic can be used as the basis for outlier detection. Currently MR-Egger and Rücker's $Q'$ statistic use 1st order weights. Preliminary work suggests that modified 2nd order weighting can be applied to MR-Egger regression to improve its performance - both in terms of causal effect estimation and heterogeneity quantification - just as for an IVW analysis, but further development and validation of this method is required.

R code is provided in Supplementary On line Methods to implement modified 2nd order weighting within an IVW analysis

**Key messages:**

- Two-sample summary data Mendelian randomization requires the specification of inverse variance weights for model fitting, heterogeneity quantification and outlier detection amongst a set of causal estimates.
- Heterogeneity indicates a possible violation of the necessary IV or modelling assumptions.
- 1st order weights can inflate the type I error rate of Cochran's $Q$ statistic for detecting heterogeneity about the IVW estimate when the NOME assumption is strongly violated (as judged by a low $F$-statistic), and the true causal effect of interest is non-zero.
- 2nd order weights can reduce the power of Cochran's $Q$ statistic for detecting heterogeneity about the IVW estimate when the NOME assumption is violated.
- Modified 2nd order weights (developed in this paper) preserve the type I error rate of Cochran's $Q$ statistic, whilst maintaining its statistical power, and naturally correct for regression dilution bias due to NOME violation.
- 'Exact' modified 2nd order weights should be used for global tests of heterogeneity. 'Iterative' modified 2nd order weights should be used to assess the outlier status of individual SNPs.

## Acknowledgements

We thank the reviewers and associate editor for providing helpful comments and suggestions that greatly improved this paper. We are particularly grateful to a reviewer for pointing out the connections in our work to the econometric literature. Jack Bowden is supported by an MRC Methodology Research Fellowship (grant MR/N501906/1). Jack Bowden, Debbie A Lawlor and George Davey Smith all work in a Unit that receives support from the University of Bristol and UK Medical Research Council (MCUU12013/1 and MCUU12013/5). The authors declare no conflict of interest.

# References

[1] Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* 2003; **32**:1–22.

[2] Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology* 2013; **37**:658–665.

[3] Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization *Statistics in Medicine* **36**: 1783–1802

[4] Zhao Q, Wang J, Bowden J, Small D. Two-sample instrumental variable analyses using heterogeneous samples. Technical report, University of Pennsylvania 2017. `https://arxiv.org/abs/1709.00081`

[5] Vansteelandt S, Bowden J, Babanezhad M, Goetghebeur E. On instrumental variables estimation of causal odds ratios. *Statistical Science* 2011; **26**:403–422.

[6] Zhao Q, Wang J, Hemani G, Bowden J, Small D. Statistical inference in two-sample summary data Mendelian randomization using a robust adjusted profile score. Technical report, University of Pennsylvania 2018. `http://www-stat.wharton.upenn.edu/~qyzhao/papers/mr_raps.pdf`

[7] Swanson SA, Hernan MA. The challenging interpretation of instrumental variable estimates under monotonicity *International Journal of Epidemiology* 2017. DOI: https://doi.org/10.1093/ije/dyx038

[8] Del Greco M F, Minelli C, Sheehan NA, Thompson JR. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Statistics in Medicine* 2015; **34**:2926–2940.

[9] Davey Smith G, Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics* 2014, **23**: 89–98.

[10] Stearns FW. One hundred years of pleiotropy: a retrospective. *Genetics* 2010; **186**:767–773.

[11] Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* 2015; **44**:512–525.

[12] Borges MC, Lawlor DA, de Oliveira C, White J, Horta BL, Barros AJD. The Role of Adiponectin in Coronary Heart Disease Risk: A Mendelian Randomization Study. *Circulation Research* 2016; **119**:491–499

[13] White J, Sofat R, Hemani G, Shah T, Engmann J, Dale C, Shah S et al. Plasma urate concentration and risk of coronary heart disease: a Mendelian randomisation analysis. *The Lancet Diabetes & Endocrinology* 2016; **4**:327–336.

[14] Noyce AJ, Kia DA, Hemani G, Nicolas A, Price TR, De Pablo-Fernandez E, Haycock PC, Lewis PA, Foltynie T, Davey Smith G, IPDGC collaborators, Schrag A, Lees AJ, Hardy J, Singleton A, Nalls MA, Pearce N, Lawlor DA, Wood NW. Estimating the causal influence of BMI on risk of Parkinson's disease: a Mendelian randomization study. *PLoS Medicine* 2017; **14**: e1002314

[15] Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *International Journal of Epidemiology* 2016 **45**:1866–86.

[16] Haycock PC, Burgess S, Nounu A, Zheng J, Okoli GN, Bowden J, Wade KH, Timpson NJ, Evans DM, Willeit P, Aviv A. Association between telomere length and risk of cancer and non-neoplastic diseases: a Mendelian randomization study. *JAMA oncology* 2017; **3**: 636–651.

[17] Burgess S, Bowden J. Integrating summarized data from multiple genetic variants in Mendelian randomization: bias and coverage properties of inverse-variance weighted methods Technical report, University of Cambridge 2015. `https://arxiv.org/pdf/1512.04486.pdf`

[18] Thomas D, Lawlor, D, Thompson J. Re: Estimation of Bias in Nongenetic Observational Studies Using Mendelian Triangulation by Bautista et al. *Annals of Epidemiology* 2007; **17**: 511–513.

[19] Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for Mendelian randomization analyses using MR-Egger regression: the role of the $I^2$ statistic. *International Journal of Epidemiology* 2016; **45**: 1961–74.

[20] Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* 1994; **89**:1314–1328.

[21] Thompson J, Minelli C, Del Greco M F. Mendelian randomization using public data from genetic consortia. *International Journal of Biostatistics* 2016 **12**:2

[22] Windmeijer, F. Two-Stage Least Squares as Minimum Distance Working paper, department of economics, University of Bristol, 2017. Available at: `http://www.efm.bris.ac.uk/economics/working_papers/pdffiles/dp17683.pdf`

[23] Ference BA, Julius S, Mahajan N, Levy PD, Williams KASr, Flack JM. Clinical effect of naturally random allocation to lower systolic blood pressure beginning before the development of hypertension. *Hypertension* 2014; **63**: 1182-88.

[24] Cumby, RE, Huizinga J, Obstfeld M. Two-step two-stage least squares estimation in models with rational expectations. *Journal of Econometrics* 1982; **21**: 333-355.

[25] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.

[26] Thompson SG, Sharp S. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* 1999; **18**:2693–2708.

[27] International Consortium for Blood Pressure Genome-Wide Association Studies. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 2011; **478**:103-9.

[28] CARDIoGRAMplusC4D. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nature Genetics* 2003; **45**:25–33.

[29] Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology* 2016; **40**: 304–14.

[30] Verbanck M, Chen CY, Neale B, Do R. Widespread pleiotropy confounds causal relationships between complex traits and diseases inferred from Mendelian randomization. bioRxiv. 2017 `http://www.biorxiv.org/content/biorxiv/early/2017/06/30/157552.full.pdf`.

[31] Corbin LJ, Richmond RC, Wade KH, Burgess S, Bowden J, Davey Smith G, Timpson NJ. Body mass index as a modifiable risk factor for type 2 diabetes: Refining and understanding causal estimates using Mendelian randomisation. *Diabetes* 2016 db160418.

[32] Hartwig FP, Davey Smith G, Bowden J. Robust inference in two-sample Mendelian randomisation via the zero modal pleiotropy assumption. *International Journal of Epidemiology* 2017; **46**:1985–1998

[33] Rücker G, Schwarzer G, Carpenter J, Binder H, Schumacher M. Treatment-effect estimates adjusted for small study effects via a limit meta-analysis. *Biostatistics* 2011; **12**:122–142.

# On-line supplementary material

## R code

```
# Load in data set ‘data’ and vector of
# SNP-exposure associations BetaXG and S.Es seBetaXG
# SNP-outcome associations BetaYG and S.Es seBetaYG

BetaXG   = data$BetaXG
BetaYG   = data$BetaYG
seBetaYG = data$seBetaYG
seBetaXG = data$seBetaXG
```

```
# mean F statistic

F   = BetaXG^2/seBetaXG^2

# Degree of freedom for Cochran's Q

DF  = length(BetaYG)-1

# IVW analysis
# 1st order and 2nd order weights

BIV       = BetaYG/BetaXG
W1        = 1/(seBetaYG^2/BetaXG^2)
BIVw1     = BIV*sqrt(W1)
W2        = 1/(seBetaYG^2/BetaXG^2 +
            (BetaYG^2)*seBetaXG^2/BetaXG^4)
BIVw2     = BIV*sqrt(W2)
sW2       = sqrt(W2)
sW1       = sqrt(W1)

IVWfitR1  = summary(lm(BIVw1 ~ -1+sW1)) # 1st order
IVWfitR2  = summary(lm(BIVw2 ~ -1+sW2)) # 2nd order


##########################################
# Modified 2nd order weights (iterative) #
##########################################

Bhat1      = IVWfitR1$coef[1] # initialise
Bhat2      = IVWfitR2$coef[1]
Bhat3      = IVWfitR1$coef[1]

for(gg in 1:4){
W3        = 1/(seBetaYG^2/BetaXG^2 + (Bhat1^2)*seBetaXG^2/BetaXG^2)
BIVw3     = BIV*sqrt(W3)
sW3       = sqrt(W3)
IVWfitR3 = summary(lm(BIVw3 ~ -1+sW3))
Bhat3     = IVWfitR3$coef[1]
phi_IVW3 = IVWfitR3$sigma^2

if(gg==1){phi_IVW31  = IVWfitR3$sigma^2}

print(Bhat3)
print(phi_IVW3)

}

# General inference #

IVWfitR3

################################################
# In order to stop the analysis adjusting      #
# for under-dispersion - making the estimates  #
# more precise than under a standard           #
# fixed effects meta-analysis, the following   #
# code can be used                             #
################################################

SE_IVWR3 = IVW2FitR3$coef[1,2]/min(IVW2FitR3$sigma, 1).

# Global test for heterogeneity

phi_IVW3  = IVWfitR3$sigma^2
QIVW3     = DF*phi_IVW3                        # Cochran’s Q
Qp3       = pchisq(QE3, DF, lower.tail=FALSE)  # Cochran’s Q p-value

# Individual contribution to Q

Q3ind     = W3*(BIV - Bhat3)^2

######################################
# Modified 2nd order weights (Exact) #
######################################

# Full maximisation

PL = function(a){
b = a[1]
w = 1/(seBetaYG^2/BetaXG^2 + (b^2)*seBetaXG^2/BetaXG^2)
q =  sum(w*(BIV - b)^2)
}

Bhat4     = optimize(PL,interval=c(-2,2))$minimum  # (change interval if needed)

# Global test for heterogeneity

W4        = 1/(seBetaYG^2/BetaXG^2 + (Bhat4^2)*seBetaXG^2/BetaXG^2)
QIVW4     = sum(W4*(BIV - Bhat4)^2)
Qp4       = pchisq(QIVW4, DF, lower.tail=FALSE)

# Individual contribution to Q

Q4ind     = W4*(BIV - Bhat4)^2

# Exact IVW estimate (accounting for heterogeneity)

PL2 = function(a){
b   = a[1]
PHI = max(phi_IVW31,1)
w   = 1/(PHI*seBetaYG^2/BetaXG^2 + (b^2)*seBetaXG^2/BetaXG^2)
q   =  sum(w*(BIV - b)^2)
}

Bhat5     = optimize(PL2,interval=c(-2,2))$minimum
```

# Additional Figures

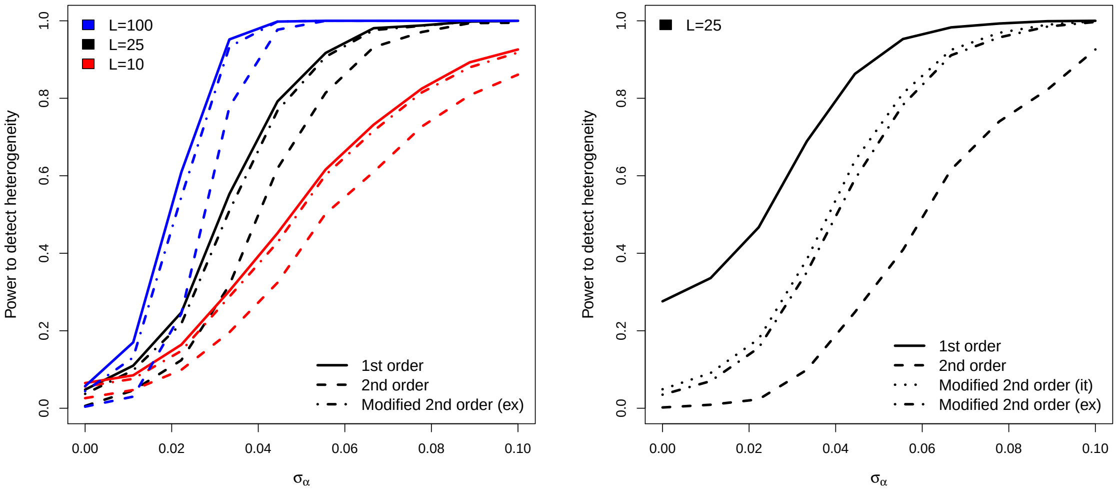

Figure 6: *Left: Power of Cochran's Q statistic to detect heterogeneity as a function of the pleiotropy standard deviation ($\sigma_\alpha$) and number of SNPs (L) using 1st order, 2nd order and modified 2nd weights under an additive pleiotropy model.*