Statistics in Medicine - Accepted paper

Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome

Fabiola Del Greco M[a], Cosetta Minelli[b], Nuala A Sheehan[c], John R Thompson[c]

[a] *Center for Biomedicine, EURAC research, Bolzano, Italy, E-mail: fabiola.delgreco@eurac.edu*
[b] *Respiratory Epidemiology, Occupational Medicine and Public Health, NHLI, Imperial College, London, United Kingdom*
[c] *Department of Health Sciences, University of Leicester, Leicester, United Kingdom*

Abstract

Mendelian randomisation (MR) estimates causal effects of modifiable phenotypes on an outcome by using genetic variants as instrumental variables but its validity relies on the assumption of no pleiotropy, i.e. the genes influence the outcome only through the given phenotype. Excluding pleiotropy is difficult but the use of multiple instruments can indirectly address the issue: if all genes represent valid instruments, their MR estimates should vary only by chance. The Sargan test detects pleiotropy when individual genotype, phenotype and outcome data are measured in the same subjects. We propose an alternative approach to be used when only summary data are available or when data on gene-phenotype and gene-outcome come from different subjects. The presence of pleiotropy is investigated using the between-instrument heterogeneity Q test (together with the I^2 index) in a meta-analysis of MR Wald estimates, derived separately from each instrument. For a continuous outcome, we evaluate the approach through simulations and illustrate it using published data. For the scenario where all data come from the same subjects, we compare it with the Sargan test. The Q test tends to be conservative in small samples. Its power increases with the degree of pleiotropy and the sample size, as does the precision of the I^2 index, in which case results are similar to those of the Sargan test. In MR studies with large sample sizes based on summary data, the between-instrument Q test represents a useful tool to explore the presence of heterogeneity due to pleiotropy or other causes.

Keywords: Mendelian randomisation; instrumental variables; pleiotropy; heterogeneity Q test; I^2 index

1. Introduction

Genetic variants are increasingly being used as instrumental variables to provide estimates of the causal association between modifiable intermediate phenotypes and outcomes.

This approach, called Mendelian randomisation (MR) is not affected by the confounding and reverse causation typical of observational studies [1, 2, 3] and represents a method for assessing causality in epidemiology when randomised clinical trials are not possible. If a genetic variant, carried by an individual from birth and hence not modified by classical confounding factors or the presence of the outcome, is associated with increased levels of the intermediate phenotype and the same variant is associated with the outcome, then one can infer that the association between intermediate phenotype and outcome is causal [1, 2]. MR can be used both to test for causality and to provide an unconfounded estimate of the effect of the intermediate phenotype (hereafter referred to as phenotype) on the outcome. The current availability of large collections of genome-wide association (GWA) data on phenotypes and outcomes are a tremendous resource for exploiting the potential of the approach, which in part explains the recent increase in MR studies published in the literature.

The most commonly used estimators for continuous outcomes are the two-stage least square (2SLS) estimator and the Wald estimator, which provide the same estimate of the phenotype-outcome causal effect in the case of a single instrument with linear associations between instrument, phenotype, confounder and outcome and no statistical interactions [4, 5, 6, 7]. With the 2SLS estimator, the MR estimate is derived by a two-stage regression, where a linear regression of the phenotype on the instrument is performed in the first stage, followed by a linear regression of the continuous outcome on the predicted value of the phenotype. This analysis requires individual data on the genetic variant (G), phenotype (X) and continuous outcome (Y) measured in the same subjects. On the other hand, the Wald estimator derives the MR estimate indirectly as the ratio of the estimate of the $G - Y$ association over the estimate of the $G - X$ association [7], so that it can be used when only summary data are available for either association or when evidence on $G - X$ and $G - Y$ associations come from different studies.

An important statistical problem for MR is weak instrument bias, which is evident when the $G - X$ association has low power. The strength of that association can be measured using the F statistic (F) of the $G-X$ regression, which depends on the sample size, N, the number of instruments, m, and their effect sizes: $\mathrm{F} = \frac{R^2(N-m-1)}{m(1-R^2)}$, where R^2 is the proportion of the variance of the phenotype that is explained by the instruments. In the situation where data on G, X and Y come from the same subjects, the weak instrument bias of the MR estimator has been shown to be in the direction of the confounded observational $X - Y$ association, and corresponds to approximately 100/F percent of the observational estimate. Thus, if F is 20, the bias is about 5% of the observational estimate. Instruments with F higher than 10 are conventionally described as 'strong' [3, 4, 8, 9, 10, 11].

The validity of the MR approach relies on three assumptions about the genetic variant used as an instrumental variable (IV): 1) the IV should be associated with the phenotype; 2) the IV should not be associated with any unobserved confounder of the association between phenotype and outcome; 3) the IV should be conditionally independent of the outcome given

the phenotype and any confounder, i.e. the genetic variant influences the outcome only through the given phenotype [7]. The first assumption does not represent a problem since only genetic variants associated with the phenotype are selected as instruments. The second assumption is unlikely to be an issue either in the context of MR, given that the genetic variant should not be associated with typical confounding factors [6]. The third assumption is the crucial one because its violation can lead to misleading tests and biased MR estimates.

The most likely cause of violation of the third IV assumption is pleiotropy, which occurs when a genetic variant influences multiple intermediate phenotypes that separately affect the outcome of interest. The assumption can be violated through other mechanisms including canalization and population stratification. Developmental canalization occurs when a genetic variant expressed during fetal development or post-natal growth stimulates compensatory processes that protect against the effect of that variant on the outcome in adulthood [1], while population stratification occurs when ancestral sub-populations with different allele frequency and outcome distributions confound the $G-Y$ association [1, 5, 6].

1.1. Investigation of Pleiotropy

Ruling out pleiotropic effects on biological grounds is challenging even for well-studied genes. However, the use of multiple independent genetic instruments not only increases the power of the MR analysis [10], but it also allows investigation of pleiotropy [6, 12]. If all of the genetic variants are valid instruments, their individual MR estimates will only vary by chance and so a larger between-instrument heterogeneity would indicate a violation of the IV assumptions, most likely due to pleiotropy in one or more of the genetic variants. When using data on $G-X$ and $G-Y$ that were collected on the same subjects, formal testing of pleiotropy can be performed using the Sargan test [13], an over-identification test associated with the two-stage regression that tests the null hypothesis that all instruments provide the same MR estimate [12]. The test requires that at least one of the instruments is valid (e.g. not affected by pleiotropy), and the test statistic is NR_u^2, where R_u^2 is the proportion of the variance of the residuals of the second-stage regression explained indirectly by the instruments. Under the null hypothesis, the test statistic follows a chi-square distribution with $m-1$ degrees of freedom [4]. Rejection of the null hypothesis indicates lack of validity of at least one instrument, which could be due to pleiotropy or to a violation of the third IV assumption due to some other cause.

In this study we investigate a simple approach to testing for the presence of pleiotropy, or other causes of violation of the third IV assumption, that can be used in MR studies based on multiple instruments when only summary data for $G-X$ and $G-Y$ associations are available, or when evidence on the two associations come from different studies. The approach is based on the between-instrument heterogeneity observed in a meta-analysis of the Wald estimates obtained separately for each of the instruments, with presence and magnitude of the heterogeneity evaluated using the heterogeneity Q test and the I^2 index. The Q test [14] is the statistical test most commonly used to assess the presence of heterogeneity

in a meta-analysis, but it suffers from low statistical power when the number of estimates to be pooled in the meta-analysis is small [15]. To overcome this problem and to provide an estimate of the magnitude of the heterogeneity, Higgins and colleagues proposed the use of the I^2 index, defined as the percentage of total variation in the estimates explained by heterogeneity rather than sampling error, which is independent of the number of estimates to be pooled [16, 17]. As with the Sargan test, this approach cannot detect pleiotropy if all instruments share exactly the same pleiotropic effects, since all estimates would be similarly biased and there would be no heterogeneity.

Through simulations of MR studies with continuous outcome, we measure the type I error and power of our approach. We investigate scenarios with different degree of pleiotropy, number of instruments and sample size, considering both situations where the evidence on $G-X$ and $G-Y$ come from the same study and from separate studies. The paper is organised as follows. The method is described in Section 2. In Section 3 we investigate the performance of our approach through simulation work, comparing it to the Sargan test when appropriate ($G-X$ and $G-Y$ data from the same study). An illustrative example using published summary data on birth weight and fasting glucose levels in adults is provided in Section 4.

2. Methods

2.1. Mendelian randomisation analysis

For the Mendelian randomisation analysis of summary data, the Wald estimator can be used to obtain a MR estimate of the effect of X on Y for each instrument G_k separately,

$$\hat{\beta}_{XY}^{(k)} = \frac{\hat{\beta}_k}{\hat{\alpha}_k},$$

that is the regression coefficient of G_k on Y ($\hat{\beta}_k$) is divided by the regression coefficient of G_k on X ($\hat{\alpha}_k$) [3, 18, 19].

The estimated variance of each Wald estimate can be approximated using the delta method [3, 18, 19] based on the first terms of the Taylor series expansion:

$$Var\left(\hat{\beta}_{XY}^{(k)}\right) \approx \left(\hat{\beta}_{XY}^{(k)}\right)^2 \left(\frac{Var(\hat{\alpha}_k)}{\hat{\alpha}_k^2} - 2r\frac{\sqrt{Var(\hat{\alpha}_k)Var(\hat{\beta}_k)}}{\hat{\alpha}_k\hat{\beta}_k} + \frac{Var(\hat{\beta}_k)}{\hat{\beta}_k^2} \right),$$

where r is the correlation between the estimated regression coefficients for $G-X$ and $G-Y$, which is equal to zero when X and Y come from separate studies. When individual subject data are available from a single study, the Wald estimator gives identical results to the 2SLS estimator and the Taylor series variance is extremely close to the variance estimates from 2SLS given by programs such as ivregress in Stata. A simpler variance approximation assumes $Var(\hat{\alpha}_k) = 0$ so that the variance of the MR estimator becomes $Var(\hat{\beta}_k)/\hat{\alpha}_k^2$ [20, 21];

this is close to the Taylor series variance for large samples but underestimates the variance when the sample size is small.

In our analysis, the pooled MR estimate was obtained using inverse variance weighted fixed-effect meta-analysis across instruments [21]:

$$\hat{\beta}_{XY} = \frac{\sum_{k=1}^{m} \hat{\beta}_{XY}^{(k)} / Var(\hat{\beta}_{XY}^{(k)})}{\sum_{k=1}^{m} 1/Var(\hat{\beta}_{XY}^{(k)})}$$

with standard error:

$$se(\hat{\beta}_{XY}) = \left(\sum_{k=1}^{m} \frac{1}{Var(\hat{\beta}_{XY}^{(k)})} \right)^{-1/2}.$$

2.2. Testing and estimating the between-instrument heterogeneity

In large samples, the Q statistic follows a chi-square distribution with $m-1$ degrees of freedom under the null hypothesis of homogeneity [16]. It is defined as:

$$Q = \sum_{k=1}^{m} w_k (\hat{\beta}_{XY}^{(k)} - \mu_F)^2,$$

where m is the number of estimates to be pooled (here corresponding to the number of instruments), w_k is the weight for the estimate $\hat{\beta}_{XY}^{(k)}$ and represents the precision (reciprocal of the variance) of the estimate, and μ_F is a weighted mean estimate calculated as $\mu_F = \sum w_k \hat{\beta}_{XY}^{(k)} / \sum w_k$.

The I^2 index [16, 17], defined as the percentage of total variance in the estimates to be pooled explained by heterogeneity rather than sampling error, is related to the Q statistic through the equation:

$$I^2 = \begin{cases} \frac{Q-(m-1)}{Q} \times 100, & \text{for } Q \geq m-1 \\ 0, & \text{for } Q < m-1 \end{cases}$$

with its 95% confidence interval equal to

$$1 - \frac{1}{\exp(\ln(H)) \pm 1.96 se(\ln(H))},$$

where $H = \sqrt{\frac{Q}{m-1}}$, and

$$se(\ln(H)) = \begin{cases} \frac{1}{2} \times \frac{\ln(Q - \ln(m-1))}{\sqrt{2Q} - \sqrt{2m-3}}, & \text{for } Q > m \\ \sqrt{\frac{1}{2(m-2)} \left(1 - \frac{1}{3(m-2)^2}\right)}, & \text{otherwise.} \end{cases}$$

As a rough guide to judge the degree of heterogeneity, Higgins and colleagues have suggested that I^2 values below 25% indicate mild heterogeneity, values over 50% suggest severe heterogeneity and between 25% and 50% the heterogeneity can be described as moderate [17].

2.3. Data simulation

We simulated genetic studies with sample size N composed of m biallelic genetic variants G_k, in Hardy-Weinberg equilibrium [22] and with no linkage disequilibrium between them (statistically independent). A continuous phenotype X was created that was affected by all of the G_k, together with a continuous outcome Y affected by X and possibly, by a secondary continuous phenotype Z (pleiotropic effect), and some continuous confounders U and C.

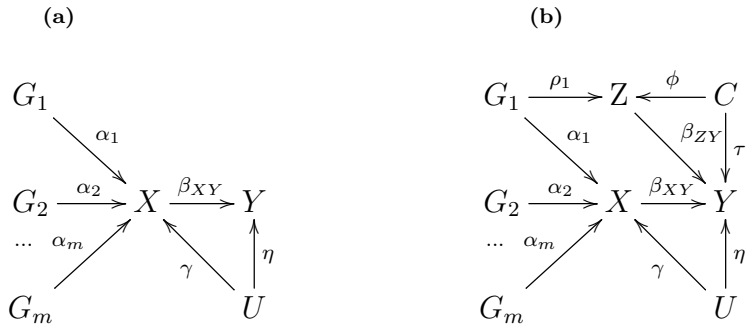

Figure 1: Mendelian randomisation with multiple independent instruments (a) without pleiotropy; (b) with pleiotropy in one instrument

The simulation in the absence of pleiotropy followed the pattern illustrated in Figure 1a. The genotypes g_{ik}, coded as 0, 1, or 2, were randomly generated from a binomial distribution:

$$g_{ik} \sim Binomial(2, f_k),$$

with a random allele frequency f_k generated uniformly between 0.1 and 0.9. The value of U, for the ith subject, was generated as $u_i \sim Normal(0, 1)$, and the value of the X created with an additive genetic effect:

$$x_i \sim Normal\Big(\sum_{k=1}^{m} \alpha_k g_{ik} + \gamma u_i, \sigma_X\Big),$$

where α_k is the effect of a single copy of the effect allele on X, γ is the effect of U on X, and σ_X is the standard deviation of the random error. The value of Y was generated from:

$$y_i \sim Normal(\beta_{XY} x_i + \eta u_i, \sigma_Y),$$

where β_{XY} is the causal effect that we want to estimate, η is the effect of U on Y, and σ_Y is the standard deviation of the random error.

We fixed the proportion of the variance of the phenotype that is explained by each genetic variant to be 1% ($R^2_{G_kX} = 0.01$), to reflect the reality that genetic variants in genetic association studies typically have a small effect on a phenotype. This implies that under an additive model of inheritance, the effect size of the single genetic variant on the phenotype was $\alpha_k = \sqrt{\frac{0.01}{2f_k(1-f_k)}}$ [23], and α_k varied between 0.141 and 0.236. We set the proportion of the variance of X explained by U to be 0.25 ($\gamma = 0.5$) and made $\sigma_X = \sqrt{1 - 0.25 - m0.01}$ [23] so that the total variance of X was 1. We set σ_Y equal to σ_X and set $\eta = \gamma$ and $\beta_{XY} = 1$ so that X explained about 50% of the variance in Y.

Type I error probabilities for the Q test were obtained for 5, 10 and 20 instruments with the $G - X$ and $G - Y$ data coming from one single study that evaluates both associations or from two separate studies of equal size. In both scenarios we used sample sizes of $1,000$, $5,000$, $10,000$ and $20,000$. For comparison, the simulations for one single study were also analysed using 2SLS and the Sargan test.

The pattern followed by the simulation in the presence of pleiotropy is illustrated in Figure 1b. To investigate the power of the between-instrument Q test, the pleiotropic instruments were allowed to affect two intermediate phenotypes X and Z.
The genotypes and X were simulated as before and Z, which is affected by m' of the genetic variants, was simulated according as:

$$z_i \sim Normal\Big(\sum_{k=1}^{m'} \rho_k g_{ij} + \phi c_i, \sigma_Z\Big),$$

where ρ_k is the effect of the pleiotropic variant on Z, c_i is the value of the confounder of Z and Y generated from a standard normal distribution, ϕ is its effect size and σ_Z is the standard deviation of Z; so that,

$$y_i \sim Normal(\beta_{XY} x_i + \beta_{ZY} z_i + \eta u_i + \tau c_i, \sigma_Y),$$

where β_{ZY} is the effect of Z on Y, τ is the effect of C on Y, and σ_Y is the standard deviation of the random error in Y.

We set the value of the single genetic effect on the pleiotropic phenotype equal to the effect on the primary phenotype ($\rho_k = \alpha_k$), and the proportion of variance explained by the confounders equal to 0.25 ($\gamma = \eta = \phi = \tau = 0.5$). We also set $\sigma_Y = \sigma_Z = \sigma_X$.

For a variant with a pleiotropic effect, the regression coefficient of Y on G_k adjusted for all confounders is $\rho_k\beta_{ZY} + \alpha_k\beta_{XY}$, so the relative importance of the pleiotropic pathway compared with the pathway of interest is $\rho_k\beta_{ZY}/\alpha_k\beta_{XY}$. When there are several independent instruments the combined effect of all of the instruments on Y has the form $\sum \rho_k\beta_{ZY} + \sum \alpha_k\beta_{XY}$ where the first summation is over the pleiotropic variants and the second summation is over all pleiotropic and non-pleiotropic variants. Thus, if there are m

instruments and the first m' of these are pleiotropic, we measured the amount of pleiotropy by,

$$\mathbf{P} = \frac{\beta_{ZY} \sum_{k=1}^{m'} \rho_k}{\beta_{XY} \sum_{k=1}^{m} \alpha_k} \times 100.$$

Under this definition $\mathbf{P} = 100\%$ implies that the pleiotropic effect of G_k on Y via Z is the same as the effect of interest via X. In the simulations we had a single pleiotropic pathway so that all pleiotropic variants acted through the same variable Z and we kept $\beta_{ZY} = \beta_{XY}$ fixed and varied the magnitude of the pleiotropy by changing ρ_k.

The power to detect heterogeneity due to pleiotropy was evaluated with the degree of pleiotropy, $\mathbf{P}$, set to 50%, 100% and 200%. Pleiotropy was introduced into 20% of the total number of instruments. When pleiotropy was present in more than one instrument the same degree of pleiotropy was given to each instrument.

We repeated each analysis $10,000$ times. As well as the type I error and power of the Q test, we recorded the percentage of simulations for which I^2 fell in each of the categories, $[0-25)$; $[25-50)$; $[50-100]$.

3. Results

3.1. Type I error

The results for the type I error of the Q test in the absence of pleiotropy are reported in Table 1. Since the results are based on $10,000$ simulations, the type I errors will have a standard error of 0.2% when the type I error is 5% falling to 0.1% when the type I error is 1%. The test is conservative as the type I error is always lower than the nominal level of 5%. The type I error gets closer to its nominal value with increasing sample size but for a given sample size it tends to decrease with increasing number of instruments. The differences between one study and two studies are small and consistent with the natural variation in the simulations. In scenarios with a single study or two studies, the distribution of I^2 values is positively skewed with the mass of the distribution concentrated on values smaller than 25%. However, the interpretation of I^2 as an indicator of heterogeneity is not independent of sample size or number of instruments and high I^2 values may be observed even in the absence of pleiotropy, with small numbers of instruments and large sample sizes. A formal test based on I^2 would have the same properties as the Q test as the two statistics are connected by the formula given in section 2.2, $\mathrm{I}^2 = 100(Q-(m-1))/Q$. Thus, when there are $m = 5$ instruments, Q would be significant at the 5% level if it is above $\chi^2_4(0.95) = 9.49$ and I^2 would be significant if it is above $100(9.49-4)/9.49 = 58\%$.

To investigate the cause of the conservatism of the Q test we plotted the ordered Q statistics generated under the null against the expected values from a chi-square distribution with

Table 1: Type I error in a nominal 5% Q test and corresponding values of I^2 in the absence of pleiotropy, for scenarios with $G - X$ and $G - Y$ data from a single study or from two separate studies. N = sample size, IVs = number of instruments. All the values are reported as percentage.

			Single Study				**Two Studies**		
		Type I	**I^2**			**Type I**	**I^2**		
N	**IVs**	**error**	**$0-25$**	**$25-50$**	**$50-100$**	**error**	**$0-25$**	**$25-50$**	**$50-100$**
$1,000$	5	0.6	88.6	10.3	1.1	0.6	90.7	7.9	1.4
$5,000$	5	3.5	77.1	16.6	6.3	3.0	79.3	15.3	5.4
$10,000$	5	4.5	75.0	17.3	7.7	3.8	77.1	16.4	6.5
$20,000$	5	4.4	75.4	16.3	8.3	4.0	75.5	16.9	7.6
$1,000$	10	0.3	95.2	4.8	0.0	0.2	95.8	4.1	0.1
$5,000$	10	2.6	83.5	15.6	0.9	2.8	84.5	13.4	2.1
$10,000$	10	3.9	80.8	17.3	1.9	3.4	81.8	15.6	2.6
$20,000$	10	4.5	79.7	17.9	2.4	4.3	80.2	16.6	3.2
$1,000$	20	0.1	99.0	1.0	0.0	0.1	99.0	1.0	0.0
$5,000$	20	2.5	90.1	9.3	0.3	2.2	90.7	9.2	0.1
$10,000$	20	3.4	86.5	13.1	0.4	3.2	88.5	11.3	0.2
$20,000$	20	4.2	86.6	13.9	0.5	4.1	86.8	12.7	0.5

$m - 1$ degrees of freedom. The results are shown in Figure 2. In all cases the Q statistics are less dispersed than would be expected under a chi-square distribution with the under-dispersion much more noticeable at low sample sizes. The results for single studies and for two separate studies are both shown on each plot but they are so similar that they are difficult to distinguish by eye. Two factors affect these plots. First, small samples are affected by weak instrument bias that makes some of the MR estimates for individual instruments very unstable, and second, there is a positive correlation between the estimates and their variances, as can be seen from the Taylor approximation to the variance given in section 2.1 in which the variance is proportional to the square of the Wald estimator; this same correlation is also present in the 2SLS analysis and in a slightly weaker form when the variance approximation that assumes $Var(a_k) = 0$ is used (data not shown). This correlation means that when an individual MR estimate is randomly high it will tend to have a large variance and therefore a low weight, w_i, in the Q statistic. Large positive deviations are down-weighted and as a result the Q statistic is under-dispersed. With small samples, there is greater weak instrument bias which means that the sampling distribution of the Wald estimates is more skewed towards large estimates and the impact is greater. In larger samples the weak instrument bias is less, the sampling distribution is more symmetrical and the

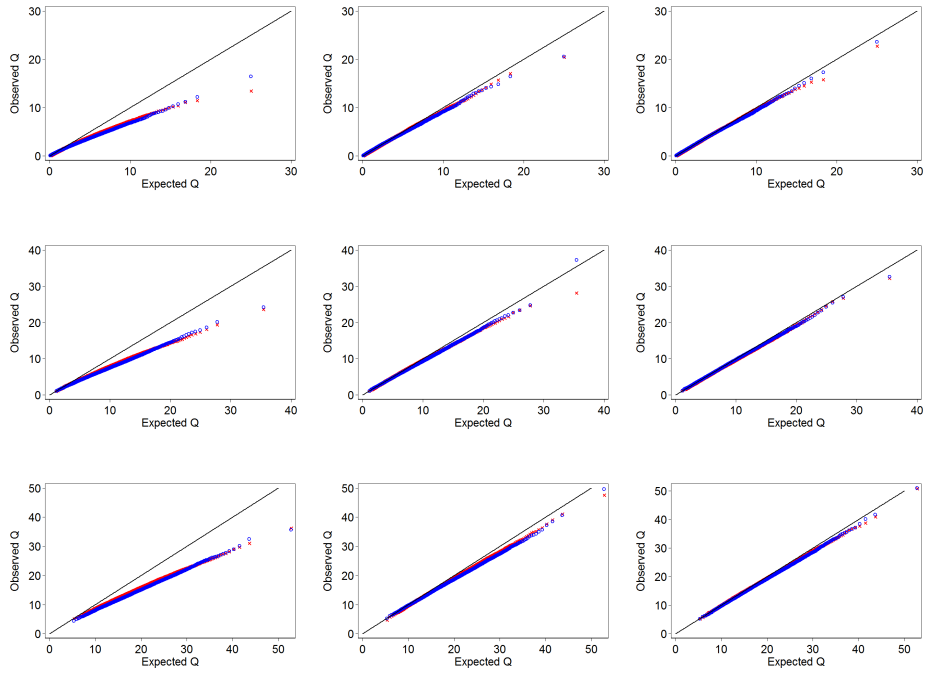

Figure 2: Quantile-Quantile plots of the Q statistic. The columns represent sample sizes of $1,000$, $5,000$ and $10,000$, the rows represent 5, 10 and 20 instruments. Results for a single study are shown as circles and for two studies are shown as crosses.

effect of the correlation tends to cancel between randomly high and randomly low individual IV estimates.

3.2. Power

Table 2 shows the apparent power of the Q statistic when the pleiotropy index $\mathbf{P} = 100\%$, but this has to be interpreted considering that the type I error does not reach the nominal level. Corresponding results for $\mathbf{P} = 50\%$ and $\mathbf{P} = 200\%$ are given in supplementary Tables S1 and S2. The power of the Q test to detect pleiotropy increases with sample size and, for a fixed small sample size $(1,000 - 5,000)$, decreases with the number of instruments when $G - X$ and $G - Y$ data came from a single study. An opposite trend is observed when data came from two separate studies.

Figure 3 shows how the power increases with the strength of the pleiotropy as measured by the index $\mathbf{P}$. The power is greater when the $G-X$ and $G-Y$ associations come from a single study reflecting the greater precision of the individual IV estimates. Even relatively large sample sizes will have low power to detect heterogeneity below 100%, which corresponds to a pleiotropic pathway of equal strength to the pathway that we are seeking to measure.

Table 2: Power of the Q test for a 5% significance level and corresponding values of $\mathbf{I^2}$ when the pleiotropy is 100% and affects 20% of the variants, for scenarios with $G-X$ and $G-Y$ data from a single study or from two separate studies. N = sample size, IVs = number of instruments. All the values are reported as percentage.

		Single Study				**Two Studies**			
			$\mathbf{I^2}$				$\mathbf{I^2}$		
N	**IVs**	**Power**	$\mathbf{0-25}$	$\mathbf{25-50}$	$\mathbf{50-100}$	**Power**	$\mathbf{0-25}$	$\mathbf{25-50}$	$\mathbf{50-100}$
1,000	5	4.4	62.5	26.4	11.1	1.7	80.1	16.5	3.4
5,000	5	86.6	2.2	5.8	92.0	37.8	20.2	27.5	52.3
10,000	5	99.7	0	0.1	99.9	80.4	3.2	8.6	88.2
20,000	5	100	0	0	100	99.3	0	0.3	99.7
1,000	10	2.2	79.8	18.6	1.6	0.8	86.1	10.3	0.6
5,000	10	83.6	3.6	17.5	78.9	44.6	21.1	40.5	38.4
10,000	10	99.7	0	0.4	99.6	89.4	1.5	11.2	87.3
20,000	10	100	0	0	100	100	0	0.1	99.9
1,000	20	0.6	95.3	4.7	0	0.4	96.8	3.2	0
5,000	20	77.0	9.4	46.2	44.4	46.9	27.1	56.8	16.1
10,000	20	99.7	0.1	3.8	96.1	92.8	1.8	26.9	71.3
20,000	20	100	0	0	100	100	0	0.3	97.7

The use of a less conservative significance level (10%) to increase the power of the Q test, as has been recommended [24], makes no difference to power when the sample is very large (20,000) for any level of pleiotropy, and it makes a difference of limited magnitude for smaller sample sizes (Table S3).

The results for the I^2 index also reflect those of the Q test. The ability of I^2 to detect pleiotropy increases with increasing sample size, but decreases with increasing number of instruments, for both scenarios of a single study and two studies (Table 2). Only when the sample size reaches 20,000 do we see severe ($> 50\%$) levels of heterogeneity in almost all simulated samples. Table 2 also shows that the ability of I^2 to detect pleiotropy is better for the scenario where $G-X$ and $G-Y$ data came from the same study.

Data on the effects of the pleiotropy on the pooled MR estimate is given in the supplement. The magnitude of the bias in the MR estimate is proportional to the degree of pleiotropy (Table S4). In general, the bias increases with increasing number of instruments in both scenarios with one single study and two separate studies. The overall bias in MR estimates is smaller when $G-X$ and $G-Y$ data came from two separate studies, partly due to the

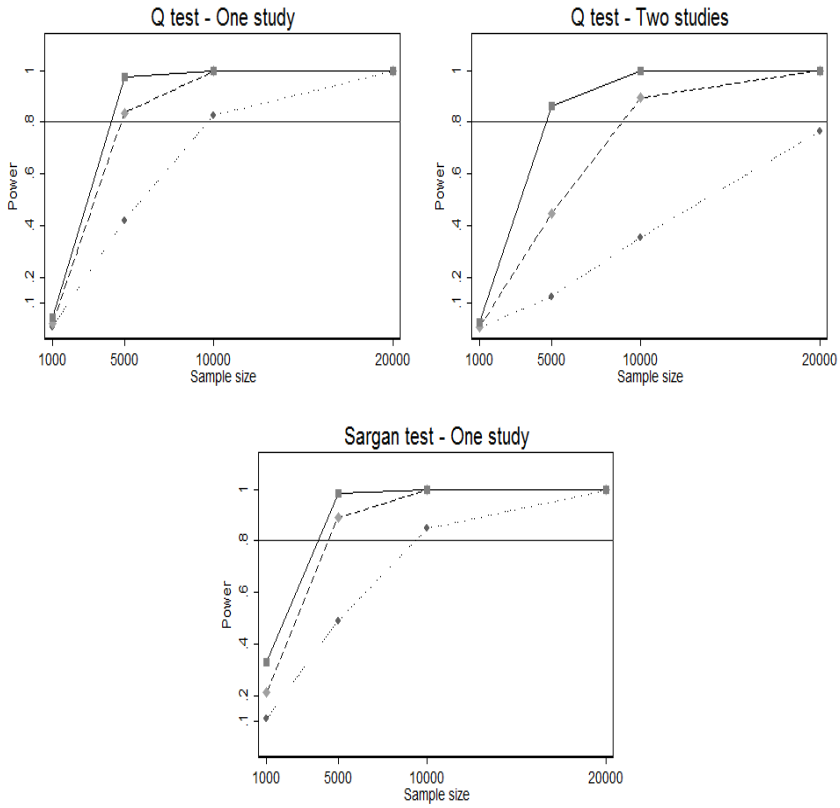

Figure 3: Power of the Q test and the Sargan test for a 5% significance level in scenario with 10 instruments for different sample sizes and pleiotropy, $\mathbf{P} = 50\%$ (dotted line), $\mathbf{P} = 100\%$ (dashed line) and $\mathbf{P} = 200\%$ (solid line).

presence of small-sample bias that is in the opposite direction (towards the null) to the bias induced by the pleiotropic effect (away from the null). Finally, the coverage for the MR estimate decreases with increasing sample size, number of instruments and magnitude of pleiotropy.

3.3. Comparison with the Sargan test

Table 3 shows the results for the type I error probability of the Sargan test when using the 2SLS regression based on individual data from a single study. The type I error probability of the Sargan test is always close to the nominal significance level, and is more stable than the Q test for all sample sizes and numbers of instruments. When individual level data are available, the 2SLS analysis is able to use all of the instruments in a single regression. Thus if there are 5 instruments, they will together explain 5% of the variance while the individual variants each explain 1%. Although the average F statistic for the 5 instruments is of similar average size to the F statistic for the single instrument regression, the Q statistic requires that all 5 regressions have F$>$ 10 while the Sargan test only requires a single F statistic to

Table 3: Type I error in a nominal 5% Sargan test by sample size (N) and number of instruments (IVs), in the absence of pleiotropy for the scenario with individual data from one single study. Values are reported as percentage.

N	IVs	Type I error
1,000	5	4.9
5,000	5	4.9
10,000	5	4.9
20,000	5	4.8
1,000	10	5.4
5,000	10	5.1
10,000	10	4.9
20,000	10	5.4
1,000	20	5.1
5,000	20	5.0
10,000	20	4.9
20,000	20	5.0

be over 10.

The results for the power of the Sargan test for scenarios with 10 instruments are reported in Figure 3, while results for the scenarios with 5 and 20 instruments are given in the supplementary Figure S1. As with the Q test, the power increases with increasing sample size, extent of pleiotropy, and number of instruments. For high levels of pleiotropy (200%), the power reaches 80% with 10 instruments and a sample size smaller than 5,000. The Sargan test out-performs the Q test, reflecting the better calibration of the type I errors. This advantage of the Sargan test over the Q test is more evident with larger numbers of instruments, although the two tests performed similarly when the study is very large.

4. Illustrative example: birth weight and glucose levels in adulthood

Many observational studies have been performed to investigate the association between low birth weight and adult fasting glucose levels, but they have led to different conclusions [25, 26, 27]. However, estimates from these studies could have been affected by confounding. We used the MR approach described in Section 2.1 to estimate a causal association of low birth weight with fasting glucose levels, using published results of two meta-analyses of GWA studies (equivalent to the scenario of two separate studies, i.e. a meta-analysis of $G - X$ and a separate meta-analysis of $G - Y$), and we explored the possible presence of

pleiotropy with the Q test and the I^2 index. The results of the GWA meta-analyses for birth weight and glucose are available on the websites of the Early Growth Genetics (EGG) Consortium [28] and the Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) [29], respectively.

As instruments, we chose seven independent genetic variants from the GWA meta-analysis on birth weight which had p-value lower than 10^{-5} in the discovery sample and which were replicated, i.e. the p-value in the combined analysis of discovery and replication studies was smaller than the genome-wide significance threshold of 5×10^{-8}. In order to control for the selection bias due to the winner's curse, i.e. the overestimation of the genetic effects in the discovery stage associated with the selection of 'top hits' [29], for the MR analysis we used the estimates obtained from the replication analysis performed on 25 studies, with a total of $42,415$ individuals. Even though these seven instruments all together explained a very low percentage of the variability of birth weight (0.7%), due to the large sample size only one instrument had an F statistic smaller than 10 (Table 4). The F statistic for all SNPs together was 33.9.

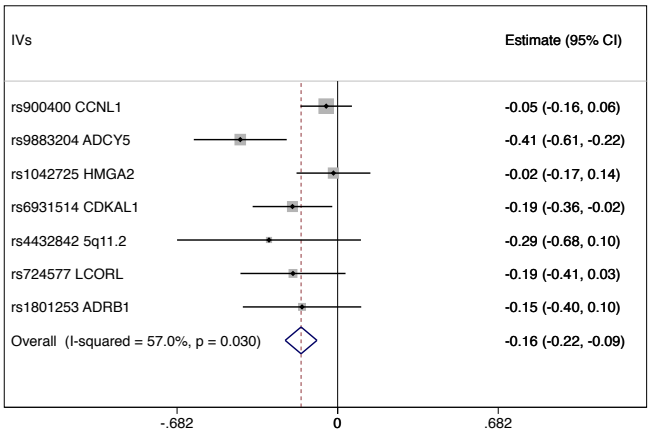

Figure 4: Forest plot of the MR estimates from the seven instruments. The size of the squares is proportional to the precision of the MR estimates for each polymorphism, with the horizontal lines indicating their 95% confidence intervals. The combined MR estimate is represented by the centre of the diamond, with the lateral tips indicating its 95% confidence interval. The solid vertical line is the line of no effect.

We looked at the association of these instruments with fasting glucose levels in the meta-analysis of 21 studies, with a total sample size of $46,186$ individuals. All the data used in the MR analysis are reported in Table 4.

The results of the MR analysis using one instrument at a time are reported in Table 4 and Figure 4. The meta-analysis of the seven MR estimates results in a statistically significant

combined effect of -0.155 (95% confidence interval: -0.223 to -0.088; p-value: 6.8×10^{-6}), indicating a protective effect of birth weight on fasting glucose levels in adults, with a reduction of 0.155 mmol/L in fasting glucose levels per standard deviation unit increase in birth weight. With a standard deviation for birth weight of 484 gr in the genetic association study [28], this corresponds approximately to an effect of -0.0003 mmol/L on fasting glucose levels per 1 gr increase in birth weight.

The Forest plot in Figure 4 shows the strong effect of two genetic variants: rs9883204 in the *ADCY5* gene and rs6931514 in the *CDKAL1* gene, which are already known as type 2 diabetes loci [27, 30, 31, 32]. There is statistically significant evidence of heterogeneity across instruments, with a Q test p-value of 0.03, and an I^2 index of 57% (95% CI: 0% to 81%). These results suggest that the third IV assumption may be violated for some of the instruments. The Forest plot also shows that the heterogeneity is primarily due to the variant in the *ADCY5* gene that showed the largest effect. Exclusion of this variant from the MR analysis leads to a non-statistically significant Q test (p-value= 0.450) and a reduction of the I^2 index to 0% (95% CI: 0% to 75%), with still a statistically significant MR estimate of -0.098; (95% CI: -0.168 to -0.027; p-value: 6.6×10^{-3}).

Table 4: Illustrative example: Data and findings of the MR analyses for the individual instruments. IV = instrument, Chr = chromosome, AF = frequency of the reference allele from the meta-analysis on glucose levels, N = sample size, R^2 = percentage of the variance of birth weight explained by the instruments, F = F statistic, β = effect estimate, se = standard error, p = p-value.

					Birth weight						**Fasting glucose**				**MR analysis**		
IV	**Gene**	**Chr**	**Allele**	**AF**	**N***	**R^2**	**F**	β	**se**	**p**	**N**	β	**se**	**p**	β	**se**	**p**
rs900400	*CCNL1*	3	C	0.39	34,329	0.25	84.9	-0.072	0.007	7.5×10^{-22}	45,727	0.004	0.004	0.371	-0.049	0.056	0.383
rs9883204	*ADCY5*	3	C	0.76	34,721	0.12	42.7	-0.058	0.009	2.4×10^{-11}	45,726	0.024	0.005	9.3×10^{-8}	-0.414	0.101	4.0×10^{-5}
rs1042725	*HMGA2*	12	T	0.50	41,828	0.10	42.4	-0.045	0.007	1.1×10^{-11}	46,186	0.001	0.004	0.819	-0.018	0.080	0.824
rs6931514	*CDKAL1*	6	G	0.29	42,415	0.10	43.7	-0.050	0.007	5.9×10^{-12}	45,056	0.010	0.004	0.019	-0.192	0.086	2.6×10^{-2}
rs4432842	*5q11.2*	5	C	0.30	26,808	0.02	6.5**	-0.024	0.009	8.0×10^{-3}	46,171	0.007	0.004	0.080	-0.292	0.199	0.143
rs724577	*LCORL*	4	C	0.73	29,057	0.06	17.4	-0.039	0.009	1.2×10^{-5}	45,062	0.007	0.004	0.069	-0.190	0.114	9.6×10^{-2}
rs1801253	*ADRB1*	10	G	0.26	23,071	0.05	12.2	-0.037	0.010	3.9×10^{-4}	42,074	0.006	0.005	0.213	-0.151	0.128	0.238

The R^2 are calculated using the formula $2\text{AF}(1-\text{AF})\beta^2$ [23]. *N varies since a different number of studies contributed to the data for the individual genetic variant. **This genetic variant is considered as weak instrument, because its F is smaller than 10.

The sensitivity analysis excluding the instrument with the strongest effect therefore confirmed the negative association between birth weight and glucose levels in adulthood. Although sensitivity analyses can help identify sources of heterogeneity and may point to variants with possible pleiotropic effects, the investigation of causes of pleiotropy relies on biological knowledge.

5. Discussion

In a MR study, the third IV assumption states that the instrument should affect the outcome only through the phenotype of interest and a violation of this assumption could lead to false conclusions about the causality of the phenotype and outcome association. The assumption is violated in the presence of pleiotropy but pleiotropy is difficult to rule out even for well-studied genes. The use of multiple instruments (genetic variants) with independent effects on the intermediate phenotype represents an indirect approach to evaluating pleiotropy and other causes of heterogeneity, since their MR estimates should vary only by chance when all of the instruments are valid. Heterogeneity can be tested using the Sargan over-identification test [4, 13], but this test can only be used when individual level data are available on the genetic variant, phenotype, and the outcome. This precludes its use in situations, which are likely to become more and more common, where summary data on gene-phenotype and gene-outcome associations are obtained from separate sources, often represented by different international genetics consortia. In this paper we propose an alternative approach to assessing possible violations of the third assumption, based on the between-instrument heterogeneity in a meta-analysis of Wald estimates across multiple instruments, that can be used in this case. In particular, our aim was to evaluate the between-instrument Q test as a tool to provide statistical evidence for the possible presence of pleiotropy.

Through simulations we evaluated our approach, in terms of type I error probability and power of the Q test, for scenarios differing in the magnitude of pleiotropy, number of instruments, sample size and whether gene-phenotype and gene-outcome data come from the same or from separate studies. Our results suggest that the Q test performs well when the sample size is large and when the gene-phenotype and gene-outcome data come from the same study. When data come from two separate studies, a large sample ($> 20,000$) is needed to detect a weak or moderate pleiotropic effect. However, such large sample sizes are not uncommon in genome-wide association that combine data across studies collaborating within an international consortium. A similar behaviour was observed for magnitude and precision of the I^2 index, a measure of the heterogeneity across estimates in a meta-analysis.

Recently published MR studies have used up to 20 instruments [33, 34, 35], with the main purpose of generating more precise MR estimates [11]. For that reason, we assessed whether the power to detect pleiotropy with our approach increased when increasing the number of instruments used. Although the combined effect of all of the instruments will explain more variance in the phenotype, the Q test requires that each of the variants used separately is able to produce a reasonable estimate of the causal effect of phenotype on the continuous outcome and thus increasing the number of instruments does not compensate for sample size.

The type I error rate of our approach is conservative, especially for small sample size. This conservatism is a result of the correlation between the Wald estimator and its variance, which means that large positive deviations in the individual MR estimates are down-weighted by having a large variance. The chi-square approximation becomes better when increasing the

sample size, since this tends to cancel the effects of such correlation. In practice, the use of summary data on gene-phenotype and gene-outcome associations from large genome-wide association consortia, which is the setting of interest in this paper, can address this issue by guaranteeing large sample sizes in future MR studies. Our approach assumes that the Q statistic has a chi-square distribution under the null hypothesis of no pleiotropy, which requires that the meta-analysed statistics (MR estimates of the individual genetic variants) are statistically independent. However, even when the genetic variants are independent (no linkage disequilibrium) and the $G-X$ and $G-Y$ data come from separate studies, there remains the theoretical possibility of a weak correlation between MR estimates induced by the fact that the genetic associations are estimated on the same individuals for all variants. Both the simulations and the illustrative example confirm the lack of precision in the estimate of the I^2 index when the number of estimates to be pooled is small, with large confidence intervals containing the value of zero even in the presence of high I^2 values [36, 37, 38, 39, 40]. The confidence intervals of the I^2 index become narrower when increasing the number of instruments for both scenarios with one single study and two separate studies.

In terms of pooled MR estimates across instruments, our simulations show not only biased estimates induced by pleiotropy, but also the presence of a bias related to the sample size in scenarios with no pleiotropy. However, the small-sample bias is also present when performing a two-stage least squares regression analysis with individual data. The small-sample bias seems to be related to the strength of the instrument, and gets smaller when increasing the variance of the phenotype explained by the instruments (data not shown).

Our findings show that, in MR studies based on large sample sizes, the between-instrument Q test represents a good tool to detect heterogeneity among MR estimates. However, it does not give information on the source of that heterogeneity, which could be due to pleiotropy as well as other reasons, such as confounding by population stratification. This is true also for the Sargan test, which does not provide information on the conditions leading to the rejection of the null hypothesis, and cannot identify which instruments are not valid. Moreover, neither test would be able to detect pleiotropy in the situation where all of the instruments share the same pleiotropic effect, leading to similarly biased MR estimates (Tables S4-S5). The Sargan test out-performs the Q test when the study is small, given the better calibration of the type I errors. The two tests perform similarly when the sample is very large.

Data on birth weight trait has been contributed by the EGG Consortium and has been downloaded from www.egg-consortium.org. Data on glycaemic traits have been contributed by MAGIC investigators and have been downloaded from www.magicinvestigators.org.

References

[1] Davey-Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease. *Int J Epidemiol* 2003; **32**: 1-22

[2] Davey-Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol* 2004; **333**: 30-42

[3] Thomas D, Conti D. Commentary: the concept of Mendelian randomization. *International Journal of Epidemiology* 2004; **33**: 21-25
[4] Wooldridge JM. *Econometric analysis of cross section and panel data*. Cambridge: MIT Press, 2002
[5] Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in medical research* 2007; **16**: 309-330
[6] Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey-Smith G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* 2008; **27**: 1133-1163
[7] Didelez V, Meng S, Sheehan N. Assumptions of IV Methods for Observational Epidemiology. *Statistical Science* 2010; **25**: 22-40
[8] Bound J, Jaeger D, Baker R. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 1995; **90**: 443-450
[9] Bowden RJ, Turkington DA. *Instrumental variables*. Cambridge: University press, 2007
[10] Pierce B, Ahsan H, VanderWeele T. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *International Journal of Epidemiology* 2011; **40**: 740-752
[11] Sheehan NA, Didelez V. Commentary: Can 'many weak' instruments ever be 'strong'?. *International Journal of Epidemiology* 2011; **40** (3): 752-754
[12] Palmer TM, Harbord RM, Sheehan NA, Tobias JH, Timpson NJ, Davey-Smith G, et al. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research* 2012; **21**: 223-242
[13] Sargan JD. The estimation of economic relationships using instrumental variables. *Econometrica* 1958; **26**: 393-415
[14] Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954; **10**: 101-129
[15] Hardy RJ. Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* 1998; **17**: 841-856
[16] Higgins JP. Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; **21**: 1539-1558
[17] Higgins JP,Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003; **327**: 557-560
[18] Bautista LE, Smeeth L, Hingorani AD, Casas JP. Estimation of bias in nongenetic observational studies using "mendelian triangulation". *Annals of Epidemiology* 2006; **16**: 675-680
[19] Thomas DC, Lawlor DA, Thompson JR. Re: Estimation of bias in nongenetic observational studies using "mendelian triangulation" by Bautista et al. *Annals of Epidemiology* 2007; **17**: 511-513
[20] Ehret G (The International Consortium for Blood Pressure Genome-wide Association Studies). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk *Nature* 2011; **478**: 103-109
[21] Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology* 2013; 1-8
[22] Lange K. *Mathematical and statistical methods for genetic analysis*. 2nd edition New York: Springer, 2003
[23] Falconer DS, Mackay TFC. *Introduction to Quantitative Genetics*. 4th edition Harlow, Essex, UK: Longman 1996
[24] Petitti DB, Approaches to heterogeneity in meta-analyses. *Statistics in medicine* 2001; **20**: 3625-3633
[25] Barker DJP. *Mother, babies, and health in later life*. 2nd edition Edinburgh: Churchill Livingstone 1998
[26] Whincup PH, Kaye SJ, Owen CG, Huxley R, Cook DG, Anazawa S, et al. Birth Weight and Risk of Type 2 Diabetes. A Systematic Review. *JAMA* 2008; **300**: 2886-2896
[27] Sayers SM, Mott SA, Mann KD, Pearce MS, Singh GR. birth weight and fasting glucose and insulin levels: results from the Aboriginal Birth Cohort Study. *Med J Aust* 2013; **199**:112-116
[28] Horikoshi M, Yaghootkar H, Mook-Kanamori DO, Sovio U, Taal HR, Hennig BJ, et al. New loci

associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nat Genet* 2013; **45**: 76-82

[29] Dupuis J, Langenberg C, Prokopenko I, Saxena R, at al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* 2010; **42**:105-116

[30] Freathy RM, Bennett AJ, Ring SM, Shields B, Groves CJ, Timpson NJ, et al. Type 2 diabetes risk alleles are associated with reduced size at birth. *Diabetes* 2009; **58**: 1428-1433

[31] Zhao J, Li M, Bradfield JP, Wang K, Zhang H, Sleiman P, et al. Examination of type 2 diabetes loci implicates $CDKAL1$ as a birth weight gene. *Diabetes* 2009, **58**: 2414-2418

[32] Freathy RM, Mook-Kanamori DO, Sovio U, Prokopenko I, Timpson NJ, Berry DJ, et al. Variants in $ADCY5$ and near $CCNL1$ are associated with fetal growth and birth weight. *Nat Genet* 2010; **42**: 430-435

[33] Kivimaki M, Magnussen CG, Juonala M, Kahonen M, Kettunen J, Loo BM, et al. Conventional and Mendelian randomization analyses suggest no association between lipoprotein(a) and early atherosclerosis: the Young Finns Study. *International Journal of Epidemiology* 2011; **40**: 470-478

[34] De Silva NM, Freathy RM, Palmer TM, Donnelly LA, Luan J, Gaunt T, et al. Mendelian randomization studies do not support a role for raised circulating triglyceride levels influencing type 2 diabetes, glucose levels, or insulin resistance. *Diabetes* 2011; **60**:1008-1018

[35] Ken-Dror G, Humphries SE, Kumari M, Kivimaki M, Drenos F. A genetic instrument for Mendelian randomization of fibrinogen. European *Journal of Epidemiology* 2012; **27**: 267-79

[36] Huedo-Medina T, Sanchez-Meca J, Marin-Martinez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistics or I^2 index? *Psychological Methods* 2006; **11**: 193-206

[37] Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analysis. *BMJ* 2007a; **335**: 914-916

[38] Ioannidis JP, Patsopoulos NA, Evangelou E. Heterogeneity in Meta-Analyses of Genome-Wide Association Investigations. *Plos One* 2007b; **2**: e841

[39] Mittlböck M, Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analysis. *Statistics in Medicine* 2010; **37**: 4321-4333

[40] Thorlund K, Imberger G, Johnston BC, Walsh M, Awad T, et al. Estimates and their 95% confidence interval in large meta-analyses. *Plos One* 2012; **7**: e39471