John R. Thompson*, Cosetta Minelli and Fabiola Del Greco M

Mendelian Randomization using Public Data from Genetic Consortia

DOI 10.1515/ijb-2015-0074

Abstract: Mendelian randomization (MR) is a technique that seeks to establish causation between an exposure and an outcome using observational data. It is an instrumental variable analysis in which genetic variants are used as the instruments. Many consortia have meta-analysed genome-wide associations between variants and specific traits and made their results publicly available. Using such data, it is possible to derive genetic risk scores for one trait and to deduce the association of that same risk score with a second trait. The properties of this approach are investigated by simulation and by evaluating the potentially causal effect of birth weight on adult glucose level. In such analyses, it is important to decide whether one is interested in the risk score based on a set of estimated regression coefficients or the score based on the true underlying coefficients. MR is primarily concerned with the latter. Methods designed for the former question will under-estimate the variance if used for MR. This variance can be corrected but it needs to be done with care to avoid introducing bias. MR based on public data sources is useful and easy to perform, but care must be taken to avoid false precision or bias.

Keywords: genetic risk score, Mendelian randomization, genome-wide association study

1 Introduction

Mendelian randomization (MR) is the name given to an instrumental variable analysis in which one or more genetic variants are used as the instrument [1]. In principle, it offers a very powerful way of using non-randomized data to establish causal relationships between an exposure and an outcome, but in practice it has two major limitations. First, individual genetic effects tend to be weak so that large sample sizes are required to detect those effects with the accuracy required by MR [2]. Second, it is vital that we are able to select genetic instruments that act on the final outcome only through the intermediate exposure [3], that is, the genes must not have pleiotropic effects that change the same outcome via different pathways. The weakness of the effects of individual genetic variants has led investigators to replace single genes by the combined effects of sets of variants. Unfortunately the extra variants make it even more difficult to guarantee that there is no pleiotropy.

In recent years a large number of consortia have made public the meta-analysed results from genome-wide association studies of specific traits. Wiki-Genes (http://www.wikigenes.org/e/art/e/185.html) lists over a hundred such genetic consortia. By no means all have made summary data public, but data on top hits are often given in the supplement to their main paper and most consortia will supply information on specific variants if requested. Typically these meta-analyses cover many hundreds of thousands of variants and report the separate effect sizes of each genetic variant on the trait, together with the p-values, standard errors or confidence intervals. Results are not given for the combined effects of sets of variants, but if variants are chosen that are independent of one another, the coefficients in a joint regression will be the same as those for the separate variants so that a joint genetic risk score can approximated from the published results.

***Corresponding author: John R. Thompson,** Department of Health Sciences, Leicester, UK, E-mail: trj@le.ac.uk
Cosetta Minelli, Population Health and Occupational Disease Section, National Heart and Lung Institute, Imperial College, London, UK
Fabiola Del Greco M, Center for Biomedicine, EURAC, Bolzano, Italy

In 2011 an influential paper developed a genetic risk score for blood pressure based on the results from their own consortium and then applied that score to other traits using publicly available findings from other consortia [4]. In this way they were able to show, amongst other things, that a genetic risk score for blood pressure shows a significant association with coronary artery disease but not with kidney disease.

There has been some investigation of the statistical properties of Mendelian randomization based on multiple genetic instruments when exposure and outcome are measured on the same subjects [2, 5–7] and recently Burgess et al. have considered MR based on summary data for multiple instruments but again primarily in the context of exposure and outcome measured in the same study [8, 9].

In this paper we consider the properties of different ways of performing a MR analysis using a genetic risk score estimated from one study and applied to a second study. The key point underlying this work is that there is an important difference between the estimate the effect of the theoretically best genetic risk score for one variable on a second variable and the estimate the effect of the fitted genetic risk score in a particular sample on a second variable. While the point estimator for both situation is the same, their standard errors are different.

2 Methods

2.1 The Mendelian randomization ratio estimator

Suppose that a study or meta-analysis reports the results of regression analyses for each of m genetic variants, G_j, that are associated with their trait, X. These results might be in the form of the estimated regression coefficients, $a_{X\,j}$, and their variances, $V_{X\,j}$, or other statistics from which these quantities can be derived. It is important that the selection of the genetic variants is not based on the same data that is used to calculate $a_{X\,j}$ for otherwise the estimated coefficients will be biased away from zero due to the Winner's curse [10], so $a_{X\,j}$ and $V_{X\,j}$ might be taken from a replication study. These estimated regression coefficients will be modelled as,

$$a_{Xj} \sim N\left(\alpha_{Xj},\, V_{Xj}\right) \quad j = 1,\, \ldots,\, m$$

where α_{Xj} is the true regression coefficient and the $V_{X\,j}$ will be treated as known.

A second study or meta-analysis publishes similar data for the same variants but a different outcome, Y. The variants are unlikely to be top hits for Y so now we will need access to the full set of results in order to look-up the required estimates. The Winner's curse is no longer a concern because the variants were not chosen for their effect in the second study. Suppose that the regression coefficients and variances from the look-up are $b_{Y\,j}$ and the $V_{Y\,j}$, we can model them as,

$$b_{Yj} \sim N\left(\beta_{Yj},\, V_{Yj}\right) \quad j = 1,\, \ldots,\, m$$

where the β_{Yj} represent the true coefficients and the $V_{Y\,j}$ are treated as known.

A Mendelian randomization for a continuous outcome targets the unconfounded regression coefficient, ϕ, of X on Y. Provided that the assumptions for Mendelian randomization hold for every genetic variant, there are m relationships $\beta_{Yj} = \phi\alpha_{Xj}$, each of which creates a ratio, or Wald, estimator $\hat{\phi}_j = b_{Yj}/a_{Xj}$ [11, 12]. These can be averaged with weights inversely proportional to their variances in order to create an overall Mendelian randomization estimate $\hat{\phi}$.

When the selected genetic variants are independent, we can estimate the variance of $\hat{\phi}$ without needing to make assumptions about the unknown pattern of linkage disequilibrium and the coefficients α_{Xj} that apply to each variant separately will also be the coefficients in the joint regression of X on the genetic risk score S_{Xi}. That is,

$$E\{X_i|S_{Xi}\} = \mu + S_{Xi} = \mu + \sum_{j=1}^{m} \alpha_{Xj} g_{ij}$$

$$E\{Y_i|S_{Xi}\} = \nu + \phi E\{X_i|S_{Xi}\} = \nu + \phi\mu + \phi \sum_{j=1}^{m} \alpha_{Xj} g_{ij}$$

where the confounder is omitted because it is assumed independent of each G_j and g_{ij} represents the measured genotype of the i^{th} subject for variant G_j coded as the number of effect alleles, 0,1 or 2. This genetic risk score S_{Xi}, assumes a per allele effect of each variant and ignores any interactions. Dominant or recessive genetic effects could be created but the necessary estimates of the coefficients are rarely published. In this model S_{Xi} represents the ideally weighted combination of the variants for use as a combined instrument in a Mendelian randomization.

We could estimate the variances of the ratio estimates of each variant using a Taylor series [13],

$$Var\{\hat{\phi}_j\} \approx [\beta_{Yj}/\alpha_{Xj}]^2 [V_{Xj}/\alpha_{Xj}^2 + V_{Yj}/\beta_{Yj}^2)]$$

where the covariance term is omitted because X and Y come from different studies and their regression coefficients are independent. To estimate this variance we could just replace α_{Xj} and β_{Yj} by $a_{X\,j}$ and $b_{Y\,j}$. The weights, w_j, needed for averaging the ϕ_j would then be the inverse of these variances and the resulting Mendelian randomization estimate of ϕ would have variance,

$$Var\{\hat{\phi}\} = \sum w_j^2 Var\{\hat{\phi}_j\} / \left[\sum w_j\right]^2 = 1/\sum w_j$$

However, as we will see in the simulations, inverse variance weighting does not work well in this context.

Should we want to test the hypothesis that $\phi = 0$, we would need the variance when this hypothesis is true, that is when $\beta_{Yj} = 0$. In that case the variance reduces to V_{Yj}/α_{Xj}^2.

2.2 The ICBP estimator

The International Consortium for Blood Pressure Genome-Wide Association Studies [4] considered a subtly different question to Mendelian randomization. Their analysis takes the ratio estimates b_{Yj}/a_{Xj} and averages them using weights $w_j = [V_{Yj}/a_{Xj}^2]^{-1}$. This produces,

$$\hat{\phi} = \frac{\sum_{j=1}^{m} \frac{b_{Yj}/a_{Xj}}{V_{Yj}/a_{Xj}^2}}{\sum_{j=1}^{m} \frac{1}{V_{Yj}/a_{Xj}^2}} = \frac{\sum_{j=1}^{m} \frac{b_{Yj}a_{Xj}}{V_{Yj}}}{\sum_{j=1}^{m} \frac{a_{Xj}^2}{V_{Yj}}}$$

$$Var\{\hat{\phi}\} = \left[\sum_{j=1}^{m} \frac{a_{Xj}^2}{V_{Yj}}\right]$$

We can think of this estimator either as an approximation to Mendelian randomization based on a simplified variance for $\hat{\phi}_j$ that ignores the uncertainty in $a_{X\,j}$, or as the correct estimator of the regression coefficient on Y on the genetic risk score,

$$S_{Xi}^{\star} = \sum_{j=1}^{m} a_{Xj} g_{ij}$$

where $S_{Xi}^{\star}$ differs from S_{Xi} because the $a_{X\,j}$ have replaced the α_{Xj}. This ICBP estimator correctly addresses the question of what would be the regression coefficient if Y were regressed on the genetic risk score based on the particular study that provided X. Thus instead of estimating ϕ we would in fact be estimating the regression coefficient for the model,

$$E\{Y_i|S^*_{Xi}, U_i\} = \nu^* + \phi^* S^*_{Xi}$$

and the actual value of this coefficient is,

$$\phi^* = \phi \frac{\sum_{j=1}^{m} \alpha_j a_j \sigma_j^2}{\sum_{j=1}^{m} a_j^2 \sigma_j^2}$$

where σ_j^2 is the variance of G_j (see supplementary methods). This distinction between regression on S_X and S^*_X is blurred by the fact that, in both cases, the natural estimators for ϕ and ϕ^* based on the individual variants are b_{Yj}/a_{Xj}, although the variances of these estimators do differ.

2.3 Bias adjustment

A problem arises with both the Mendelian randomization ratio estimator and the ICBP estimator because the sampling distribution of b_{Yj}/a_{Xj} has a noticeable skewness especially when the study measuring X is small or the true effect size α_{Xj} is small. Employing the delta method based on a Taylor series [13],

$$E\{b_{Yj}/a_{Xj}\} \approx \phi[1 + V_{Xj}/\alpha_{Xj}^2]$$

So we can obtain a less biased estimate of ϕ by replacing b_{Yj}/a_{Xj} with $[b_{Yj}/a_{Xj}]/[1 + V_{Xj}/a_{Xj}^2]$.

2.4 Improved estimation of α_{Xj}

Much of the instability in MR estimates is due to the difficulty of estimating α_{Xj} accurately, because when the estimate, a_{Xj}, approaches zero the ratio, b_{Yj}/a_{Xj}, will become large and unstable. Were ϕ known, under the assumptions for Mendelian randomization, we would have two related relationships for each variant,

$$a_{Xj} \sim N(\alpha_{Xj}, V_{Xj})$$
$$b_{Yj}/\phi \sim N(\alpha_{Xj}, V_{Yj}/\phi^2)$$

So a better estimate of α_{Xj} would be the weighted average,

$$\hat{\alpha}_{Xj} = \frac{\frac{a_{Xj}}{V_{Xj}} + \frac{b_{Yj}/\phi}{V_{Yj}/\phi^2}}{\frac{1}{V_{Xj}} + \frac{1}{V_{Yj}/\phi^2}} = \frac{\frac{a_{Xj}}{V_{Xj}} + \frac{\phi b_{Yj}}{V_{Yj}}}{\frac{1}{V_{Xj}} + \frac{\phi^2}{V_{Yj}}}$$

Of course, ϕ is not known but we could create a two-stage procedure by estimating its value, for example by using the ICBP estimator, and then using that estimate in place of ϕ to improve the estimates of the α_{Xj}. Because Mendelian randomization is so sensitive to the accuracy of the estimates of α_{Xj}, basing those estimates on both X and Y may produce less bias and greater stability, provided the assumptions of Mendelian randomization hold for each variant.

2.5 Simulation

To investigate the properties of the different estimators, a simulation study was performed that was based on the model shown diagrammatically in Figure 1. The simulations were conducted twice for each scenario as if there were two independent but identically designed studies with the G_j and X taken

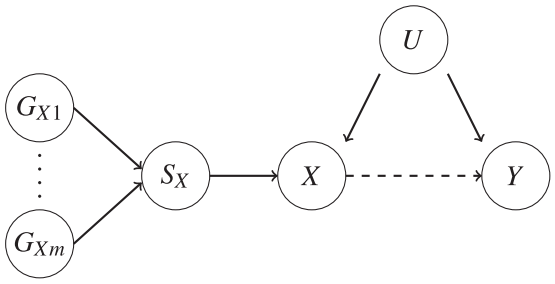

Figure 1: The model for the simulations based on genetic variants G_1 to G_m combined into a risk score, S_X that drives and exposure X and possibly an outcome Y. The association between X and Y is confounded by an unobserved U.

from one study and G_j and Y were taken from the other. In each case we considered three sample sizes, 1,000, 5,000 and 20,000.

In constructing the genetic risk score we used either 5, 10 or 50 independent variants. The minor allele frequencies of the variants were randomly selected to lie between 0.1 and 0.9, and the coefficients were adjusted so that each variant explained the same percentage of the variance in X. In the case of a score based on 5 genes, each gene explained 1% of the variance and in the cases of 10 and 50 genes, each gene explained 0.5% of the variance. So the scores based on 5 and 10 variants both explained 5% of the variance in X and the 50 genes explained 25%. The unconfounded effect of X on Y, ϕ, was varied between 0, 0.3, 0.6 or 0.9. The variances of X and Y were fixed at one so that for different ϕ, X explained 0%, 9%, 36% or 81% of the variance in Y. The confounder explained a third of the non-genetic variance of X and a third of the variance of Y that was not due to X. All scenarios were repeated 10,000 times.

3 Results

3.1 Simulation

Table 1 summarizes the performance of the ICBP estimator [4] when the Mendelian randomization model holds and all genes act on Y through X. When we want to estimate the coefficient of the genetic risk score, $S^*_{Xi} = \sum_{j=1}^{m} a_{Xj} g_{ij}$, then the expected value of that coefficient varies depending on the results of the first study. Alternatively, we might want to estimate the regression coefficient for the regression of the genetic risk score, $S_{Xi} = \sum_{j=1}^{m} \alpha_{Xj} g_{ij}$, on Y in which case the true answer is always the value of ϕ used in the simulation. Regression on S_{Xi} is the appropriate analysis for a Mendelian randomization.

Table 1 shows that the estimation of ϕ^* is very good, as one might expect since this is the situation that the ICBP estimator is designed to tackle. The coverage stays at its nominal level, the bias is small and RMSE decreases with sample size and with the percent of the variance explained by the genes. The results for 5 and 10 genes are very similar as both explain the same percentage of the variance of X, while the RMSEs for 50 genes are much smaller. The RMSEs for a sample size of 20,000 are about half those for a sample size of 5,000 consistent with negligible bias and an increase of four times in the sample size.

The ICBP results for estimating the MR coefficient, ϕ, used in the simulation, are less impressive. The coverage is correct for $\phi = 0$ but falls as ϕ increases. The bias is negative and increases in magnitude with ϕ and is especially noticeable with small samples. It decreases with sample size but increases with the number of genes. A bias (x1,000) of -50 for $\phi = 0.3$ or -100 for $\phi = 0.6$ or -150 for $\phi = 0.9$ would represent a 17% error in estimation. In summary, the ICBP analysis performs well when estimating the regression coefficient for S^*_{Xi} but is only approximate when used in a Mendelian randomization.

Table 2 compares the results of several different estimators of the MR coefficient, ϕ, for the middle sample size of 5,000. First we consider the use of the Taylor series variance estimate in an inverse-variance weighted analysis as described in section 2.1. When compared with Table 1, performance is

Table 1: Performance of the ICBP estimator when used to estimate the regression coefficients of two different risk scores.

Sample Size†	Genes§	ϕ	Statistic to be estimated					
			ϕ^* = Coefficient of S^*_{Xi}			ϕ = Coefficient of S_{Xi}		
			Bias‡	RMSE‡	Coverage	Bias‡	RMSE‡	Coverage
1,000	5	0.0	−0.9	139.1	95.2	−0.9	139.1	95.2
1,000	5	0.3	4.0	153.9	94.9	−12.5	159.6	93.9
1,000	5	0.6	−0.5	164.7	94.4	−35.8	186.5	90.2
1,000	5	0.9	1.9	159.6	94.8	−48.6	204.9	85.1
1,000	10	0.0	−1.6	134.3	94.9	−1.6	134.3	94.9
1,000	10	0.3	−2.4	146.1	94.7	−44.8	156.1	92.6
1,000	10	0.6	0.5	151.9	95.0	−83.8	185.7	87.3
1,000	10	0.9	0.1	148.3	94.9	−127.6	219.2	77.0
1,000	50	0.0	0.2	58.6	95.2	0.2	58.6	95.2
1,000	50	0.3	−0.1	63.6	94.6	−48.3	81.3	86.1
1,000	50	0.6	1.2	67.9	94.4	−94.6	120.7	65.3
1,000	50	0.9	1.0	68.6	93.3	−143.8	166.4	40.5
5,000	5	0.0	−0.0	62.5	95.4	−0.0	62.5	95.4
5,000	5	0.3	−0.4	68.7	95.2	−3.8	71.5	93.9
5,000	5	0.6	−0.9	72.7	95.0	−7.5	82.6	91.1
5,000	5	0.9	1.1	72.1	94.6	−8.8	92.7	86.0
5,000	10	0.0	−0.5	63.1	95.0	−0.5	63.1	95.0
5,000	10	0.3	0.1	67.1	95.2	−9.2	70.0	94.1
5,000	10	0.6	0.2	71.9	94.7	−18.4	81.6	90.7
5,000	10	0.9	−0.8	70.0	94.7	−27.7	93.2	84.2
5,000	50	0.0	−0.0	27.9	94.7	−0.0	27.9	94.7
5,000	50	0.3	0.5	29.9	94.9	−10.6	33.0	92.0
5,000	50	0.6	0.4	32.5	94.2	−21.3	42.6	84.3
5,000	50	0.9	−0.1	33.0	93.0	−33.3	53.8	71.1
20,000	5	0.0	0.4	31.7	94.8	0.4	31.7	94.8
20,000	5	0.3	−0.5	34.2	94.8	−1.5	35.5	94.3
20,000	5	0.6	−0.5	36.5	94.9	−2.3	41.1	91.5
20,000	5	0.9	0.1	35.8	94.6	−2.4	45.6	86.9
20,000	10	0.0	0.1	31.6	94.8	0.1	31.6	94.8
20,000	10	0.3	0.2	34.0	94.9	−2.1	35.2	94.3
20,000	10	0.6	−0.5	35.5	95.1	−5.0	40.5	91.0
20,000	10	0.9	−0.0	35.7	94.5	−7.1	46.2	86.0
20,000	50	0.0	0.1	14.2	94.9	0.1	14.2	94.9
20,000	50	0.3	0.0	15.3	94.7	−2.8	16.2	93.1
20,000	50	0.6	0.1	16.6	94.1	−5.6	19.7	88.4
20,000	50	0.9	0.2	17.0	92.7	−8.3	23.4	80.3

Note: † samples sizes for the studies of X and Y assumed equal.
§ 5 genes each explaining 1% of the variance in X. 10 and 50 genes each explaining 0.5% of the variance
‡ bias and RMSE, root mean square error, (x1,000).

actually worse than that for ICBP estimator, despite the improved variance estimation for the individual ratios. The problem lies in the correlation between the ratio estimates and the variance estimates. When the estimated ratio for a particular variant is randomly high, its estimated variance will also be larger. This correlation causes randomly large ratios to be down-weighted in an inverse-variance weighted analysis and so creates a bias towards zero. The effect of this correlation is removed if we use weights that do not depend on the estimated variances as in the second block of Table 2 that contains the results for a simple average with equal weights.

Table 2: Performance of different estimators for ϕ when the Mendelian randomization model is appropriate.

Sample Size†	Genes§	ϕ	Bias‡	RMSE‡	Coverage	Bias‡	RMSE‡	Coverage
				Inverse-variance			Simple average	
5,000	5	0.0	−0.0	60.7	96.3	0.0	64.6	95.8
5,000	5	0.3	−13.0	70.8	95.1	6.0	74.1	95.4
5,000	5	0.6	−25.4	84.8	92.9	11.8	87.2	95.6
5,000	5	0.9	−35.3	97.9	91.1	20.4	100.2	95.2
5,000	10	0.0	−0.4	59.5	96.7	−0.1	68.8	95.8
5,000	10	0.3	−28.2	71.8	94.2	13.6	78.4	96.2
5,000	10	0.6	−55.0	94.5	88.3	27.5	94.1	96.1
5,000	10	0.9	−81.5	119.4	81.2	40.9	110.5	96.5
5,000	50	0.0	−0.0	26.3	96.3	−0.2	30.6	95.6
5,000	50	0.3	−30.5	42.4	84.7	14.3	37.6	94.4
5,000	50	0.6	−60.6	70.1	59.0	28.1	50.8	91.1
5,000	50	0.9	−91.1	99.9	35.6	40.5	63.6	88.7
				Bias adjustment			Improved α_j	
5,000	5	0.0	0.0	63.2	95.7	0.0	63.6	96.0
5,000	5	0.3	−0.6	72.0	95.4	0.6	72.0	95.7
5,000	5	0.6	−1.5	83.5	95.3	−0.3	82.7	95.5
5,000	5	0.9	0.5	94.0	95.2	−0.4	92.4	95.0
5,000	10	0.0	−0.2	64.8	95.9	−0.1	66.3	96.4
5,000	10	0.3	−1.5	72.1	96.2	0.6	72.4	96.7
5,000	10	0.6	−3.0	82.2	96.0	−1.6	80.7	96.2
5,000	10	0.9	−4.7	91.8	96.1	−8.4	89.0	95.5
5,000	50	0.0	−0.2	28.9	95.5	−0.2	29.4	96.1
5,000	50	0.3	−0.9	32.5	95.7	0.3	32.5	96.5
5,000	50	0.6	−2.4	38.7	95.1	−3.2	37.9	95.2
5,000	50	0.9	−5.0	44.4	95.0	−12.5	44.0	93.1

Notes: † samples sizes for the studies of X and Y assumed equal.
§ 5 genes each explaining 1% of the variance in X. 10 and 50 genes each explaining 0.5% of the variance
‡ bias and RMSE, root mean square error, (x1,000).

The third block of Table 2 shows further improvement by using the Taylor series adjustment to the estimate of the ratio as described in section 2.3 and the final block shows the result of using estimates of α_{Xj} that use information on both the gene-X and gene-Y relationships as described in section 2.4. In both cases the performance is improved.

3.2 Birth weight and glucose levels in adulthood

Horikoshi et al. published the results of a meta-analysis of genome-wide studies of birth weight [14]. They identified seven loci that replicated with a p-value below 5×10^{-8}. Table 3 shows the replication results for the seven SNPs. The Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) have made their genome-wide results freely available on the internet [15] and in Table 3 we show the association between the lead SNPs for birth weight and fasting glucose level; similar data can be found in the supplementary table 5 of Horikoshi et al. [14]. The results were used to define a risk score for birth weight, which was used as the instrument in a Mendelian randomization by looking at its association with fasting glucose levels in adults.

Many epidemiological studies have found that low birth weight babies are at increased risk of diabetes [16] and if this relationship is causal we would expect that genes that are negatively related to birth weight

Table 3: Markers associated with birth weight [14] and their association with glucose levels in adulthood [15].

Locus	SNP	EA	Birth weight replication			Glucose GWAS		
			Coeff	SE	*p*-value	Coeff	SE	*p*-value
CCNL1	rs900400	C	−0.072	0.007	7.5e-22	+0.0035	0.0040	0.371
ADCY5	rs9883204	C	−0.058	0.009	2.4e-11	+0.024	0.0045	9.3e-8
HMGA2	rs1042725	T	−0.045	0.007	1.1e-11	+0.0008	0.0036	0.819
CDKAL1	rs6931514	G	−0.050	0.007	5.9e-12	+0.0096	0.0041	0.019
5q11.2	rs4432842	C	−0.024	0.009	8.0e-3	+0.0070	0.0040	0.080
LCORL	rs724577	C	−0.039	0.009	1.2e-5	+0.0074	0.0041	0.069
ADRB1	rs1801253	G	−0.037	0.010	3.9e-4	+0.0056	0.0045	0.213

Note: EA = Effect allele, SE = standard error.

would show a positive association with glucose levels and vice versa. The results in Table 3 do show such an inverse relationship.

Using the ICBP estimator $\hat{\phi} = -0.155\,(\text{se} = 0.032)$ with a pvalue $= 1.0 \times 10^{-6}$. This is the correct p-value as it uses the standard error calculated under the null. When we are interested in producing confidence intervals for a MR estimate it is important to incorporate the extra uncertainty due to the estimation of the weights in the risk score. We have already noted that the inverse-variance weighted average performs badly because it is biased towards zero, here it gives, $\hat{\phi} = -0.133\,(\text{se} = 0.034)$. The simple average of the individual ratios is less prone to bias and gives $\hat{\phi} = -0.186\,(\text{se} = 0.044)$. Adjusting for the bias in the individual ratios bring the solution down slightly, $\hat{\phi} = -0.176\,(\text{se} = 0.044)$ and much the same effect is seen if a two-stage procedure is used to improve the individual estimates of α_{Xj} giving $\hat{\phi} = -0.155\ (\text{se} = 0.034)$. It is evident that the simple ICBP analysis performed well and although it does underestimate the standard error, that under estimation is slight.

The key question that this analysis does not address is whether the assumptions required by a Mendelian randomization hold for these variants. The evidence that they are truly associated with birth weight is strong. The likely confounders between birth weight and glucose level in adulthood relate to lifestyle and these are unlikely to be associated with these genes, so the only likely confounder is ethnicity. The meta-analysis of birth weight was conducted across populations of European origin and each meta-analysis adjusted for internal population stratification using genomic control, so confounding by ethnicity is unlikely to be a major problem.

The chief concern with the validity of this Mendelian randomization is pleiotropy. Biological knowledge about these genes is limited although, for instance, HMGA2 has previously been associated with height while ADRB1 has been associated with blood pressure and heart failure. Findings such as these suggest that the genes act through different pathways and so some of these genes might have secondary effects with a long-term influence on glucose levels. Genes that exhibit such pleiotropy would give different ratio estimates from those obtained from valid instruments. The ratios b_{Yj}/a_{Xj} for the seven variants are, −0.05, −0.41, −0.02, −0.19, −0.29, −0.19, −0.15 each with an ICBP standard error of about 0.08. The difference between the second and third genes, ADCY5 and HMGA2, is 0.39 with a p-value of 4×10^{-4}. Adjusting for the possible 21 pairs of genes, the Bonferroni adjusted p-value is 8×10^{-3}. This is still significant and suggests that this analysis might be influenced by pleiotropy.

4 Discussion

Genome-wide associations measured by large consortia offer enormous potential for performing Mendelian randomizations. Not only can we investigate the effects of an exposure, X, on an outcome Y using a genetic risk score for X, but we could reverse the investigation and look at the effects of Y on X using a risk score for Y [17–19], or perhaps we could look at the effect of X on Y using SNPs that show an association with a third

factor or which are known to act through particular pathways. If we want to perform such analyses there is a range of estimators that could be used and as we have seen they are not all equally good.

The ICBP estimator is not designed for Mendelian randomization but provides a reasonable approximation provided that the sample sizes are large and the effect of X on Y is not too great. The ICBP estimator actually addresses a slightly different question to Mendelian randomization as it is concerned with the regression coefficient on Y of the particular risk score that best predicts X in the data supplied by the first consortium; this it does very well.

Burgess et al. have investigated the use of the ICBP estimator in the context of summary data on the exposure and outcome coming from the same study [8, 9]. As one might expect, they too conclude that the ICBP performs well across a range of scenarios.

When we are interested in Mendelian randomization we should allow for the uncertainty in the estimates provided by both consortia and this can be approximated by using a Taylor series for the variance. However, this variance estimate creates a problem because there is a correlation between the individual estimates of ϕ obtained from each variant and their corresponding variances. As we saw in Table 2, even a simple average will give better results than an inverse-variance weighted average. Table 2 also shows that a Taylor series adjusted estimate of the ratio that allows for this skewness can improve the final estimate of ϕ.

Mendelian randomization requires that all of the variants individually estimate the same ϕ and it should be possible to use the data to test this basic assumption. When both outcomes are measured on the same subjects the assumption can be assessed using the Sargan test [20]. This option is not available when the analysis is based on publicly available summary data but it would still be possible to test the assumption by measuring the variability in the individual ratios, perhaps using a chi-squared statistic of the type used to test for heterogeneity in a meta-analysis [21].

The methods described in this paper have all assumed that there is a linear relationship between the risk score and each of the continuous outcomes. However, many genetic consortia have looked at binary, disease related outcomes. Binary responses are usually analysed with a logistic link so that there is a non-linear relationship between the outcome and the genetic variants. The regression coefficients for the individual regressions are no longer unbiased estimates of the coefficients in the joint regression (although this effect will be small unless the genes jointly explain a lot of the variance) and the estimates of ϕ lose their causal interpretation, although this bias is also likely to be small [22].

Perhaps the trickiest issue for anyone planning to combine data from separate consortia is to satisfy themselves that the two studies are sufficiently similar. It would be a concern if the size of the effects of the variants on the exposure were different, perhaps because of measuring an average effect in the presence of a gene-environment interaction. At present most GWAS have been conducted on populations of European descent living in industrialised countries and so the exchangeability of the study populations is unlikely to be a major issue, but careful thought will be needed before mixing data from studies in widely differing settings.

Genetic consortia are making available summary data on more and more traits allowing for the possibility of increasingly complex Mendelian randomizations. Provided that these analyses are performed carefully they have the potential to produce important clues as to the causality behind the associations discovered in epidemiological studies.

Acknowledgement: 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. This work was in part supported by a travel grant from the Royal Society.

References

1. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med 2008;27(8):1133–63.

2. Burgess S, Thompson SG. Bias in causal estimates from Mendelian randomization studies with weak instruments. Stat Med 2011;30(11):1312–23.
3. Didelez V, Sheehan NA. Mendelian Randomization as an instrumental variable approach to causal inference. Stat Meth Med Res 2007;16:309–30.
4. Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 2011;478(7367):103–9.
5. Pierce BL, Ahsan H, Vanderweele TJ. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. Int J Epidemiol 2011;40(3):740–52.
6. Burgess S, Thompson SG, Consortium CC. Avoiding bias from weak instruments in Mendelian randomization studies. Int J Epidemiol 2011;40(3):755–64.
7. Palmer TM, Lawlor DA, Harbord RM, Sheehan NA, Tobias JH, Timpson NJ, et al. Using multiple genetic variants as instrumental variables for modifiable risk factors. Stat Methods Med Res 2012;21(3):223–42.
8. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Gen Epidemiol 2013;37:658–65.
9. Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. Int J Epidemiol 2013;42(4):1134–44.
10. Zollner S, Pritchard JK. Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am J Hum Genet 2007;80(4):605–15.
11. Wald A. The fitting of straight lines if both variables are subject to error. Ann Math Stat 1940;11:284–300.
12. Durbin J. Errors in variables. Rev Int Stat Inst 1954;22:23–32.
13. Kendall M, Stuart A. The advanced theory of statistics, Volume 1. London: C. Griffin, 1977.
14. Horikoshi M, Yaghootkar H, Mook-Kanamori DO, Sovio U, Tall 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(1):76–82.
15. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 2010;42(2):105–16.
16. 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. J Am Med Assoc 2008;300(24):2886–97.
17. Welsh P, Polisecki E, Robertson M, Jahn S, Buckley BM, de Craen AJ, et al. Unraveling the directional link between adiposity and inflammation: a bidirectional Mendelian randomization approach. J Clin Endocrinol Metab 2010;95(1):93–9.
18. Lyngdoh T, Vuistiner P, Marques-Vidal P, Rousson V, Waeber G, Vollenweider P, et al. Serum uric acid and adiposity: deciphering causality using a bidirectional Mendelian randomization approach. PLoS One 2012;7(6):e39321.
19. Vimaleswaran KS, Berry DJ, Lu C, Tikkanen E, Pilz S, Hiraki LT, et al. Causal relationship between obesity and vitamin D status: bi-directional Mendelian randomization analysis of multiple cohorts. PLoS Med 2013;10(2):e1001383.
20. Sargan JD. The Estimation of Economic Relationships Using Instrumental Variables. Econometrica 1958;26:392–415.
21. Del Greco-M F, Minelli C, Sheehan NA, Thompson JR. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. Stat Med 2015;34:2926–40.
22. Harbord R, Didelez V, Palmer T, Meng S, Sterne J, Sheehan N. Severity of bias of a simple estimator of the causal odds ratio in Mendelian randomization studies. Stat Med 2013;32:1246–58.

Supplementary Methods

Regression on a general risk score

Assume that Y is actually formed from its dependence on m genes, so that,

$$y_i = \phi \sum_{j=1}^{m} \alpha_j g_{ij} + \varepsilon_i$$

but we regress y_i on $x_i = \sum_{j=1}^{m} w_j g_{ij}$, where w_j are arbitrary weights that might or might not be equal to α_j. The variance covariance matrix of two values y_i and x_i will be,

$$\begin{bmatrix} \sum \phi^2 \alpha_j^2 \sigma_j^2 + \sigma^2 & \sum \phi \alpha_j w_j \sigma_j^2 \\ \sum \phi \alpha_j w_j \sigma_j^2 & \sum w_j^2 \sigma_j^2 \end{bmatrix}.$$

where σ^2 is the residual variance and σ_j^2 is the variance of G_j, which for variants in Hardy-Weinberg equilibrium will be equal to $2f_j(1-f_j)$ where f_j is the allele frequency.
The regression coefficient of Y on X will have expectation,

$$\phi \frac{\sum \alpha_j w_j \sigma_j^2}{\sum w_j^2 \sigma_j^2}$$

When we regress on the observed coefficients from the first study, a_j, this reduces to,

$$\phi^\star = \phi \frac{\sum \alpha_j a_j \sigma_j^2}{\sum a_j^2 \sigma_j^2}$$