Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2011; all rights reserved. Advance Access publication 16 March 2011

International Journal of Epidemiology 2011;**40**:755–764
doi:10.1093/ije/dyr036

Avoiding bias from weak instruments in Mendelian randomization studies

Stephen Burgess, Simon G Thompson and CRP CHD Genetics Collaboration[†]

MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Rosinson Way, Cambridge CB2 OSR, UK
E-mail: stephen.burgess@mrc-bsu.cam.ac.uk

[†] The members of the CRP CHD Genetics Collaboration are listed in the Supplementary Appendix 1 available as supplementary data at *IJE* online.

Accepted 8 February 2011

Background Mendelian randomization is used to test and estimate the magnitude of a causal effect of a phenotype on an outcome by using genetic variants as instrumental variables (IVs). Estimates of association from IV analysis are biased in the direction of the confounded, observational association between phenotype and outcome. The magnitude of the bias depends on the *F*-statistic for the strength of relationship between IVs and phenotype. We seek to develop guidelines for the design and analysis of Mendelian randomization studies to minimize bias.

Methods IV analysis was performed on simulated and real data to investigate the effect on bias of size of study, number and choice of instruments and method of analysis.

Results Bias is shown to increase as the expected *F*-statistic decreases, and can be reduced by using parsimonious models of genetic association (i.e. not over-parameterized) and by adjusting for measured covariates. Using data from a single study, the causal estimate of a unit increase in log-transformed C-reactive protein on fibrinogen (μmol/l) is shown to increase from −0.005 ($P = 0.99$) to 0.792 ($P = 0.00003$) due to injudicious choice of instrument. Moreover, when the observed *F*-statistic is larger than expected in a particular study, the causal estimate is more biased towards the observational association and its standard error is smaller. This correlation between causal estimate and standard error introduces a second source of bias into meta-analysis of Mendelian randomization studies. Bias can be alleviated in meta-analyses by using individual level data and by pooling genetic effects across studies.

Conclusions Weak instrument bias is of practical importance for the design and analysis of Mendelian randomization studies. *Post hoc* choice of instruments, genetic models or data based on measured *F*-statistics can exacerbate bias. In particular, the commonly cited rule of thumb that $F > 10$ avoids bias in IV analysis is misleading.

Keywords Mendelian randomization, instrumental variables, causal inference, weak instruments, bias, meta-analysis

Introduction

In observational studies, an association between outcome and a phenotype (a modifiable risk factor or exposure of interest) may not be causal. It may be due to confounding factors which affect both an individual's phenotype and outcome, or due to reverse causation where the outcome affects the phenotype.[1] Although we can measure known confounders, we can never be certain that all confounders have been identified and so cannot interpret an estimate of association from an observational study as a causal effect.[2]

Mendelian randomization uses genetic variants (G) which are associated with the phenotype (X) to estimate a causal effect.[3] The genetic variants must not be associated with any confounding factor (U) and not directly associated with the outcome (Y).[2,3] These assumptions define the genetic variants as an instrumental variable (IV)[4,5] and can be summarized graphically (Figure 1).[2] From the IV assumptions, and as genetic variation is determined at conception, the variation in X explained by G is independent of any confounding or reverse causation, and so any difference in Y associated with G indicates a causal effect of X on Y.[6] We assume throughout this article that the IV assumptions are satisfied by the instruments used.

As well as testing for a causal association, we can estimate its magnitude.[3] As genetic effects on phenotypes are typically small, Mendelian randomization estimates of association have much wider confidence intervals (CIs) than conventional epidemiological estimates.[7] The rationale for using Mendelian randomization is that an unbiased, imprecise estimate is preferable to a precise, biased estimate of causal association.[8]

Although IV techniques are asymptotically unbiased in the presence of confounding, IV estimates suffer from finite sample bias, known as weak instrument bias.[9–11] (Bias is the difference between the average estimated value of a parameter and its true value.) This bias is in the direction of the observational confounded association, and its magnitude depends on the strength of association between genetic instrument and phenotype.[12,13] Under certain conditions, the relative bias of the IV estimator to the observational estimator is ~1/*F*, where *F* is the *F*-statistic in the regression of X on G.[14] When *F* is 10, this means that the bias of the IV estimator is 10% of the bias of the observational estimator, leading to the 'rule of thumb' that the *F*-statistic should be at least 10 to avoid bias.[3,14]

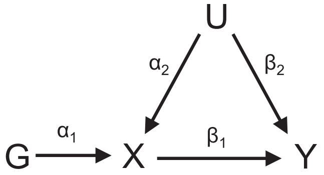

Figure 1 Directed acyclic graph of Mendelian randomization assumptions

In this article, we consider continuous outcomes; with binary outcomes, weak instrument bias affects the causal estimate in a similar way, although it is not the only bias present.[15] Unless otherwise stated, we use an additive per allele model. We use the two-stage least squares (2SLS)[16] and limited information maximum likelihood (LIML)[17] methods to calculate IV estimates. In 2SLS, we first regress the phenotype (x_i) on the genetic instruments ($g_{ik},\ k = 1, \dots, K$) where each instrument $g_{ik} = 0, 1, 2$ represents the number of variant alleles of single nucleotide polymorphism (SNP) k in individual i. We then regress the outcome (y_i) on the fitted values of phenotype from the first stage regression ($\hat{x}_i$).

$$\begin{aligned} x_i &= \alpha_0 + \sum_{k=1}^{K} \alpha_k g_{ik} + \epsilon_{xi} \\ y_i &= \beta_0 + \beta_{IV}\, \hat{x}_i + \epsilon_{yi} \end{aligned} \tag{1}$$

The causal estimate from the 2SLS method is $\hat{\beta}_{IV}$. (Note that although this gives the correct point estimate, the standard error (SE) is not correct; the use of 2SLS software is recommended for estimation in practice.[18]) The LIML estimator is the 'maximum likelihood counterpart of 2SLS'.[19] It is calculated by a maximum likelihood procedure on the assumption of homoscedastic errors. When there is a single instrument, these estimators coincide and equal the ratio (or Wald) estimate.[3]

We use data from the CRP CHD Genetics Collaboration[20] to estimate the causal association of C-reactive protein (CRP) on fibrinogen, which are both markers for inflammation. As the distribution of CRP is positively skewed, we take its logarithm and assume a linear association of log(CRP) on fibrinogen. Although log(CRP) and fibrinogen are highly positively correlated ($r = 0.45 - 0.55$ in the studies below), it is thought that long-term elevated levels of CRP are not causally associated with an increase in fibrinogen.[21]

Bias of IV estimates in small studies

As a motivating example, we consider the Copenhagen General Population Study (CGPS),[22] a cohort study with complete data on CRP from a high-sensitivity assay, fibrinogen and three SNPs from the CRP gene region (rs1205, rs1130864 and rs3093077) for 35 679 participants. We calculate the observational estimate [simply regressing fibrinogen on log(CRP)] and IV estimate of association using all three SNPs as instrumental variables. We then analyse the same data as if it came from multiple studies by dividing the study randomly into substudies of equal size, calculating estimates of association in each substudy and meta-analysing the results

Table 1 Estimates of effect (SE) of log(CRP) on fibrinogen (μmol/l) from CGPS ($N = 35\,679$)

Number of substudies	Observational estimate	2SLS IV estimate	LIML IV estimate	Mean *F*-statistic
1	1.6799 (0.0143)	−0.0468 (0.1510)	−0.0531 (0.1515)	152.0
5	1.6796 (0.0143)	−0.0092 (0.1478)	−0.0541 (0.1508)	31.44
10	1.6789 (0.0143)	0.0871 (0.1426)	−0.0068 (0.1485)	16.44
16	1.6781 (0.0143)	0.2300 (0.1372)	0.1641 (0.1426)	10.81
40	1.6761 (0.0143)	0.4562 (0.1266)	0.3093 (0.1385)	4.833
100	1.6713 (0.0142)	0.8279 (0.1078)	0.6575 (0.1279)	2.516
250	1.6695 (0.0141)	1.2711 (0.0826)	1.1796 (0.1022)	1.646

Estimates are divided randomly into substudies of equal size and combined using fixed-effect meta-analysis: observational estimate using unadjusted linear regression, IV estimate from Mendelian randomization using 2SLS and LIML methods. *F*-statistics from linear regression of log(CRP) on three genetic IVs averaged across substudies.

using a fixed-effect model. We divide into $5, 10, 16, 40, 100$ and 250 substudies.

We see from Table 1 that the observational estimate stays almost unchanged whether the data are analysed as one study or as several studies. However, the IV estimate increases from near zero until it approaches the observational estimate and the SE of the estimate decreases. We can see that even where the number of substudies is 16 and the average *F*-statistic is around 10, there is a serious bias with a positive causal estimate ($P = 0.09$ using 2SLS) despite the causal estimate with the data analysed as one study being near zero.

Why does weak instrument bias occur?

Although asymptotically the genetic variants are independent of confounders, confounders will not be perfectly balanced between genotypic subgroups in finite samples. If the instrument is strong, then the difference in phenotype between subgroups will be due to the genetic instrument, and the difference in outcome (if any) will be due to this difference in phenotype. However, if the instrument is weak, it explains little variation in the phenotype, the chance difference in confounders may explain more of the difference in phenotype between subgroups than the instrument. If the effect of the instrument is near zero, then the estimate of the 'causal association' approaches the association between phenotype and outcome caused by changes in the confounders, i.e. the observational confounded association.[12]

To illustrate the point algebraically, we take a phenotype X which depends linearly on an IV G and a confounder U, and an outcome Y which depends linearly on X and U. We assume that there are no other error terms in the model and that all the variability comes via U. The causal effect of X on Y is β_1.

$$\begin{aligned} X &= \mu_X + \alpha_1 G + \alpha_2 U \\ Y &= \mu_Y + \beta_1 X + \beta_2 U \end{aligned} \tag{2}$$

We assume that G is dichotomous, dividing the population into two genotypic subgroups labelled 1 and 2 such that $\alpha_1 > 0$. We let $\bar{x}_j$ be the average value for phenotype in subgroup j, similarly $\bar{y}_j$ and $\bar{u}_j$. An expression for the IV (ratio) estimate for the causal effect in this case depends on $\Delta = \bar{u}_2 - \bar{u}_1$:[11]

$$\beta_{\text{IV}} = \frac{\text{difference in outcome}}{\text{difference in phenotype}} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1} = \beta_1 + \frac{\beta_2 \Delta}{\alpha_1 + \alpha_2 \Delta} \tag{3}$$

We know that Δ has mean zero as G is an IV. Hence, β_{IV} tends to β_1 asymptotically as the sample size increases. If α_1 is large compared with $\alpha_2\Delta$, that is the proportion of variation explained by the instrument compared with that explained by the chance imbalance in confounders is large, then β_{IV} is close to β_1. However, if α_1 is small compared with $\alpha_2\Delta$, then whether Δ is positive or negative, it can be seen from (3) that the bias $\beta_{\text{IV}} - \beta_1$ tends to (β_2/α_2), which is the confounded association between the phenotype and outcome.[14]

Hence, the IV estimator is biased towards the observational confounded association, and the magnitude of the bias depends on the strength of association between X and G.[17]

How can we minimize weak instrument bias?

Increasing the *F*-statistic

The *F*-statistic is a measure of instrument strength. It is related to the proportion of variance in the phenotype explained by the genetic variants (R^2), sample size (n) and number of instruments (k) by the formula $F = (\frac{n-k-1}{k})(\frac{R^2}{1-R^2})$. It is sometimes known as the Cragg–Donald *F*-statistic.[23,24] The bias from weak instruments depends on the strength of the instrument through the *F*-statistic.[14,23] As the *F*-statistic depends on the sample size, then bias can be reduced by increasing sample size. Similarly, if there are instruments that are not contributing much to explaining

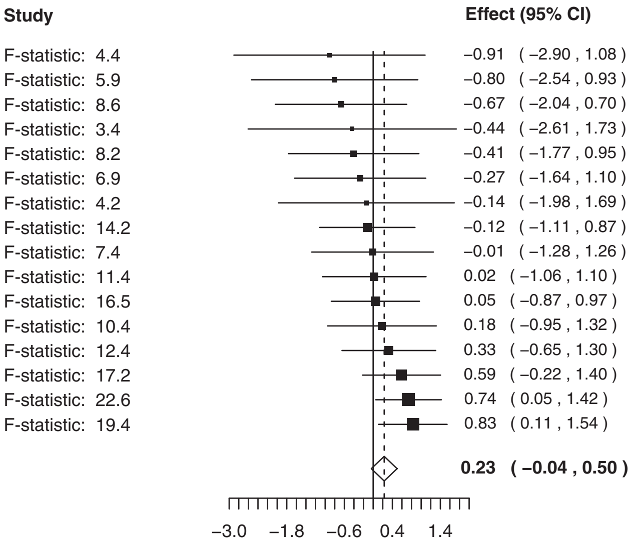

Figure 2 Forest plot of causal estimates of log(CRP) on fibrinogen (μmol/l) using data from CGPS divided randomly into 16 equally sized substudies (each $N \simeq 2230$). Studies ordered by causal estimate. F-statistic from regression of phenotype on three IV. Size of markers is proportional to weight in meta-analysis

the variation in the phenotype, then excluding these instruments will increase the F-value. In general, employing fewer degrees of freedom to model the genetic association, that is using parsimonious models, will increase the F-statistic and reduce weak instrument bias.[25] For example, an additive per allele or an additive haplotype model is more parsimonious than a model with one coefficient for each genotypic subgroup; provided the former does not misrepresent the data, bias will be reduced.

However, it is not enough to simply rely on an F-statistic measured from data to inform us about bias.[26] Returning to the previous example where we divided the CGPS study into 16 equally sized substudies with mean F-statistic 10.81, Figure 2 shows the forest plot of the estimates of these 16 substudies using the 2SLS method with their corresponding F-values. We see that the substudies which have greater estimates are the ones with higher F-values. The correlation between F-values and point estimates is 0.83 ($P < 0.001$). The substudies with higher F-values also have tighter CIs and so receive more weight in the meta-analysis. If we exclude from the meta-analysis substudies with an F-statistic <10, then the pooled estimate increases from 0.2300 (SE 0.1372, $P = 0.09$) to 0.4322 (SE 0.1574, $P = 0.006$). Equally, if we only use the IVs as instruments in each substudy with an F-statistic >10 when regressed in a univariate regression on the phenotype, then the pooled estimate increases to 0.2782 (SE 0.1470, $P = 0.06$). So neither of these approaches is useful in reducing bias.

Although the expectation of the F-statistic is a good indicator of bias, with low expected F-statistics indicating greater bias, the observed F-statistic shows considerable variation. In the 16 substudies of Figure 2, the F-statistic ranges from 3.4 to 22.6. From above, we see that the observed F-statistic will be large when the difference in phenotype between the two genotypic subgroups ($\bar{x}_2 - \bar{x}_1 = \alpha_1 + \alpha_2 \Delta$) is large. This occurs when Δ is large and positive and corresponds with a value of β_{IV} biased in the direction of the confounded estimate. The observed F-statistic will be small when $\alpha_1 + \alpha_2 \Delta$ is small. This occurs when Δ is negative and will often correspond with a value of β_{IV} biased in the opposite direction to the confounded estimate.

In more realistic examples, assuming similar instruments in each study, larger studies would have higher expected F-statistics due to sample size which would correspond to truly stronger instruments and less bias. However, the sampling variation of causal effects and observed F-statistics in each study would still tend to follow the pattern of Figure 2, with larger observed F-statistics corresponding to more biased causal estimates.

So although it is desirable to use strong instruments, the measured strength of instruments in data is not a good guide to the true instrument strength. Any guidance that relies on providing a threshold, such as excluding studies from a meta-analysis if $F < 10$, is flawed and may introduce more bias than it prevents.

Choice of instruments

Including more instruments, where each instrument explains extra variation in the phenotype, should give more information on the causal parameter. In order to investigate how using more instruments affects bias in the IV estimator, we perform 100 000 simulations in a model where for each participant indexed by i, the phenotype x_i depends linearly on six dichotomous genetic instruments ($g_{ik} = 0$ or 1, $k = 1, \ldots, 6$), a normally distributed confounder u_i and an independent normally distributed error term ε_{xi}. Outcome y_i is a linear combination of phenotype, confounder and an independent error term ε_{yi}.

$$\begin{gathered} x_i = \sum_{k=1}^{6} \alpha_{1k} g_{ik} + \alpha_2 u_i + \epsilon_{xi} \\ y_i = \beta_1 x_i + \beta_2 u_i + \epsilon_{yi} \\ u_i, \epsilon_{xi}, \epsilon_{yi} \sim \mathcal{N}(0, 1) \text{ independently} \end{gathered} \tag{4}$$

We set $\beta_1 = 0$, $\alpha_2 = 1$, $\beta_2 = 1$ so that X is observationally strongly positively associated with Y, but there is a null causal association. We draw parameters for the genetic association α_{1k} randomly from a uniform distribution on 0.15 to 0.25 independently for each genetic instrument k, corresponding to mean F-values from 6.8 to 16.3. We use a sample size of 2048 divided equally between the $2^6 = 64$ genotypic

Table 2 Median and 95% range of bias using 2SLS and LIML methods

	Median 2.5–97.5% quantiles				
	2SLS		LIML		Mean F-statistic
Estimate using 1 IV		0.0005	−1.1996 to 0.5629		11.2
2 IVs	0.0242	−0.5453 to 0.4007	−0.0002	−0.6529 to 0.3946	11.2
3 IVs	0.0310	−0.3861 to 0.3367	−0.0004	−0.4805 to 0.3247	11.3
4 IVs	0.0343	−0.3092 to 0.2990	−0.0003	−0.3943 to 0.2832	11.4
5 IVs	0.0361	−0.2622 to 0.2731	−0.0002	−0.3416 to 0.2545	11.4
6 IVs	0.0373	−0.2298 to 0.2531	0.0001	−0.3059 to 0.2328	11.5
IV with greatest F		0.1249	−0.2103 to 0.3645		21.2
IV with least F		−0.3028	−3.1756 to 0.7737		3.7
IVs with $F>10$	0.0923	−0.2103 to 0.3645	0.0779	−0.2291 to 0.3612	17.0

Mean F-statistic across 100 000 simulations using combinations of six uncorrelated instruments are provided.

subgroups. The instruments are uncorrelated, so that variation explained by each of the instruments is independent, and the mean F-values do not depend greatly on the number of IVs (mean 11.2 using 1 IV, 11.5 using 6 IVs).

Table 2 shows the median and 95% range of the estimates of bias from the 2SLS and LIML methods using all combinations of all numbers of IVs as the instrument, with the mean across simulations of the F-statistic for all the instruments used. We also give results using the IV with the greatest and lowest F-values, as well as using all IVs with an F-statistic greater than 10 in univariate regression of phenotype.

As the number of instruments increases, while the variance of estimates decreases, using 2SLS, the bias increases, despite the mean F-value remaining constant. This is because there is a greater risk of imbalances in confounders between the greater number of genetic subgroups defined by the instruments. The greatest increase in median bias is from one instrument to two instruments, and coincides with the greatest increase in precision. With LIML, a similar increase in precision is observed, but no increase in bias. Using the single IV with the greatest F gives markedly biased results, despite a mean F-value of 21.2. There is a similar bias only using IVs with $F>10$.

For 2SLS, the mean bias is similar to the median presented, except in the case of a single IV where the theoretical mean is infinite.[12] For LIML, the mean bias is infinite for all numbers of IVs.[27]

As a further illustration, we consider the Framingham Heart Study (FHS), a cohort study measuring CRP and fibrinogen at baseline with complete data for nine SNPs on the CRP gene for 1500 participants. The observational estimate of the log(CRP)–fibrinogen (μmol/l) association is 1.134 (95% CI 1.052–1.217). We calculate the causal estimate of the association using the 2SLS method with different numbers of SNPs as an instrument. Figure 3 shows a plot of the IV estimates against number of instruments, where each point represents $\hat{\beta}_{\text{IV}}$ calculated using a different combination of SNPs. The range of values of $\hat{\beta}_{\text{IV}}$ reduces as we include more instruments, but the median causal estimate across the different combinations of IVs increases. The 2SLS estimate using all nine SNPs in an additive per allele model is −0.005 (95% CI −0.721 to 0.711, $P=0.99$, $F_{9,1490}=3.34$). If we relax the genetic assumptions of a per allele model and additivity between SNPs to instead use a model with one coefficient for each of the 49 genotypes represented in the data, the 2SLS estimate is 0.792 (95% CI 0.423–1.161, $P=0.00003$, $F_{48,1451}=1.66$). Using LIML, the estimate is 0.052 (95% CI −0.706 to 0.809, $P=0.89$).

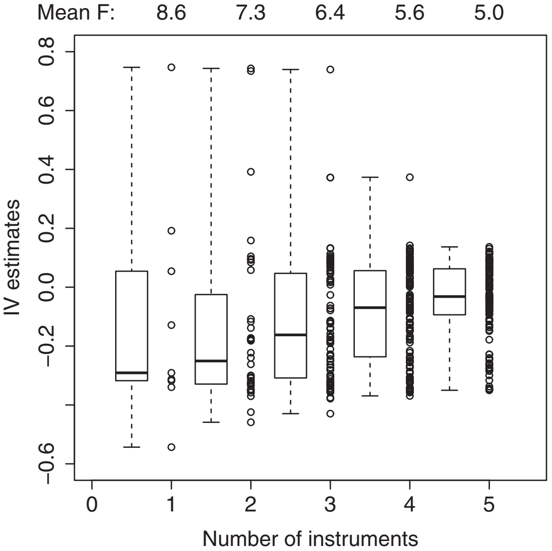

Figure 3 IV estimates for causal association in FHS of log(CRP) on fibrinogen (μmol/l) using all combinations of varying numbers of SNPs as instruments. Point estimates, associated box plots (median, interquartile range, range) and mean F-statistics across combinations

Table 3 Bias of the IV estimator, median and interquartile (IQ) range across simulations from model (5)

	Not adjusted			Adjusted		
α_1	Mean F	Median bias	IQ range	Partial F	Median bias	IQ range
0.05	1.05	0.6418	−0.1026 to 1.3859	1.58	0.4659	−0.3830 to 1.3138
0.15	1.39	0.4573	−0.2408 to 1.1406	2.09	0.2916	−0.4442 to 0.9776
0.25	2.06	0.2478	−0.3819 to 0.7446	3.09	0.1290	−0.4535 to 0.5949
0.35	3.08	0.1110	−0.4282 to 0.4821	4.62	0.0460	−0.4104 to 0.3883
0.45	4.42	0.0412	−0.4122 to 0.3414	6.63	0.0115	−0.3468 to 0.2819
0.55	6.11	0.0138	−0.3620 to 0.2691	9.16	0.0030	−0.2822 to 0.2277

Bias for different strengths of instrument without and with adjustment for confounder is provided.

Our conclusions from these analyses are as follows. Whereas including more genetic IVs will increase precision, it may also increase bias. Bias is exacerbated if some of the included IVs are weak, but cannot be avoided by data-driven selection of instruments. In addition, using over-parameterized models of genetic association with 2SLS is likely to increase bias.[28]

Adjustment for measured covariates

If we can find measured covariates which explain variation in the phenotype, and which are not on the causal pathway between phenotype and outcome, then we can incorporate such covariates in our model. This will increase precision and reduce weak instrument bias. Precision will be further increased if these covariates can be used to explain variation in the outcome.

To exemplify this, we perform 100 000 simulations in a model similar to (4), but with a single IV and with two separate terms accounting for confounding between X and Y, corresponding to measured (V) and unmeasured (U) confounders.

$$\begin{aligned} x_i &= \alpha_1 g_i + \alpha_2 u_i + \alpha_2 v_i + \epsilon_{xi} \\ y_i &= \beta_1 x_i + \beta_2 u_i + \beta_2 v_i + \epsilon_{yi} \\ u_i, v_i, \epsilon_{xi}, \epsilon_{yi} &\sim \mathcal{N}(0,1) \text{ independently} \end{aligned} \tag{5}$$

We again set $\beta_1 = 0$, $\alpha_2 = 1$, $\beta_2 = 1$ and vary the parameter for the genetic association α_1 from 0.05 to 0.55, corresponding to mean F-values from 1.05 to 6.11. We use a sample size of 200 equally divided between two genotypic groups, $g_i = 0$, 1. We calculate an estimate of causal association from the 2SLS method, both with and without adjustment for V in the G-X and $\hat{X}$-Y regressions. R^2 in the regression of X on V is 33%. The relevant measure of instrument strength with a measured confounder is the partial F-statistic for G in the regression of X on G and V.[29] Table 3 shows that adjustment for measured covariates increases the F-statistic and decreases the median bias of the IV estimator. For stronger instruments, we also see a reduction in the variability of the estimator.

As an example, we consider data on interleukin-6 (IL6), a cytokine which is involved in the inflammation process upstream of CRP and fibrinogen.[30] Elevated levels of IL6 lead to elevated levels of both CRP and fibrinogen, so IL6 is correlated with short-term variation in CRP,[31] but is independent of underlying genetic variation in CRP.[21] We assume that it is a confounder in the association of CRP with fibrinogen and not on the causal pathway (if such a pathway exists). As IL6 has a positively skewed distribution, we take its logarithm. The Cardiovascular Health Study (CHS) is a cohort study measuring CRP, IL6 and fibrinogen at baseline, as well as three SNPs on the CRP gene, with complete data for 4137 subjects. The proportion of variation in log(CRP) explained in the data by log(IL6) is 26%. We calculate the causal estimate of the CRP–fibrinogen association for each SNP separately and for all the SNPs together in an additive per allele model, both without and with adjustment for log(IL6) in the first- and second-stage regressions. Results are given in Table 4. We see that after adjusting for log(IL6) the causal estimate in each case has decreased, its SE has reduced and the F-statistic has increased. This indicates that both weak instrument bias has been reduced and precision has been improved.

Borrowing information across studies

The IV estimator would be unbiased if we knew the true values for the average phenotype in different genotypic groups. In a meta-analysis context,[32] we can combine the estimates of genotype–phenotype association from different studies to give more precise estimates of phenotype levels in each genetic group. In the 2SLS method, an individual participant data (IPD) meta-analysis for data on individual i in study m with phenotype x_{im}, outcome y_{im} and g_{ikm} for number of alleles of genetic variant k ($k = 1, 2, \ldots K$) is:

$$\begin{aligned} x_{im} &= \alpha_{0m} + \sum_{k=1}^{K} \alpha_{km} g_{ikm} + \epsilon_{xim} \\ y_{im} &= \beta_{0m} + \beta_1 \hat{x}_{im} + \epsilon_{yim} \\ \epsilon_{xim} &\sim \mathcal{N}(0, \sigma_x^2);\ \epsilon_{yim} \sim \mathcal{N}(0, \sigma_y^2) \text{ independently} \end{aligned} \tag{6}$$

Table 4 Estimate and SE of IV estimator for causal effect of log(CRP) on fibrinogen and F-statistic for regression of log(CRP) on IVs calculated using each SNP separately

	Not adjusted		Adjusted	
IV estimate	Estimate (SE)	F-statistic	Estimate (SE)	Partial F
Using rs1205	0.219 (0.201)	79.6	0.173 (0.196)	100.2
Using rs1417938	−0.457 (0.407)	27.6	−0.458 (0.362)	37.2
Using rs1800947	0.354 (0.325)	28.6	0.324 (0.316)	36.5
Using all SNPs	0.186 (0.194)	24.4	0.127 (0.188)	32.2

All SNPs together in additive per allele model, adjusting with and without adjustment for log(IL6) in CHS.

Table 5 Estimates of effect of log(CRP) on fibrinogen (μmol/l) from each of five studies separately and from meta-analysis of studies

Study	Participants	Causal estimate	95% CI	F-statistic	df	Observational estimate (SE)
CCHS	7999	−0.286	−1.017 to 0.445	29.6	(3,7995)	1.998 (0.030)
EAS	650	0.745	0.113 to 1.396	6.9	(3,646)	1.115 (0.056)
HPFS	405	0.758	−0.071 to 1.587	5.3	(3,401)	1.048 (0.081)
NHS	385	−0.906	−2.154 to 0.341	6.1	(3,381)	0.562 (0.114)
SHEEP	1044	0.088	−0.588 to 0.763	10.5	(3,1040)	1.078 (0.051)
Different genetic effects		0.021	−0.362 to 0.403	14.4	(15, 10463)	
Common genetic effects		−0.093	−0.534 to 0.348	56.6	(3, 10475)	
Summary estimates		0.234	−0.107 to 0.575			

Studies included number of participants, causal estimates using 2SLS with 95% CI, F-statistic with degrees of freedom (df) from additive per allele regression of phenotype on SNPs used as IVs, observational estimate (SE). Meta-analyses conducted using IPD with different study-specific genetic effects, common pooled genetic effects and using summary estimates with inverse-variance weighting.

The phenotype levels are regressed on the instruments using a per allele additive linear model separately in each study, and then the outcome levels are regressed on the fitted values of phenotype ($\hat{x}_{im}$). The terms α_{0m} and β_{0m} are study-specific intercept terms. Here, we assume homogeneity of variances across studies; we can use generalized method of moments (GMM)[16] or Bayesian methods[33] to allow for possible heterogeneity.

If the same genetic variants are measured and assumed to have the same effect on the phenotype in each study, we can use common genetic effects (i.e. $\alpha_{km} = \alpha_k$) across studies by replacing the first line in model (6) with:

$$x_{im} = \alpha_{0m} + \sum_{k=1}^{K} \alpha_k g_{ikm} + \epsilon_{xim} \tag{7}$$

where the coefficients α_k are the same in each study. If the assumption of fixed genetic effects is correct, this will improve the precision of the $\hat{x}_{im}$ and reduce weak instrument bias. Model (7) can be used even if, for example, the phenotype is not measured in one study, under the assumption that the data are missing at random (MAR).[34]

To illustrate, we consider the Copenhagen City Heart Study (CCHS), Edinburgh Artery Study (EAS), Health Professionals Follow-up Study (HPFS), Nurses Health Study (NHS) and Stockholm Heart Epidemiology Program (SHEEP), which are cohort studies or case–control studies measuring CRP and fibrinogen levels at baseline.[20] In case–control studies, we use the data from controls alone since these better represent cross-sectional population studies. These five studies measured three SNPs: rs1205, rs1130864 and rs3093077 (or rs3093064, which is in complete linkage disequilibrium with rs3093077). We estimate the causal association using the 2SLS method with different genetic effects (Model 6), common genetic effects (Model 7) and by a fixed-effect meta-analysis of summary estimates from each study.

Table 5 shows that the studies analysed separately have apparently disparate causal estimates with wide CIs. The meta-analysis estimate assuming common genetic effects across studies is further from the confounded observational estimate and closer to the estimate from the largest study with the strongest instruments (CCHS) than the model with different genetic effects, suggesting that the latter suffers bias from weak instruments.

Table 6 Estimates of causal effect (SE) of log(CRP) on fibrinogen from CGPS study

Substudies	Summary	*P*-value	IPD different genetic effects	*P*-value	IPD common genetic effects	*P*-value
1	−0.0468 (0.1510)	0.76				
5	−0.0092 (0.1478)	0.95	−0.0273 (0.1479)	0.85	−0.0473 (0.1511)	0.75
10	0.0871 (0.1426)	0.54	0.0370 (0.1430)	0.80	−0.0457 (0.1510)	0.76
16	0.2300 (0.1372)	0.09	0.1530 (0.1372)	0.26	−0.0482 (0.1512)	0.75
40	0.4562 (0.1266)	<0.001	0.2986 (0.1272)	0.02	−0.0433 (0.1511)	0.77
100	0.8279 (0.1078)	<0.001	0.6782 (0.1056)	<0.001	−0.0450 (0.1506)	0.77
250	1.2711 (0.0826)	<0.001	1.1499 (0.0793)	<0.001	−0.0413 (0.1505)	0.78

Estimates divided randomly into substudies and combined: using 2SLS summary study estimates by fixed-effect meta-analysis, using individual patient data (IPD) with different and common genetic effects across substudies.

The estimate from meta-analysis of study-specific causal estimates is greater than that from meta-analysis using the individual participant data. Although the CCHS study has about 8 times the number of participants as SHEEP and 12 times as many as EAS, its causal estimate has a larger SE. The standard errors in the 2SLS method, calculated by sandwich variance estimators using strong asymptotic assumptions, are known to be underestimated, especially with weak instruments.[35] Also, Figure 2 shows that causal estimates nearer to the observational association have lower variance. So a meta-analysis of summary outcomes may be biased due to overestimated weights in the studies with more biased estimates.

In the example at the beginning of the article, if we use the IPD data to combine the substudies in the meta-analysis rather than combining summary estimates, then Table 6 shows that the pooled estimates are less biased. If we additionally assume common genetic effects across studies, then we recover close to the original estimate based on analysing the full dataset as one study and weak instrument bias has been eliminated.

Discussion

Our conclusions from the investigations in this article are summarized in the box of key messages, and amplified below.

This article exemplifies the effect of weak instrument bias on causal estimates in real and simulated data. We have seen how the magnitude of the bias depends on the instrument strength through the mean or expected F-statistic, with lower mean F-statistics corresponding to greater bias. However, a novel finding is that, for a study of fixed size and underlying instrument strength, an observed F-statistic greater than the expected F-value corresponds to an estimate closer to the observational association with greater precision; conversely, an observed F-statistic less than the expected F-value corresponds with an estimate further from the observational association with less precision. So simply relying on an F-statistic from an individual study is over-simplistic and simple threshold rules such as ensuring $F > 10$ may cause more bias than they prevent.

Using the 2SLS method, we demonstrated a bias-variance trade-off for number of instruments used in IV estimation. For a fixed mean F-statistic, as the number of instruments increases, the precision of the IV estimator increases, and the bias also increases. Using the LIML method, bias did not increase with the number of instruments. Nevertheless, we seek parsimonious models of genetic association, for example using additive per allele effects and including only the most important IVs, based on biological knowledge and external information. Provided the data are not misrepresented, these should provide the best estimates of causal association. It is also possible to summarize multiple SNPs using a gene score.[25] If this is done using pre-specified weights, this makes strong assumptions about the effects of different SNPs which may itself introduce bias. The use of a data-derived weighted gene score is equivalent to 2SLS.[36] Again, *post hoc* use of F-statistics to choose between instruments may cause more bias than it prevents.

Ideally, issues of weak instrument bias should be addressed prior to data collection, by specifying sample sizes, instruments and genetic models using the best prior evidence available to ensure that the expected values of F-statistics are large. Where this is not possible, our advice would be to conduct sensitivity analyses using different IV methods, numbers of instruments and genetic models to investigate the impact of different assumptions on the causal estimate.

Generally, the LIML estimate is less biased than the 2SLS estimate. Difference between the 2SLS and LIML IV estimates is the evidence of possible bias from weak instruments. When a single instrument is used, the IV estimate is close to median unbiased. Although using one instrument will result in an estimator with greater variance, it will on average be less biased.

Another technique that helps reduce weak instrument bias is adjustment for covariates. Including predictors of the phenotype in the first-stage regression, or predictors of the outcome in the second-stage regression, increases precision of the causal estimate. The former will also increase the *F*-statistic for the genetic IVs, and thus reduce weak instrument bias.

In a meta-analysis context, bias is a more serious issue, as it arises not only from the bias in the individual studies but also from the correlation between causal effect size and variance which results in studies with effects closer to the observational estimate being over-weighted. By using a single IPD model, we reduce the second source of bias. Additionally, we can pool information on the genetic association across studies to strengthen the instruments. The assumptions of homogeneity of variances and common genetic effects across studies will often be overly restrictive. Allowing for heterogeneity across studies in phenotype variance, genetic effects and in the causal effects themselves is possible in a Bayesian framework.[33]

Supplementary data

Supplementary data are available at *IJE* online.

Funding

Both authors are supported by the UK Medical Research Council (grant U.1052.00.001)

Conflict of interest: None declared.

KEY MESSAGES

- Bias from weak instruments can result in seriously misleading estimates of causal effects. Studies with instruments having high mean *F*-statistics are less biased on average. However, if a study by chance has a higher *F*-statistic than expected, then the causal estimate will be more biased.
- Data-driven choice of instruments or analysis can exacerbate bias. In particular, any guideline such as $F > 10$ is misleading. Methods, instruments and data to be used should be specified prior to data analysis. Meta-analysis based on summary study-specific estimates of causal association are susceptible to bias.
- Bias can be alleviated in a single study by using the LIML rather than 2SLS method and by adjusting for measured confounders, and in a meta-analysis by using IPD modelling. We advocate parsimonious modelling of the genetic association (e.g. per allele additive SNP model rather than one coefficient per genotype). This should be accompanied by sensitivity analyses to assess potential bias.

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] Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res* 2007;**16:**309–30.

[3] Lawlor D, Harbord R, Sterne J, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* 2008;**27:**1133–63.

[4] Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000;**29:**722–29.

[5] Wehby G, Ohsfeldt R, Murray J. 'Mendelian randomization' equals instrumental variable analysis with genetic instruments. *Stat Med* 2008;**27:**2745–49.

[6] Davey Smith G, Ebrahim S. What can Mendelian randomisation tell us about modifiable behavioural and environmental exposures? *BMJ* 2005;**330:**1076–79.

[7] Davey Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol* 2004;**33:**30–42.

[8] Bautista LE, Smeeth L, Hingorani AD, Casas JP. Estimation of bias in nongenetic observational studies using ''Mendelian triangulation''. *Ann Epidemiol* 2006;**16:**675–80.

[9] Richardson DH. The exact distribution of a structural coefficient estimator. *J Am Stat Assoc* 1968;**63:**1214–26.

[10] Sawa T. The exact sampling distribution of ordinary least squares and two-stage least squares estimators. *J Am Stat Assoc* 1969;**64:**923–37.

[11] Nelson C, Startz R. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *J Bus* 1990;**63:**125–40.

[12] Bound J, Jaeger D, Baker R. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J Am Stat Assoc* 1995;**90:**443–50.

[13] Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology* 2006;**17:**260–67.

[14] Staiger D, Stock J. Instrumental variables regression with weak instruments. *Econometrica* 1997;**65:**557–86.

[15] Didelez V, Meng S, Sheehan NA. Assumptions of IV methods for observational epidemiology. *Stat Sci* 2010; **25:**22–40.

[16] Baum C, Schaffer M, Stillman S. Instrumental variables and GMM: estimation and testing. *Stata J* 2003;**3:**1–31.

[17] Davidson R, MacKinnon JG. *Estimation and Inference in Econometrics*. USA: Oxford University Press, 1993.

[18] Angrist JD, Pischke JS. Instrumental variables in action: sometimes you get what you need. *Mostly Harmless Econometrics: an Empiricist's Companion*. Princeton, USA: Princeton University Press, 2009.

[19] Hayashi F. *Econometrics*. Princeton, USA: Princeton University Press, 2000.

[20] CRP CHD Genetics Collaboration. Collaborative pooled analysis of data on C-reactive protein gene variants and coronary disease: judging causality by Mendelian randomisation. *Eur J Epidemiol* 2008;**23:**531–40.

[21] CRP CHD Genetics Collaboration. Association between C reactive protein and coronary heart disease: Mendelian randomisation analysis based on individual participant data. *BMJ* 2011;**342:**d548.

[22] Zacho J, Tybjaerg-Hansen A, Jensen JS, Grande P, Sillesen H, Nordestgaard BG. Genetically elevated C-reactive protein and ischemic vascular disease. *N Engl J Med* 2008;**359:**1897–908.

[23] Stock J, Wright J, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. *J Bus Econom Stat* 2002;**20:**518–29.

[24] Baum CF, Schaffer ME, Stillman S. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *Stata J* 2007;**7:**465–506.

[25] Pierce BL, Ahsan H, VanderWeele TJ. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int J Epidemiol* 2011; **40:**740–52.

[26] Hall AR, Rudebusch GD, Wilcox DW. Judging instrument relevance in instrumental variables estimation. *Int Econom Rev* 1996;**37:**283–98.

[27] Hahn J, Hausman JA, Kuersteiner GM. Estimation with weak instruments: accuracy of higher-order bias and MSE approximations. *Econometrics J* 2004;**7:**272–306.

[28] Zohoori N, Savitz DA. Econometric approaches to epidemiologic data: relating endogeneity and unobserved heterogeneity to confounding. *Ann Epidemiol* 1997;**7:** 251–57.

[29] Shea J. Instrument relevance in multivariate linear models: a simple measure. *Rev Econom Stat* 1997;**79:** 348–52.

[30] Hansson GK. Inflammation, atherosclerosis, and coronary artery disease. *N Engl J Med* 2005;**352:**1685–95.

[31] Kaptoge S, Di Angelantonio E, Lowe G *et al*. Emerging Risk Factors Collaboration. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *Lancet* 2010;**375:**132–40.

[32] Thompson J, Minelli C, Abrams K, Tobin M, Riley R. Meta-analysis of genetic studies using Mendelian randomization–a multivariate approach. *Stat Med* 2005;**24:** 2241–54.

[33] Burgess S, Thompson SG. CRP CHD Genetics Collaboration. Bayesian methods for meta-analysis of causal relationships estimated using genetic instrumental variables. *Stat Med* 2010;**29:**1298–311.

[34] Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd edn. New York: Wiley, 2002.

[35] Stock J, Yogo M. Testing for weak instruments in linear IV regression. [Working Paper Series.] SSRN eLib 2002;**11:**T0284.

[36] Baum CF. An introduction to modern econometrics using Stata (p188). College Station, USA: Stata Corp, 2006.