Bias due to participant overlap in two-sample Mendelian randomization

Stephen Burgess [*]

Department of Public Health and Primary Care, University of Cambridge

Neil M. Davies

MRC Integrative Epidemiology Unit, University of Bristol

School of Social and Community Medicine, University of Bristol

Simon G. Thompson

Department of Public Health and Primary Care, University of Cambridge

June 30, 2016

Running title: Bias due to overlap in two-sample Mendelian randomization (48 chars).

[*] Corresponding author. Address: Department of Public Health & Primary Care, Strangeways Research Laboratory, 2 Worts Causeway, Cambridge, CB1 8RN, UK. Telephone: +44 1223 748651. Correspondence to: sb452@medschl.cam.ac.uk.

Abstract

Mendelian randomization analyses are often performed using summarized data. The causal estimate from a one-sample analysis (in which data are taken from a single data source) with weak instrumental variables is biased in the direction of the observational association between the risk factor and outcome, whereas the estimate from a two-sample analysis (in which data on the risk factor and outcome are taken from non-overlapping datasets) is less biased and any bias is in the direction of the null. When using genetic consortia which have partially overlapping sets of participants, the direction and extent of bias are uncertain. In this paper, we perform simulation studies to investigate the magnitude of bias and Type 1 error rate inflation arising from sample overlap. We consider both a continuous outcome and a case-control setting with a binary outcome. For a continuous outcome, bias due to sample overlap is a linear function of the proportion of overlap between the samples. So, in the case of a null causal effect, if the relative bias of the one-sample instrumental variable estimate is 10% (corresponding to an F parameter of 10), then the relative bias with 50% sample overlap is 5%, and with 30% sample overlap is 3%. In a case-control setting, if risk factor measurements are only included for the control participants, unbiased estimates are obtained even in a one-sample setting. However, if risk factor data on both control and case participants are used, then bias is similar with a binary outcome as with a continuous outcome. Consortia releasing publicly-available data on the associations of genetic variants with continuous risk factors should provide estimates that exclude case participants from case-control samples.

Key words: Mendelian randomization, instrumental variables, summarized data, aggregated data, weak instrument bias.

1 Introduction

Mendelian randomization is the use of genetic variants as instrumental variables to assess and estimate the causal effect of a risk factor on an outcome from observational data [Davey Smith and Ebrahim, 2003; Burgess and Thompson, 2015]. A recent methodological development in Mendelian randomization is the use of summarized data on associations of genetic variants with the risk factor and with the outcome to obtain causal effect estimates [Johnson, 2013; Burgess et al., 2013]. These summarized data comprise the associations of the individual genetic variants with the risk factor and with the outcome taken from univariable regression analyses (beta-coefficients and standard errors from linear or logistic regression as appropriate). Suitable summarized data for such analyses have been made publicly available for hundreds of thousands of genetic variants by some large consortia [Burgess et al., 2015a]. Examples include associations of genetic variants with lipid fractions from the Global Lipids Genetics Consortium [The Global Lipids Genetics Consortium, 2013] and with type 2 diabetes from the DIAGRAM consortium [Morris et al., 2012]. Mendelian randomization analyses using summarized data have suggested causal effects of adiponectin on type 2 diabetes risk [Dastani et al., 2012], insulin levels on endometrial cancer risk [Nead et al., 2015], and telomere length on risk of lung adenocarcinoma [Zhang et al., 2015].

The validity of a Mendelian randomization investigation depends on the instrumental variable (IV) assumptions being satisfied for all genetic variants [Lawlor et al., 2008]. In particular, any genetic variant used as an IV is assumed to be:

1. associated with the risk factor;
2. independent of confounders of the risk factor–outcome association; and
3. independent of the outcome conditional on the risk factor and confounders of the risk factor–outcome association [Greenland, 2000; Sussman et al., 2010].

In this paper, we assume that the IV assumptions are satisfied for all genetic variants in the analysis. Variants can be weak (that is, they do not explain much variation in the risk factor), but they are all assumed to be valid IVs.

Although estimates from IV analysis in a one-sample setting (that is, genetic variants, risk factor and outcome are all measured in the same participants) are asymptotically unbiased, they can have substantial bias in finite samples [Stock et al., 2002; Burgess and Thompson, 2011]. This bias (known as weak instrument bias) acts in the direction of the confounded observational association between the risk factor and outcome [Nelson and Startz, 1990; Bound et al., 1995]. Its magnitude depends on the strength of association between the IV(s) and the risk factor [Staiger and Stock, 1997]. Weak instrument bias also leads to inflated Type 1 error rates (over-rejection of the null) [Stock and Yogo, 2002]. One way of combatting weak instrument bias in practice is a two-sample analysis strategy, in which the associations of IVs with the risk factor and with the outcome are obtained from two non-overlapping datasets [Angrist and Krueger, 1992; Inoue and Solon, 2010]. In a two-sample Mendelian randomization analysis, any bias due to weak instruments is in the direction of the null [Pierce and Burgess, 2013]. Bias in the direction of the null is less serious than bias in the direction of the observational association, as it is conservative, and so will not lead to inflated Type 1 error rates and false positive findings. This bias may lead to lower power to detect a causal effect and increased probability of a Type 2 error, although the standard errors typically also attenuate, mitigating this somewhat [Burgess et al., 2016b].

The use of summarized data in Mendelian randomization is motivated by the increasing availability of suitable data in large sample sizes. A fortuitous side-effect is that genetic associations with the risk factor and with the outcome are often obtained from separate datasets, leading to a two-sample IV analysis. However, due to the nature of major international genetics consortia, often the datasets are not completely disjoint, and some studies and participants are in common between the two datasets. When there is some overlap, it is unclear whether bias due to weak instruments would be in the direction of the null (as in the case of zero overlap) or in the direction of the observational association (as in the case of complete overlap). While the sliding scale of bias towards the observational association as the proportion of overlap increases has previously been demonstrated [Pierce and Burgess, 2013], it is unclear in specific investigations whether this bias should be of concern.

The aim of this paper is to investigate the direction and degree of bias in a "two-sample" instrumental variable analysis in which there is overlap between the two samples. This is achieved by theoretical considerations and a series of simulation studies with realistic parameters for Mendelian randomization. The structure of the paper is as follows. First, we explain why weak instrument bias occurs and the reason for the direction of the bias (Section 2). Next, we present simulation studies and discuss their results (Section 3 for a continuous outcome, Section 4 for a binary outcome). We present illustrations of the potential bias due to participant overlap in example Mendelian randomization investigations using summarized data from two large consortia (Section 5). Finally, we discuss the wider relevance of these results, and in particular the case of a two-sample investigation in which the genetic variants were discovered in one of the datasets under analysis (Section 6).

2 Weak instrument bias

We initially assume that the outcome is continuous and that all regression analyses use a linear model. In the simplest case with a single IV, the causal effect of the risk factor on the outcome can be estimated as a ratio of regression coefficients [Martens et al., 2006]. The ratio estimate is the coefficient from regression of the outcome on the IV divided by the coefficient from regression of the risk factor on the IV [Didelez and Sheehan, 2007]. In a one-sample setting, weak instrument bias arises due to correlation between the regression coefficients in the numerator and denominator as a result of confounding between the risk factor and outcome [Nelson and Startz, 1990]. In a two-sample setting, the numerator and denominator in the ratio method will be uncorrelated. Bias with a single IV is difficult to quantify, as the expected value of the ratio estimate is undefined, because of the small but finite probability that the denominator in the ratio estimator (the IV–risk factor association) is arbitrarily close to zero [Hahn et al., 2004]. However, if the IV–risk factor association is close to zero, then an IV analysis is unlikely to be performed in practice, as the first IV assumption (the only one that can be tested directly) appears to be violated. In simulation studies, the median ratio estimate across simulations is usually close to the true value of the causal effect even with complete sample overlap (a one-sample analysis) except in the case of extremely weak instruments (those for which the expected strength of association with the risk factor corresponds to a p-value above 0.05 [Burgess and Thompson, 2011]), indicating that the practical consequences of bias with a single IV are unlikely to be substantial [Angrist and Pischke, 2009].

2.1 Two-stage least squares method

When there are multiple IVs, the two-stage least squares (2SLS) estimate can be obtained in a one-sample setting with individual-level data by i) regressing the risk factor on the IVs (first-stage regression), and then ii) regressing the outcome on fitted values of the risk factor from the first-stage regression (second-stage regression). Weak instrument bias can be explained as arising from overfitting in the first-stage regression model, which occurs at least in part due to chance correlations of the IVs with confounders [Burgess and Thompson, 2011]. In a one-sample setting, the fitted values from the first-stage regression are therefore correlated with the outcome in finite samples even in the absence of a causal effect. This leads to finite-sample bias of the 2SLS estimate. The expected magnitude of this bias depends on the strength of association between the IVs and the risk factor through the concentration parameter. The concentration parameter (μ) is related to the expected value of the F statistic from the regression of the risk factor on the IVs: for large samples, $\mathbb{E}(F) = \frac{\mu}{K} + 1$, where K is the number of IVs [Cragg and Donald, 1993]. We refer to the expected value of this F statistic as the F parameter, to emphasize that this is a characteristic of the population, and not simply a function of the observed data.

In a two-sample setting with individual-level data, the 2SLS estimate can be calculated by obtaining estimates of the first-stage regression parameters in one dataset, and constructing fitted values of the risk factor in the second dataset using these estimates and the values of the IVs in the second dataset (measurements of the risk factor in the second dataset are not required). The outcome and the fitted values of the risk factor in the second-stage regression are no longer correlated due to confounding. This approach is also known as split-sample 2SLS [Angrist and Krueger, 1995]. The fitted values of the risk factor are also equivalent (up to an additive constant) to values of an externally-weighted allele score (an analysis approach used in Mendelian randomization [Burgess and Thompson, 2013]) in the second dataset using the first sample to obtain the external weights. Any bias due to weak instruments is in the direction of the null [Pierce and Burgess, 2013; Burgess et al., 2016b]. This arises for the same reason as regression dilution bias in observational studies [Frost and Thompson, 2000]. If the fitted values of the risk factor are imprecisely estimated (as will be the case with weak instruments), then they will suffer from non-differential measurement error, and bias in the second-stage regression will be in the direction of the null.

2.2 Inverse-variance weighted method and equivalence to 2SLS method

If individual-level data are not available, but instead summarized data on the associations of the genetic variants with the risk factor and with the outcome, then the 2SLS method cannot be implemented. Instead, an inverse-variance weighted (IVW) method is often employed that combines the ratio estimates calculated separately for each IV, using a formula for performing a fixed-effect meta-analysis [Burgess et al., 2013]. This method assumes that the IVs are uncorrelated (that is, not in linkage disequilibrium), although extensions for correlated genetic variants have been proposed [Burgess et al., 2016b]. The estimate from the IVW method is equal to the estimate from the 2SLS method asymptotically. The two estimates are also equal in finite samples when the correlations between the IVs are exactly zero [Burgess et al., 2016b]. The level of weak instrument bias in the IVW method has been shown to be the same as that from the 2SLS method in realistic simulations [Burgess et al., 2013]. We therefore expect the results of this paper to apply equally to analyses performed using the 2SLS method and individual-level data, as to using the IVW method and summarized data.

Although we use the 2SLS method in some simulation studies of this paper for computational convenience, and we refer to theoretical results for weak instrument bias derived using the 2SLS method, the focus of this paper is the summarized data setting. Results from the 2SLS method are presented because of their similarity with those from the IVW method that can be performed using summarized data. If individual-level data were available, then several alternative approaches for mitigating weak instrument bias would be possible, such as using either the limited information maximum likelihood (LIML) or the continuously updating estimator (CUE) method

[Davies et al., 2015], using a jackknife IV estimator [Angrist et al., 1999] (or equivalently an allele score approach using leave-one-out cross-validated weights [Burgess and Thompson, 2013]), or using an allele score approach using equal or externally-specified weights [Burgess and Thompson, 2013]. However, with summarized data, only the final option (equal or externally-specified weights) is possible [Burgess et al., 2016b].

2.3 Magnitude of bias in a one-sample setting

The ordinary least squares (OLS, also known as standard least squares regression) estimate is obtained by regressing the outcome on the risk factor. This 'observational' estimate is typically biased due to confounding. The relative bias of the 2SLS estimate – the bias of the 2SLS estimate divided by the bias of the OLS estimate – is approximately and asymptotically equal to $1/\mathbb{E}(F)$ [Staiger and Stock, 1997], where $\mathbb{E}(F)$ is the 'F parameter'. An F parameter of 10 therefore corresponds to a $1/10 = 10\%$ relative bias of the 2SLS estimate compared to the OLS estimate. However, this calculation cannot be directly employed for bias correction in an applied setting, as the F statistic in a given dataset may differ substantially from the F parameter due to random variation [Burgess et al., 2011]. As weak instrument bias occurs due to chance correlations with confounders, the reference OLS estimate should ideally be unadjusted for measured confounders, unless these confounders are also adjusted for in the IV estimate.

2.4 Expected bias of 2SLS estimator

The bias of the 2SLS estimator in a one-sample setting has been considered theoretically [Nagar, 1959]. One approximation to the bias is:

$$\text{Bias of 2SLS estimator (one sample)} = \frac{\sigma_{XY}\,(K-2)}{\sigma_X^2\,\mu} \approx \frac{\sigma_{XY}}{\sigma_X^2\,\mathbb{E}(F)} \tag{1}$$

where μ is the concentration parameter, K is the number of IVs, σ_X^2 is the variance of the error in the first-stage regression model, and σ_{XY} is the covariance of the error terms in the first- and second-stage regression models [Bun and Windmeijer, 2011]. In a two-sample setting, this formula may not be directly applicable, as the sample sizes for the first- and second-stage regressions (and so the lengths of the error term vectors) may differ, in which case a covariance cannot be calculated. If the sample sizes are equal, then we can decompose the covariance into a term corresponding to individuals included in both regressions, and a term corresponding to unrelated individuals which will have expectation zero. If the sample sizes differ, dependence between these two error terms is still driven by individuals in common between these two regressions, and the presence of individuals in only one or other of the regressions will dilute this dependence. As covariance is a linear operator in both its arguments, we may therefore expect the bias of the 2SLS estimator to be approximately linear in the proportion of overlap between the samples.

3 Simulation study – continuous outcome

To investigate the degree and direction of bias in a "two-sample" Mendelian randomization investigation where there is overlap between the samples, we conduct a simulation study. The data-generating model is given below.

$$\begin{aligned} g_{ik} &\sim \text{Binomial}(2, 0.3) \text{ independently for } k = 1, \ldots, 20 \qquad (2) \\ x_i &= \sum_{k=1}^{20} \alpha g_{ik} + u_i + \epsilon_{Xi} \\ y_i &= \beta_X x_i + \beta_U u_i + \epsilon_{Yi} \\ u_i &\sim \mathcal{N}(0,1), \epsilon_{Xi} \sim \mathcal{N}(0,1), \epsilon_{Yi} \sim \mathcal{N}(0,1) \text{ independently} \end{aligned}$$

The 20 IVs ($g_{ik}, k = 1, \ldots, 20$) are indexed by k and individuals are indexed by i. The IVs are modelled as independently-distributed biallelic single nucleotide polymorphisms (SNPs) with a minor allele frequency of 0.3. The risk factor (x_i) is a linear function of the IVs, a confounder (u_i), and an independent error term (ϵ_{Xi}). The outcome (y_i) is a linear function of the risk factor, confounder, and another independent error term (ϵ_{Yi}). The IVs all have the same per allele effect on the risk factor (α). The causal effect of the risk factor on the outcome is β_X, and the effect of the confounder on the outcome is β_U.

We consider the cases of a positive causal effect ($\beta_X = 0.2$) and a null causal effect ($\beta_X = 0$), and take three values of the confounder effect on the outcome ($\beta_U = 0.6, 1, 2$). The IV strength is varied by taking three values of $\alpha = 0.04, 0.06, 0.08$. We consider cases where there is 0% overlap, increasing in increments of 10% up to a 100% overlap. This is achieved by simulating data on 20 000 individuals. Estimates were obtained using the 2SLS and IVW methods. The first-stage regression is undertaken (or for the IVW method, IV–risk factor associations are estimated) in the first 10 000 individuals, and the second-stage regression (IV–outcome associations) in individuals 10 001 to 20 000 (0% overlap), individuals 9001 to 19 000 (10% overlap), individuals 8001 to 18 000 (20% overlap), and so on. Hence all associations were estimated using 10 000 individuals. 10 000 simulations were considered for each set of parameters.

Although some aspects of the simulation study are not varied here, we repeated the simulation study varying the number of IVs, and the total sample size (results not shown). In each case, the amount of bias was almost identical between scenarios in which the F parameter (expected value of the F statistic) was similar. Hence, we would expect the results of this simulation study to be generalizable to other situations, and would view the F parameter (which depends on the sample size, number of IVs, and the proportion of variance in the risk factor explained by the IVs) as the key measure of bias.

3.1 Results

Results are given in Table I and displayed visually in Figure 1. For each set of parameters, the mean estimate from the 2SLS method across simulations is given. Mean estimates from the IVW method were equal to those from the 2SLS method to 3 decimal places, and median estimates across simulations from both methods were similar to mean estimates. The Monte Carlo standard error for the mean estimates is 0.002 or less in all scenarios. Mean F and R^2 (coefficient of determination) statistics from regression of the risk factor on the IVs based on 10 000 participants are provided to judge the strength of the instrumental variants. Ordinary least squares (OLS) estimates are also provided to help judge the level and direction of confounding.

With a positive causal effect, the results demonstrate clearly the transition from bias in the direction of the null with no overlap, to bias in the direction of the confounded association with increasing overlap. This transition happens earlier when the

Table I: **Simulation 1 with continuous outcome and different overlap proportions**

	Mean	Mean	Percentage	$\beta_X = 0.2$			$\beta_X = 0$		
α	F	R^2	overlap	$\beta_U = 0.6$	$\beta_U = 1$	$\beta_U = 2$	$\beta_U = 0.6$	$\beta_U = 1$	$\beta_U = 2$
0.04	4.4	0.9%	OLS	0.498	0.697	1.193	0.298	0.497	0.993
			0%	0.156	0.157	0.161	-0.001	0.000	0.004
			10%	0.167	0.173	0.186	0.006	0.012	0.024
			20%	0.178	0.187	0.211	0.012	0.022	0.045
			30%	0.189	0.203	0.237	0.019	0.033	0.067
			40%	0.200	0.218	0.262	0.026	0.044	0.087
			50%	0.211	0.233	0.287	0.033	0.055	0.108
			60%	0.223	0.249	0.313	0.040	0.066	0.130
			70%	0.234	0.264	0.339	0.047	0.077	0.152
			80%	0.245	0.280	0.363	0.053	0.088	0.172
			90%	0.256	0.294	0.389	0.060	0.098	0.194
			100%	0.266	0.309	0.415	0.066	0.109	0.215
0.06	8.5	1.7%	OLS	0.495	0.692	1.185	0.295	0.492	0.985
			0%	0.178	0.179	0.176	0.000	0.000	-0.002
			10%	0.184	0.186	0.189	0.003	0.005	0.009
			20%	0.189	0.193	0.203	0.006	0.011	0.020
			30%	0.194	0.201	0.215	0.009	0.016	0.031
			40%	0.200	0.209	0.228	0.013	0.022	0.041
			50%	0.206	0.216	0.242	0.016	0.027	0.052
			60%	0.211	0.224	0.255	0.019	0.032	0.063
			70%	0.216	0.231	0.268	0.023	0.038	0.074
			80%	0.221	0.239	0.282	0.026	0.043	0.086
			90%	0.226	0.246	0.295	0.028	0.048	0.097
			100%	0.232	0.254	0.309	0.031	0.054	0.109
0.08	14.4	2.8%	OLS	0.492	0.687	1.174	0.292	0.487	0.974
			0%	0.187	0.187	0.187	0.000	0.000	0.000
			10%	0.191	0.191	0.195	0.002	0.003	0.007
			20%	0.194	0.195	0.203	0.004	0.006	0.014
			30%	0.197	0.199	0.211	0.006	0.009	0.020
			40%	0.200	0.204	0.218	0.008	0.011	0.026
			50%	0.203	0.208	0.225	0.010	0.014	0.032
			60%	0.206	0.213	0.232	0.012	0.018	0.038
			70%	0.210	0.218	0.240	0.014	0.022	0.044
			80%	0.213	0.222	0.248	0.016	0.024	0.050
			90%	0.217	0.226	0.255	0.018	0.027	0.056
			100%	0.220	0.230	0.264	0.020	0.030	0.064

Mean two-stage least squares (or equivalently, inverse-variance weighted) estimates with true causal effect $\beta_X = 0.2$ (positive effect) and $\beta_X = 0$ (null effect) for 3 values of genetic associations with the risk factor (α), 3 values of the confounder effect on the outcome (β_U), and 11 values of the percentage overlap between the two samples. The mean F statistic (F), mean coefficient of determination (R^2), and mean ordinary least squares (OLS) estimate are given to judge the strength of the instrumental variables and the degree of confounding.

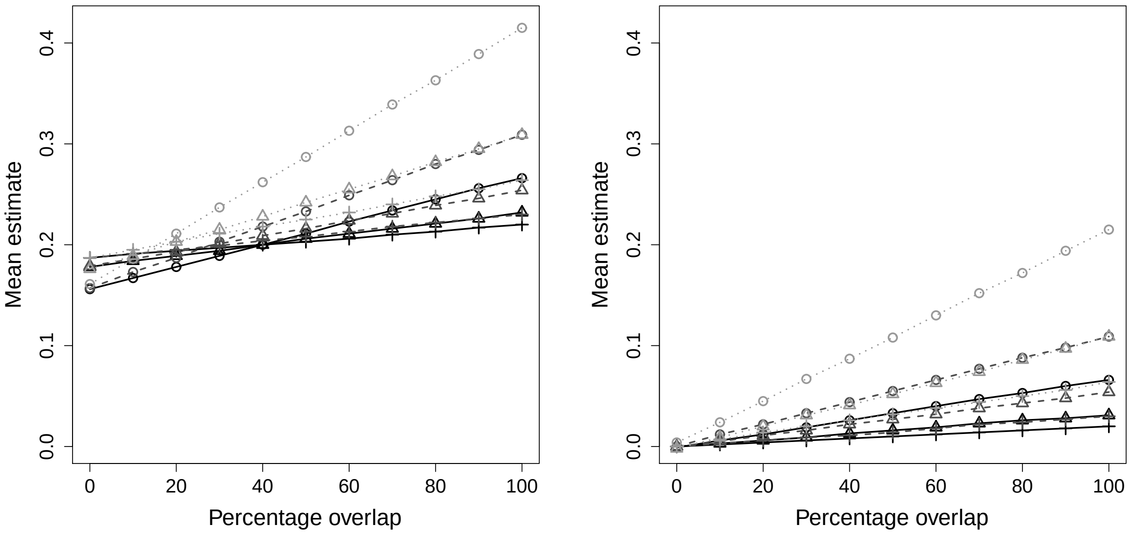

Figure 1: Mean two-stage least squares/inverse-variance weighted estimates plotted against sample overlap for different values of instrument strength ($\alpha = 0.4$, circle; $\alpha = 0.6$, triangle; $\alpha = 0.8$, plus) and different values of the confounder effect on the outcome ($\beta_U = 0.6$, black solid line; $\beta_U = 1$, mid-grey dashed line; $\beta_U = 2$, light-grey dotted line). Left panel: positive causal effect ($\beta_X = 0.2$); right panel: null causal effect ($\beta_X = 0$).

confounding is stronger, although the balance point where the biases cancel out does not seem to depend on the strength of the IVs: for $\beta_U = 0.6$, it is around 40%; for $\beta_U = 1$, it is around 28%; and $\beta_U = 2$, it is around 16%. Further simulations (not shown) suggest that the balance point also depends on the magnitude of the causal effect, so the precise balance points in this simulation will not necessarily hold in other cases. With a null causal effect, there is no bias with 0% overlap, and bias increases as the degree of overlap increases.

3.2 Deriving analytic formulae for the expected bias under the null and Type 1 error rate

The relationship between the level of overlap and the mean estimate appears to be linear throughout, both with a positive and with a null causal effect. Equally, with a null causal effect, the bias is proportional to the effect of the confounder (β_U), which in turn is proportional to the OLS estimate. This suggests that a formula can be derived for the amount of bias expected under the null hypothesis:

$$\text{Bias under null} = \text{OLS estimate} \times \frac{\text{Percentage overlap}}{100} \times \text{Relative bias} \tag{3}$$

where the relative bias is the reciprocal of the F parameter.

Given the bias and the standard error of an estimate, the expected Type 1 (false positive) error rate for a two-sided test of size 5% can be approximated analytically as:

$$\text{Type 1 error rate} = 2 - \Phi\left(1.96 - \frac{\text{bias}}{\text{standard error}}\right) - \Phi\left(1.96 + \frac{\text{bias}}{\text{standard error}}\right) \tag{4}$$

where Φ is the cumulative standard normal function. The variance of the IV estimate (the 2SLS or the IVW estimate) can be approximated as:

$$\text{Variance of IV estimate} \simeq \frac{\text{var}(R_Y)}{N\ \text{var}(X)\,\rho^2} \tag{5}$$

where N is the sample size for the IV–outcome association, R_Y is the residual outcome after subtracting the causal effect of the risk factor ($R_Y = Y - \hat{\beta}_X X$), and ρ is the correlation between the IVs and the risk factor (the R^2 statistic – the proportion of variance in the risk factor explained by the IVs – is an estimate of ρ^2) [Nelson and Startz, 1990]. The variance of R_Y is equal to the variance of Y when the causal effect of the risk factor on the outcome is zero. The relationship between the F statistic and the R^2 statistic [Dobson, 2001] is:

$$F = \frac{N - K - 1}{K}\,\frac{R^2}{1 - R^2} \tag{6}$$

where K is the number of IVs. For small values of R^2, this means that the F and R^2 statistics are approximately linearly related.

Hence, given the sample size, sample overlap percentage, OLS estimate (in standard deviation units for the risk factor and outcome, otherwise the standard deviations of risk factor and outcome are required), and an estimate of the strength of the IVs (either the F statistic or the R^2 statistic), the bias and Type 1 error rate can be calculated. R code for performing these calculations is given in Web Appendix A1.

3.3 Validating the analytic formulae

To assess the validity of the analytic formulae for the bias and Type 1 error rate, we conducted a further set of 10 000 simulations in 10 000 participants using model (2) under the causal null ($\beta_X = 0$), with the strongest level of confounding ($\beta_U = 2$) and 100% sample overlap, with a wider range of values for the IV strength ($\alpha = 0.01, 0.02, \ldots, 0.1, 0.15, 0.2$) to estimate the relative bias and empirical Type 1 error rate (proportion of simulations for which the Wald test for the IV estimate rejected the null) in each case.

These results are given in Table II. The relative bias estimates are close to $1/\mathbb{E}(F)$, as theoretically predicted [Staiger and Stock, 1997]. The expected Type 1 error rate was calculated using the mean values of the F statistic, R^2 statistic, variance of the risk factor and outcome, and the OLS estimate across simulations. The empirical and expected Type 1 error rates showed fairly close agreement, with the expected Type 1 error rate tending to be a slight underestimate of the empirical rate for all but the weakest of instruments. The Monte Carlo standard error for the empirical Type 1 error rate was around 0.2%.

Additional simulations are provided in Web Appendix A2. We varied the number of IVs ($K = 10$) and considered different overlap proportions from 0% to 100% (Web Table A1). In this case, closer correspondence was observed between the expected and observed Type 1 error rates throughout, although these were across a narrower range. We also varied the sample size ($N = 1000$, Web Table A2), and considered the case of a binary risk factor (Web Table A3). In both these cases, there was close agreement between the expected and observed values of the relative bias and the Type 1 error rate. There was some deviation in Type 1 error estimates with a smaller sample size, with Type 1 error rates underestimated with stronger instruments (F parameter 5 to 10). However, this discrepancy appeared to be due to problems of maintaining nominal Type 1 error rates with weak instruments more generally rather than mis-estimation of the relative bias, and would be resolved by using methods for inference that do not rely on the IV estimate having a normal distribution (for example, Fieller's theorem [Burgess et al., 2015b] or inversion of the Anderson–Rubin test statistic [Mikusheva, 2010]).

Table II: **Simulation 2 with continuous outcome to validate bias and Type 1 error rate formulae**

α	Mean F	Mean R^2	Mean OLS estimate	Mean IV estimate	Relative bias	(Mean F)$^{-1}$	Empirical Type 1 error	Expected Type 1 error
0.01	1.2	0.2%	0.999	0.829	0.829	0.825	84.8%	89.5%
0.02	1.8	0.4%	0.998	0.531	0.532	0.545	65.5%	69.8%
0.03	2.9	0.6%	0.996	0.333	0.334	0.346	46.9%	46.1%
0.04	4.4	0.9%	0.993	0.216	0.218	0.229	32.5%	30.6%
0.05	6.3	1.2%	0.990	0.149	0.151	0.160	24.1%	21.8%
0.06	8.6	1.7%	0.985	0.109	0.111	0.117	19.2%	16.8%
0.07	11.3	2.2%	0.980	0.082	0.083	0.089	15.6%	13.6%
0.08	14.4	2.8%	0.974	0.062	0.064	0.069	12.4%	11.5%
0.09	18.0	3.5%	0.967	0.050	0.052	0.056	11.0%	10.1%
0.10	22.0	4.2%	0.960	0.040	0.042	0.046	9.8%	9.0%
0.15	48.3	8.8%	0.914	0.019	0.021	0.021	7.5%	6.6%
0.20	85.0	14.5%	0.856	0.010	0.012	0.012	5.9%	5.8%

Simulation results with null causal effect $\beta_X = 0$, and confounder effect $\beta_U = 2$ to estimate the relative bias and empirical Type 1 error rate (5% nominal significance level) of the two-stage least squares (or equivalently, inverse-variance weighted) instrumental variable (IV) estimate; the relative bias is the bias of the IV estimate divided by the bias of the ordinary least squares (OLS) estimate. The relative bias is theoretically predicted to be close to the reciprocal of the mean value of the F statistic, (Mean F)$^{-1}$.

3.4 The F parameter and the F statistic

The bias of an IV estimate depends on the expected value of the F statistic (referred to here as the F parameter). However, in practice, the F parameter will be unknown and only the measured F statistic (an estimate of the F parameter) will be available in a single dataset. The F statistic can be highly variable. For example, we previously took a large study and divided it into 16 equally sized substudies at random [Burgess et al., 2011]. By construction, each of these substudies should have had the same expected F statistic. However, the measured F statistics ranged from 3.4 to 22.6 (mean was 10.8).

One practical suggestion is to use an estimate of the F statistic based on an external dataset to ensure that the estimate of bias is not dependent on the F statistic in the data under analysis. This can be achieved by taking the value of R^2 (which is independent of sample size) from the external dataset and calculating the corresponding F statistic for the sample size under investigation; the R^2 for a single SNP is $2\,\hat{\alpha}^2 \times MAF \times (1 - MAF)$, where the genetic association with the risk factor $\hat{\alpha}$ is in standard deviation units, and MAF is the minor allele frequency. Additionally, the bias calculation can be repeated taking a lower value of the F statistic to address the problem that the F statistic in the data under analysis may be an overly optimistic estimate of the F parameter. For example, one could take the lower limit from a confidence interval for the F parameter (such as the lower limit of a one-sided 95% confidence interval – only the lower tail of the interval is relevant). A method for constructing a confidence interval for the non-centrality parameter of an F distribution (from which a confidence interval for the F parameter can be obtained) has been considered previously [Venables, 1975]; this method is outlined in Web Appendix A3, and code for implementing the method is provided.

4 Simulation study – binary outcome

With a binary outcome in a case-control setting (as is common for Mendelian randomization with a disease outcome), the ratio estimate is calculated by dividing the IV–outcome coefficient from logistic regression by the IV–risk factor coefficient from linear regression [Didelez et al., 2010b]. A two-sample method can also be performed by replacing the linear model in the second-stage regression of the outcome on the fitted values of the risk factor with a logistic model. When summarized data are available, a causal effect estimate can be obtained based using the IVW method to combine the ratio estimates, as in the case of a continuous outcome. There are some technical issues relating to the interpretation of these estimates with a binary outcome and a logistic regression model due to the non-collapsibility of odds ratios (they approximate a population-averaged log odds ratio per unit change in the distribution of the risk factor [Burgess and CHD CRP Genetics Collaboration, 2013]), but each is a consistent estimator under the null, and each provides a valid test of the null hypothesis of no causal effect [Vansteelandt et al., 2011].

In principle, similar analytical formulae for bias under the null and Type 1 error rate could be developed with a binary outcome ($Y = 0, 1$). This would require a different formula for the variance of the IV estimate, as this depends on the number and ratio of participants with outcome events [Burgess, 2014]:

$$\text{Variance of IV estimate (binary)} \simeq \frac{1}{N\ \text{var}(X)\, \rho^2\, \mathbb{P}(Y=1)\, \mathbb{P}(Y=0)}. \tag{7}$$

However, if the data on the outcome are derived from a case-control sample, then typically the associations with the risk factor are estimated in control participants only [Bowden and Vansteelandt, 2011]. This is for three main reasons: to avoid reverse causation, particularly if the risk factor is measured after the outcome event in cases; to avoid biases due to outcome-dependent sampling (associations may be present in the case-control sample even if they are absent in the underlying population) [Didelez et al., 2010a]; and because the controls are a more representative sample of the population as a whole [Didelez and Sheehan, 2007]. Hence, even if there is overlap between datasets with a binary outcome, then provided that the IV–risk factor associations are estimated in the controls only, bias may not be substantial.

To investigate this, we simulated data using a similar data-generating mechanism as previously:

$$\begin{aligned} g_{ik} &\sim \text{Binomial}(2, 0.3) \text{ independently for } k = 1, \dots, 20 \\ x_i &= \sum_{k=1}^{20} \alpha g_{ik} + u_i + \epsilon_{Xi} \\ \text{logit}(\pi_i) &= -3 + \beta_X x_i + \beta_U u_i \\ y_i &\sim \text{Binomial}(1, \pi_i) \\ u_i &\sim \mathcal{N}(0,1), \epsilon_{Xi} \sim \mathcal{N}(0,1) \text{ independently} \end{aligned} \tag{8}$$

A logistic-linear relationship is assumed between the probability of an outcome event (π_i) and the risk factor and confounder. A case-control sample was generated by simulating data on 100 000 individuals, and taking the first 5000 with an event ($y_i = 1$) as cases, and the first 5000 without an event ($y_i = 0$) as controls. IV–outcome associations were estimated in all participants using logistic regression. The IV–risk factor associations were estimated both on the controls only, and in all participants (controls and cases). This was a one-sample analysis (100% sample overlap). We set $\beta_X = 0$ (null causal effect) and $\beta_U = 1$, and considered scenarios with $\alpha = 0.01, 0.02, 0.03, 0.04, 0.05, 0.08$.

4.1 Results

The mean estimates and empirical Type 1 (false positive) error rates based on 10 000 simulated datasets are given in Table III. With IV–risk factor associations estimated in the controls only, there was no detectable bias in the IV estimates even with extremely weak instruments, nor was there any inflation of Type 1 error rates. This suggests that a conventional Mendelian randomization analysis with a binary outcome in which the associations of the IV with the risk factor are only estimated in control participants provides a natural robustness against weak instrument bias, even in a one-sample setting. With IV–risk factor associations estimated in all participants, bias was similar to that with a continuous outcome, with relative bias close to $1/\mathbb{E}(F)$ on the log odds ratio scale and empirical Type 1 error rates close to the predicted values. However, the approximations were less accurate compared with the continuous outcome case, particularly for the weakest of instruments. This suggests that the same

Table III: **Simulation 3 with binary outcome to validate bias and Type 1 error rate formulae**

α	Mean F	Mean R^2	Mean observational estimate	Mean IV estimate	Relative bias	(Mean $F)^{-1}$	Empirical Type 1	Expected Type 1
			Risk factor measurements taken in controls only					
0.01	1.1	0.4%	0.481	0.001	0.002	-	4.9%	-
0.02	1.4	0.6%	0.481	0.000	0.000	-	5.2%	-
0.03	2.0	0.8%	0.479	-0.003	0.007	-	4.8%	-
0.04	2.7	1.1%	0.478	-0.001	-0.001	-	5.0%	-
0.05	3.7	1.5%	0.476	0.000	0.000	-	5.2%	-
0.08	7.9	3.1%	0.469	0.000	0.001	-	4.7%	-
			Risk factor measurements taken in all participants					
0.01	1.2	0.3%	0.481	0.360	0.748	0.837	24.4%	29.1%
0.02	1.8	0.4%	0.481	0.237	0.493	0.561	17.4%	21.1%
0.03	2.8	0.5%	0.479	0.149	0.311	0.363	12.8%	15.2%
0.04	4.1	0.8%	0.478	0.099	0.207	0.242	10.0%	11.7%
0.05	5.9	1.2%	0.476	0.068	0.142	0.170	8.4%	9.6%
0.08	13.6	2.6%	0.469	0.030	0.064	0.074	6.3%	6.9%

Mean instrumental variable (IV) estimates and empirical Type 1 error rate (5% nominal significance level) from inverse-variance weighted method with binary outcome for null causal effect ($\beta_X = 0$) and 6 values of genetic associations with the risk factor (α) in a case-control setting, with the risk factor measurements taken in control participants only and with the risk factor measurements taken in all participants. Observational estimates are log odds ratios from logistic regression of the outcome on the risk factor, and IV estimates are log odds ratios calculated using logistic regression for the IV–outcome association and linear regression for the IV–risk factor association.

analytic formulae can be used with a binary outcome as with a continuous outcome, except with a different expression for the standard error of the IV estimate. R code for performing these calculations is given in Web Appendix A1.

5 Examples: sample overlap between large consortia

We consider several Mendelian randomization analyses that could be undertaken using published summarized data from large consortia, and discuss the potential for bias due to participant overlap in each case.

1. **Body mass index and lipid traits:** We consider an analysis to estimate the causal effect of body mass index (BMI) on various lipid traits using data from the GIANT (Genetic Investigation of Anthropometric Traits) consortium [Locke et al., 2015] and the GLGC (Global Lipids Genetics Consortium) [The Global Lipids Genetics Consortium, 2013]. 55 common studies are mentioned in the papers authored by these two consortia, comprising around 71% of participants in the GLGC. The 97 genetic variants reported as associated with BMI at a genome-wide level of significance explain around 2.7% of the variance in BMI in the GIANT dataset (sample size = 339 224), corresponding to an F statistic of approximately $\frac{0.027}{1-0.027} \times \frac{339\,126}{97} = 97.0$. The lower limit of a one-sided 95% confidence interval for the F parameter is 93.7 (see Web Appendix A3 for calculation). Hence, despite the substantial overlap between the two consortia, considerable weak instrument bias would not be expected.

2. **Body mass index and coronary heart disease risk:** Although the F statistic suggests that weak instrument bias would not be substantial in any case, we consider the binary outcome of coronary heart disease (CHD), and investigate the degree of overlap between participants in the GIANT consortium above and the CARDIoGRAMplusC4D consortium [The CARDIoGRAMplusC4D Consortium, 2013]. Of the 38 studies that appear in CARDIoGRAMplusC4D, 27 appear in GIANT in some form. In many cases, both case and control participants are included in GIANT. Hence, even though CHD is a binary outcome, the sam-

ple overlap between the two consortia could lead to weak instrument bias if the F parameter for BMI were lower.

3. **Educational attainment:** A genetic score for an individual's number of years of schooling ("EduYears") constructed using 5 genetic variants associated with EduYears at a genome-wide level of significance ($P < 5 \times 10^{-8}$) explained about 0.1% of the variance in EduYears in the discovery sample of 101 069 individuals (corresponding F statistic = 20.2) [Rietveld et al., 2013]. Associations were also reported for a follow-up sample of 25 490 individuals (corresponding F statistic = 5.1). The corresponding lower limits of the one-sided 95% confidence intervals for the F parameter are 14.0 and 2.3 (see Web Appendix A3). Hence, a Mendelian randomization investigation using associations from the discovery sample should not lead to substantial weak instrument bias, but an investigation using associations from the follow-up sample may be severely affected by weak instruments. However, if associations from the discovery sample are used in a Mendelian randomization investigation, then bias (in particular, selection bias) may be more serious due to winner's curse; this issues is explored further in the discussion.

These examples suggest that sample overlap between major international consortia may be substantial. Bias from weak instruments in very large consortia may not be substantial, but in moderately large consortia, potential bias and inflated Type 1 error rates should be investigated.

6 Discussion

In this paper, we have shown that bias in a Mendelian randomization investigation with a continuous outcome in a two-sample setting is linearly related to the proportion of sample overlap between the two datasets. We have provided and validated analytical formulae for the expected bias and Type 1 error rate under the null given the F parameter (the expected value of the F statistic), the observational (OLS) estimate, the sample size, and the sample overlap percentage. With a binary outcome, provided that the IV–risk factor associations are estimated in the control participants only, bias due to sample overlap is negligible and Type 1 error rates are at nominal levels. If

IV–risk factor associations are estimated in all participants, then bias is similar with a binary outcome (on the log odds ratio scale) as with a continuous outcome. Formulae for estimating the bias and Type 1 error rate under the null are given, and R code for calculating these quantities is given in Web Appendix A1. This code is implemented in a web application at `https://sb452.shinyapps.io/overlap`.

6.1 Overlap with datasets of different size

If the datasets are of different size, then the percentage overlap in these formulae should be taken with respect to the larger dataset, as this determines the correlation between the association estimates. This is because only individuals in both studies will lead to correlation between the association estimates, and additional individuals in either the association with the risk factor or with the outcome will attenuate any correlation. For example, if the smaller dataset has 1000 participants, and the larger dataset has 10 000 participants, then the sample overlap is only 10% even if all of the participants from the smaller dataset are included in the larger dataset.

6.2 Additional bias from genetic discovery

Weak instrument bias will be accentuated if the genetic variants were initially discovered in the data under analysis. This is due to winner's curse; if several genetic variants in truth have similar magnitudes of association with the risk factor, the association of the one that is the strongest in the data under analysis is likely to be overestimated [Burgess et al., 2011; Taylor et al., 2014]. As this overestimation will generally mean that the associations with confounders are by chance stronger than expected, bias will occur if the discovery dataset is used in the estimation of the IV–risk factor or the IV–outcome associations. In a binary outcome setting, provided that only control participants were used in the discovery dataset, this should not lead to bias. However, if controls and cases were both used in the discovery dataset, this will lead to weak instrument bias.

There is no clear way to evaluate the bias due to overlap between the discovery sample and the dataset(s) used in a Mendelian randomization investigation. Hence, in such cases caution should be expressed, particularly if a genetic variant is close to the threshold statistical significance level for discovery or for inclusion in the Mendelian randomization analysis. In particular, data-driven approaches for choosing genetic variants to be included in a Mendelian randomization analysis should be avoided. Analytical approaches to correct genetic associations for winner's curse may be useful in such a situation [Bowden and Dudbridge, 2009].

6.3 Increased bias in MR-Egger method

The recently introduced MR-Egger method method has advantages over the conventional two-stage least squares and inverse-variance weighted methods in terms of some robustness to the instrumental variable assumptions being violated [Bowden et al., 2015]. Although weak instrument bias using data from large consortia may be low for conventional methods, bias for the MR-Egger method has been shown to be considerably more pronounced, both attenuation in a two-sample setting and bias towards the observational association in a one-sample setting [Bowden et al., 2016a; Burgess et al., 2016a]. This means that bias due to sample overlap may be more serious for the MR-Egger method. In a two-sample setting, bias in the MR-Egger estimate does not depend on the proportion of variance explained by the IVs, but rather on the variability between the IV associations with the risk factor. Hence, for the MR-Egger method, an I^2 heterogeneity statistic is a better indicator of bias than the F statistic [Bowden et al., 2016b]. Further research is needed to derive an analytical formula for weak instrument bias in the MR-Egger method in the one-sample setting.

6.4 Practical recommendations

If there is sample overlap that is likely to lead to substantial bias and inflated Type 1 error rates in a "two-sample" Mendelian randomization investigation, several approaches are available. If possible, the genetic associations with the risk factor could be derived from another non-overlapping data source (possibly a subset of the original studies). A disadvantage of this is the potential loss of efficiency if the genetic associations are estimated less precisely [Burgess and Thompson, 2013]. Alternatively, equal weights can be used, although again, there is a potential loss of power to detect a causal effect [Burgess et al., 2016b]. (Software code for performing a summarized data analysis equivalent to a equally-weighted allele score analysis is provided in Web Appendix A4.) Finally, sensitivity analyses can be performed using fewer but stronger genetic variants (and hence increasing the F parameter).

For consortia that publish genetic association estimates with continuous risk factors, we recommend that such estimates do not include case participants from case-control studies. This will help reduce sample overlap in future Mendelian randomization studies that investigate the causal effect of the continuous risk factor on disease risk.

Acknowledgements

This work was funded by the UK Medical Research Council (G0800270), British Heart Foundation (SP/09/002), UK National Institute for Health Research Cambridge Biomedical Research Centre, European Research Council (268834), and European Commission Framework Programme 7 (HEALTH-F2-2012-279233). Stephen Burgess is supported by the Wellcome Trust (grant number 100114). The authors have no conflicts of interest to report.

References

Angrist, J., Imbens, G., and Krueger, A. 1999. Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67.

Angrist, J. and Krueger, A. 1992. The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, 87(418):328–336.

Angrist, J. and Pischke, J. 2009. *Mostly harmless econometrics: an empiricist's companion. Chapter 4: Instrumental variables in action: sometimes you get what you need.* Princeton University Press.

Angrist, J. D. and Krueger, A. B. 1995. Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235.

Bound, J., Jaeger, D., and Baker, R. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.

Bowden, J., Davey Smith, G., and Burgess, S. 2015. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2):512–525.

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. 2016a. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314.

Bowden, J., Del Greco, F., Minelli, C., Davey Smith, G., Sheehan, N. A., and Thompson, J. R. 2016b. Assessing the suitability of summary data for Mendelian randomization analyses using MR-Egger regression: the role of the I^2 statistic. Under review.

Bowden, J. and Dudbridge, F. 2009. Unbiased estimation of odds ratios: combining genomewide association scans with replication studies. *Genetic Epidemiology*, 33(5):406–418.

Bowden, J. and Vansteelandt, S. 2011. Mendelian randomisation analysis of case-control data using structural mean models. *Statistics in Medicine*, 30(6):678–694.

Bun, M. J. and Windmeijer, F. 2011. A comparison of bias approximations for the two-stage least squares (2SLS) estimator. *Economics Letters*, 113(1):76–79.

Burgess, S. 2014. Sample size and power calculations in Mendelian randomization with a single instrumental variable and a binary outcome. *International Journal of Epidemiology*, 43(3):922–929.

Burgess, S., Bowden, J., Dudbridge, F., and Thompson, S. G. 2016a. Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. *arXiv*, 1606.03729.

Burgess, S., Butterworth, A. S., and Thompson, S. G. 2013. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665.

Burgess, S. and CHD CRP Genetics Collaboration 2013. Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. *Statistics in Medicine*, 32(27):4726–4747.

Burgess, S., Dudbridge, F., and Thompson, S. G. 2016b. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in Medicine*, 35(11):1880–1906.

Burgess, S., Scott, R., Timpson, N., Davey Smith, G., Thompson, S. G., and EPIC-InterAct Consortium 2015a. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *European Journal of Epidemiology*, 30(7):543–552.

Burgess, S., Small, D. S., and Thompson, S. G. 2015b. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*. Available online.

Burgess, S. and Thompson, S. G. 2011. Bias in causal estimates from Mendelian randomization studies with weak instruments. *Statistics in Medicine*, 30(11):1312–1323.

Burgess, S. and Thompson, S. G. 2013. Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology*, 42(4):1134–1144.

Burgess, S. and Thompson, S. G. 2015. *Mendelian randomization: methods for using genetic variants in causal estimation.* Chapman & Hall.

Burgess, S., Thompson, S. G., and CRP CHD Genetics Collaboration 2011. Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*, 40(3):755–764.

Cragg, J. and Donald, S. 1993. Testing identifiability and specification in instrumental variable models. *Econometric Theory*, 9(02):222–240.

Dastani, Z., Hivert, M.-F., Timpson, N., et al. 2012. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: A multi-ethnic meta-analysis of 45,891 individuals. *PLOS Genetics*, 8(3):e1002607.

Davey Smith, G. and Ebrahim, S. 2003. ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22.

Davies, N., von Hinke Kessler Scholder, S., Farbmacher, H., Burgess, S., Windmeijer, F., and Davey Smith, G. 2015. The many weak instrument problem and Mendelian randomization. *Statistics in Medicine*, 34(3):454–468.

Didelez, V., Kreiner, S., and Keiding, N. 2010a. Graphical models for inference under outcome-dependent sampling. *Statistical Science*, 25(3):368–387.

Didelez, V., Meng, S., and Sheehan, N. 2010b. Assumptions of IV methods for observational epidemiology. *Statistical Science*, 25(1):22–40.

Didelez, V. and Sheehan, N. 2007. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330.

Dobson, A. 2001. *An introduction to generalized linear models*. Chapman & Hall.

Frost, C. and Thompson, S. 2000. Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(2):173–189.

Greenland, S. 2000. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729.

Hahn, J., Hausman, J., and Kuersteiner, G. 2004. Estimation with weak instruments: accuracy of higher-order bias and MSE approximations. *Econometrics Journal*, 7(1):272–306.

Inoue, A. and Solon, G. 2010. Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92(3):557–561.

Johnson, T. 2013. Efficient calculation for multi-SNP genetic risk scores. Technical report, The Comprehensive R Archive Network. Available at http://cran.r-project.org/web/packages/gtx/vignettes/ashg2012.pdf [last accessed 2014/11/19].

Lawlor, D., Harbord, R., Sterne, J., Timpson, N., and Davey Smith, G. 2008. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163.

Locke, A. E., Kahali, B., Berndt, S. I., et al. 2015. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206.

Martens, E., Pestman, W., de Boer, A., Belitser, S., and Klungel, O. 2006. Instrumental variables: application and limitations. *Epidemiology*, 17(3):260–267.

Mikusheva, A. 2010. Robust confidence sets in the presence of weak instruments. *Journal of Econometrics*, 157(2):236–247.

Moreira, M., Porter, J., and Suarez, G. 2009. Bootstrap validity for the score test when instruments may be weak. *Journal of Econometrics*, 149(1):52–64.

Morris, A., Voight, B., Teslovich, T., et al. 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, 44(9):981–990.

Nagar, A. 1959. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica: Journal of the Econometric Society*, 27(4):575–595.

Nead, K. T., Sharp, S. J., Thompson, D. J., et al. 2015. Evidence of a causal association between insulinemia and endometrial cancer: a Mendelian randomization analysis. *Journal of the National Cancer Institute*, 107(9):djv178.

Nelson, C. and Startz, R. 1990. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *Journal of Business*, 63(1):125–140.

Pierce, B. and Burgess, S. 2013. Efficient design for Mendelian randomization studies: subsample and two-sample instrumental variable estimators. *American Journal of Epidemiology*, 178(7):1177–1184.

Rietveld, C. A., Medland, S. E., Derringer, J., et al. 2013. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, 340(6139):1467–1471.

Staiger, D. and Stock, J. 1997. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stock, J., Wright, J., and Yogo, M. 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics*, 20(4):518–529.

Stock, J. and Yogo, M. 2002. Testing for weak instruments in linear IV regression. *SSRN eLibrary*, 11:T0284.

Sussman, J., Wood, R., and Hayward, R. 2010. An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials. *British Medical Journal*, 340:c2073.

Taylor, A., Davies, N., Ware, J., VanderWeele, T., Davey Smith, G., and Munafò, M. 2014. Mendelian randomization in health research: Using appropriate genetic variants and avoiding biased estimates. *Economics & Human Biology*, 13:99–106.

The CARDIoGRAMplusC4D Consortium 2013. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nature Genetics*, 45(1):25–33.

The Global Lipids Genetics Consortium 2013. Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, 45:1274–1283.

Vansteelandt, S., Bowden, J., Babanezhad, M., and Goetghebeur, E. 2011. On instrumental variables estimation of causal odds ratios. *Statistical Science*, 26(3):403–422.

Venables, W. 1975. Calculation of confidence intervals for noncentrality parameters. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(3):406–412.

Zhang, C., Doherty, J. A., Burgess, S., et al. 2015. Genetic determinants of telomere length and risk of common cancers: a Mendelian randomization study. *Human Molecular Genetics*, 24(18):5356–5366.

Figure legend

Figure 1: Mean two-stage least squares/inverse-variance weighted estimates plotted against sample overlap for different values of instrument strength ($\alpha = 0.4$, circle; $\alpha = 0.6$, triangle; $\alpha = 0.8$, plus) and different values of the confounder effect on the outcome ($\beta_U = 0.6$, black solid line; $\beta_U = 1$, mid-grey dashed line; $\beta_U = 2$, light-grey dotted line). Left panel: positive causal effect ($\beta_X = 0.2$); right panel: null causal effect ($\beta_X = 0$).

Web Appendix

A1 Software code – estimation of bias

We provide R code to implement the formulae for the estimation of bias under the null and Type 1 error rate discussed in this paper. These formulae assume that there are multiple genetic variants; bias is a single genetic variant is unlikely to be substantial [Angrist and Pischke, 2009]. We assume that the units for the risk factor and outcome (for a continuous outcome) are the same for the ordinary least squares (OLS) and instrumental variable (IV) estimates; the simplest case is when all variables are in standard deviation units (so `var_x` and `var_y` are both 1).

```
expf = (N-K-1)/K * rsq/(1-rsq) # expf is expected value of the F statistic
                               # N is sample size
                               # K is number of genetic variants
                               # rsq is the expected value of R^2
                               # (otherwise expf can be specified directly)
bias = olsbias*overlap.prop*(1/expf)
                               # bias is the bias of the IV estimate under the null
                               # olsbias is the bias of the OLS estimate (observational
                               #  estimate for binary outcome)
                               # overlap.prop is the proportion of overlap
                               #  between the samples (between 0 and 1)
var  = var_y/(N*var_x*rsq)     # var is the variance of the IV estimate
                               # (continuous outcome)
                               # var_x is the variance of the risk factor
                               # var_y is the variance of the outcome
var  = 1/(N*rsq*var_x*prop.case*(1-prop.case))
                               # var is the variance of the IV estimate
                               # (binary outcome)
                               # prop.case is the proportion of cases (between 0 and 1)
type1err = 2-pnorm(1.96+bias/sqrt(var))-pnorm(1.96-bias/sqrt(var))
                               # type1err is the estimated Type 1 error rate
                               #  rate under the null for a nominal 5% two-sided
                               #  significance level
```

This code is implemented in a web application at `https://sb452.shinyapps.io/overlap`.

A2 Further validation of the analytic formulae for bias and Type 1 error rate

To further assess the validity of the analytic formulae for the bias and Type 1 error rate, we simulated data using the same data-generating model (2) under the causal null ($\beta_X = 0$) for a different level of confounding ($\beta_U = 1$), a different number of IVs ($K = 10$), and a different range of values of the IV strength ($\alpha = 0.02, 0.03, 0.05$) for cases where there is 0% overlap, increasing in increments of 10% up to a 100% overlap. In each case, we report the mean 2SLS (or equivalently IVW) estimate and empirical Type 1 error rate from 10 000 simulations, and the estimated bias and Type 1 error rate from the analytic formulae above, using mean values of the OLS estimate and F statistic estimated across the simulations (the relative bias is estimated as the reciprocal of the average value of the F statistic across simulations – this would not be available for a single dataset in practice).

Results are given in Table A1. The Monte Carlo standard error for the observed mean estimate is around 0.002, and for the Type 1 error rate is around 0.2 to 0.3%. We see that the analytical formulae slightly overestimate the observed bias and Type 1 error inflation. However, the estimated bias and Type 1 error rate are close to the observed values throughout.

	$\alpha = 0.02$		$\alpha = 0.03$		$\alpha = 0.05$	
	Mean $F = 1.8$		Mean $F = 2.9$		Mean $F = 6.3$	
	Mean OLS = 0.500		Mean OLS = 0.499		Mean OLS = 0.497	
Percentage	Expected	Observed	Expected	Observed	Expected	Observed
overlap	Bias (Error)	Mean (Error)	Bias (Error)	Mean (Error)	Bias (Error)	Mean (Error)
0%	0.000 (5.0)	-0.002 (4.9)	0.000 (5.0)	0.002 (5.0)	0.000 (5.0)	-0.002 (5.2)
10%	0.027 (5.2)	0.024 (4.9)	0.017 (5.1)	0.018 (5.3)	0.008 (5.0)	0.005 (5.1)
20%	0.054 (5.6)	0.049 (5.5)	0.035 (5.4)	0.034 (5.2)	0.016 (5.2)	0.012 (5.2)
30%	0.082 (6.4)	0.076 (6.2)	0.052 (5.9)	0.049 (5.8)	0.024 (5.4)	0.018 (5.4)
40%	0.109 (7.5)	0.103 (7.4)	0.069 (6.6)	0.064 (6.8)	0.032 (5.7)	0.026 (5.5)
50%	0.136 (9.0)	0.128 (8.2)	0.086 (7.5)	0.079 (7.2)	0.040 (6.1)	0.033 (5.7)
60%	0.163 (10.8)	0.153 (9.6)	0.104 (8.6)	0.096 (7.9)	0.048 (6.6)	0.040 (6.1)
70%	0.191 (12.9)	0.178 (11.3)	0.121 (10.0)	0.111 (9.1)	0.056 (7.2)	0.047 (6.6)
80%	0.218 (15.4)	0.205 (13.7)	0.138 (11.5)	0.126 (10.6)	0.064 (7.9)	0.055 (6.8)
90%	0.245 (18.3)	0.231 (16.3)	0.156 (13.3)	0.143 (12.2)	0.072 (8.7)	0.063 (7.8)
100%	0.272 (21.4)	0.258 (18.3)	0.173 (15.3)	0.159 (13.5)	0.079 (9.6)	0.068 (8.4)

Web Table A1: Comparison of the expected bias with a null causal effect ($\beta_X = 0$) and Type 1 error rate (5% significance level) calculated using formulae (3) and (4) with the observed mean and empirical Type 1 error rate from a simulation study for a two-sample IV estimate with sample overlap and 10 IVs.

We also repeated the simulation of Table II, except with a sample size of 1000 rather than 10 000. Results for $\alpha = 0.01, 0.02, 0.03$ have been omitted, as the instruments are so weak that they are barely associated with the risk factor (even an F statistic of 1.3 corresponds to a two-sided p-value of 0.34 for an F distribution on 20 and 979 degrees of freedom, and an R^2 statistic of 2.7% for a sample size of 1000 is little more than would be explained on average by a variable that was truly independent from the risk factor), and so would not be used as instruments in practice. Generally the expected estimates of relative bias (reciprocal of F parameter) are close to the observed values. However, while Type 1 error rate is overestimated for the weakest of instruments, it is underestimated for large values of α. This may be due to overrejection of the null by the two-stage least squares method using a Wald test with weak instruments [Stock and Yogo, 2002]. There are several solutions to this problem, including the use of Fieller's theorem [Burgess et al., 2015b], inversion of the Anderson–Rubin test statistic [Mikusheva, 2010], and bootstrapping [Moreira et al., 2009] for improved inference properties with weak instruments; none of these methods assume that the IV estimate is normally distributed.

Web Table A2: **Repeat of Simulation 2 with continuous outcome to validate bias and Type 1 error rate formulae with reduced sample size**

α	Mean F	Mean R^2	Mean OLS estimate	Mean IV estimate	Relative bias	(Mean $F)^{-1}$	Empirical Type 1 error	Expected Type 1 error
0.04	1.3	2.7%	0.994	0.747	0.751	0.743	80.6%	84.9%
0.05	1.5	3.0%	0.989	0.651	0.658	0.657	75.0%	78.9%
0.06	1.8	3.5%	0.986	0.560	0.568	0.570	67.7%	71.0%
0.07	2.0	4.0%	0.980	0.484	0.494	0.490	62.5%	62.4%
0.08	2.3	4.5%	0.974	0.418	0.429	0.428	54.9%	54.6%
0.09	2.7	5.2%	0.967	0.355	0.367	0.369	48.6%	46.7%
0.10	3.1	6.0%	0.959	0.306	0.319	0.321	42.3%	40.2%
0.15	5.7	10.5%	0.913	0.161	0.177	0.174	24.5%	20.8%
0.20	9.4	16.1%	0.856	0.098	0.115	0.106	17.2%	13.0%

Simulation results with null causal effect $\beta_X = 0$, and confounder effect $\beta_U = 2$ to estimate the relative bias and empirical Type 1 error rate (5% nominal significance level) of the two-stage least squares (or equivalently, inverse-variance weighted) instrumental variable (IV) estimate; the relative bias is the bias of the IV estimate divided by the bias of the ordinary least squares (OLS) estimate. The relative bias is theoretically predicted to be close to the reciprocal of the mean value of the F statistic, (Mean $F)^{-1}$. This table is a repeat of Table II from the main paper, except with a smaller sample size.

We also repeated the simulation of Table II, except with a binary risk factor. This was implemented by first simulating the risk factor as in equation (2), and then dichotomizing the risk factor at its median to take the value 0 or 1. Associations with the risk factor were still estimated using a linear model in the two-stage least squares method. Web Table A3 shows marked differences in the mean OLS and IV estimates compared with those in Table II; however, the expected estimates of relative bias and Type 1 error rate were close to their observed values for all strengths of instrument.

Web Table A3: **Repeat of Simulation 2 with continuous outcome to validate bias and Type 1 error rate formulae with binary risk factor**

α	Mean F	Mean R^2	Mean OLS estimate	Mean IV estimate	Relative bias	(Mean F)$^{-1}$	Empirical Type 1 error	Expected Type 1 error
0.01	1.1	0.2%	2.256	1.995	0.884	0.882	66.6%	66.3%
0.02	1.5	0.3%	2.254	1.449	0.643	0.654	52.8%	52.1%
0.03	2.2	0.4%	2.252	1.004	0.446	0.455	37.9%	39.4%
0.04	3.1	0.6%	2.249	0.687	0.305	0.318	27.0%	28.5%
0.05	4.3	0.9%	2.245	0.494	0.220	0.231	20.3%	22.0%
0.06	5.8	1.1%	2.240	0.368	0.164	0.172	16.0%	17.9%
0.07	7.5	1.5%	2.234	0.280	0.125	0.133	13.2%	14.9%
0.08	9.5	1.9%	2.226	0.217	0.097	0.106	11.3%	11.8%
0.09	11.7	2.3%	2.219	0.176	0.079	0.086	10.0%	10.7%
0.10	14.2	2.8%	2.211	0.142	0.064	0.071	9.0%	9.7%
0.15	30.1	5.7%	2.157	0.071	0.033	0.033	6.7%	7.5%
0.20	51.4	9.3%	2.087	0.038	0.018	0.019	5.9%	5.9%

Simulation results with null causal effect $\beta_X = 0$, and confounder effect $\beta_U = 2$ to estimate the relative bias and empirical Type 1 error rate (5% nominal significance level) of the two-stage least squares (or equivalently, inverse-variance weighted) instrumental variable (IV) estimate; the relative bias is the bias of the IV estimate divided by the bias of the ordinary least squares (OLS) estimate. The relative bias is theoretically predicted to be close to the reciprocal of the mean value of the F statistic, (Mean F)$^{-1}$. This table is a repeat of Table II from the main paper, except with a binary risk factor.

A3 Software code – bound for the F parameter

The lower bound of a one-sided 95% confidence interval for the F parameter can be obtained by: i) finding the non-centrality parameter for which the given value of the F statistic is the mean of the non-central F distribution; ii) finding the value such that the cumulative distribution function for that non-central F distribution is 5%.

We demonstrate this procedure with the example from the paper of body mass index in which the F statistic is 97.02, and the degrees of freedom for the F distribution are $\nu_1 = 97$ and $\nu_2 = 339\,127$. The mean of the non-central F distribution on (ν_1, ν_2) degrees of freedom with non-centrality parameter λ (for $\nu_2 > 2$) is:

$$\mathbb{E}(F) = \frac{\nu_2(\nu_1 + \lambda)}{\nu_1(\nu_2 - 2)}.$$

In this case, the non-centrality parameter is:

$$\lambda = \frac{97 \times 97.02 \times 339\,125}{339\,127} - 97 = 9313.88$$

The cumulative distribution function of the non-central F distribution can be calculated using the R code:

```
pf(value, df1=97, df2=339127, ncp=9313.884)
```

We find the value such that the distribution function takes the value 0.05 by an iterative search:

```
> pf(94, df1=97, df2=339127, ncp=9313.884)
[1] 0.06549997
> pf(93, df1=97, df2=339127, ncp=9313.884)
[1] 0.02182131
> pf(93.5, df1=97, df2=339127, ncp=9313.884)
[1] 0.03890898
> pf(93.75, df1=97, df2=339127, ncp=9313.884)
[1] 0.05084209
> pf(93.625, df1=97, df2=339127, ncp=9313.884)
[1] 0.04455657
> pf(93.6875, df1=97, df2=339127, ncp=9313.884)
[1] 0.04761685
> pf(93.71875, df1=97, df2=339127, ncp=9313.884)
[1] 0.04920851
> pf(93.734375, df1=97, df2=339127, ncp=9313.884)
[1] 0.05002001
```

Hence the lower limit of the one-sided 95% confidence interval is approximately 93.734.

The following function performs this search automatically:

```
findlowerflimit <- function(f, nu1, nu2) {
 lambda = f*nu1*(nu2-2)/nu2-nu1
 lower = f-1
 while (pf(lower, df1=nu1, df2=nu2, ncp=lambda)>0.05) { lower = lower-1 }
 upper = lower + 1
 while ( abs(pf((lower+upper)/2, df1=nu1, df2=nu2, ncp=lambda)-0.05)>0.0001 ) {
 if (pf((lower+upper)/2, df1=nu1, df2=nu2, ncp=lambda)>0.05)
    { upper = (lower+upper)/2 }
 if (pf((lower+upper)/2, df1=nu1, df2=nu2, ncp=lambda)<0.05)
    { lower = (lower+upper)/2 }
 }
 return((lower+upper)/2)
 }
```

Running this code gives:

```
> findlowerflimit(97.02, 97, 339127)
[1] 93.73484
```

Hence the lower limit of the one-sided 95% confidence interval is approximately 93.735. Further precision can be obtained if required, but 1 decimal place should be enough to give a good estimate of the potential bias.

For comparison, the corresponding lower limits for educational attainment are:

1. Discovery sample: observed F statistic = 20.2, sample size = 101 069, number of genetic variants = 5, lower limit for F statistic = 14.0.

2. Follow-up sample: observed F statistic = 5.1, sample size = 25 490, number of genetic variants = 5, lower limit for F statistic = 2.3.

In these cases, the lower one-sided 95% limit for the F statistic is substantially lower than the observed value of the F statistic.

A4 Software code – allele score method using summarized data with arbitrary weights

The estimate from an inverse-variance weighted method for uncorrelated genetic variants is equal to that from of an allele score method with internally-derived weights (also equal to that from a two-stage least squares method). The inverse-variance weighted method can be modified to give the same estimate as that from an allele score method with arbitrary weights [Burgess et al., 2016b]. In particular, it can provide the same estimate as that from an allele score method with equal weights.

Inputs are the genetic associations with the risk factor (`bx`, standard error `bxse`), the genetic associations with the outcome (`by`, standard error `byse`), and the allele score weights (`wts`):

```
wts      = rep(1, length(bx))
   # for equal weights, otherwise specify weights
beta_SSw = sum(wts*by/byse^2)/sum(wts*bx/byse^2)
   # equivalent to original IVW method when wts = bx
se_SSw_first   = sqrt(sum(wts^2/byse^2)/sum(wts*bx/byse^2))
   # standard error from delta method (first-order approximation)
se_SSw_second  = sqrt(sum(wts^2/byse^2)/sum(wts*bx/byse^2)^2 +
  sum(wts*by/byse^2)^2/sum(wts*bx/byse^2)^4*sum(wts^2/byse^2) -
  2*theta*sum(wts*by/byse^2)/sum(wts*bx/byse^2)^3)
   # standard error from delta method (second-order approximation)
   # theta is the correlation between the numerator and denominator of the estimate
   # if the correlation is not known, it can be taken as the observational
   # correlation between the risk factor and outcome;
   # a sensitivity analysis can also be performed for its value
```

If the genetic variants are correlated, then a valid test statistic for the causal null hypothesis can be derived:

```
beta_SSw_cor = sum(wts*by/byse^2)/sum(wts^2/byse^2)
se_SSt_cor   = sqrt(sum((wts*byse^-1)%o%(wts*byse^-1)*rho)/
  sum(wts^2/byse^2)^2)
   # standard error from delta method (first-order approximation)
   # rho is the matrix of correlations between the genetic variants
```

This test statistic can also be used with uncorrelated genetic variants [Burgess et al., 2016b], although it does not have an interpretation as an effect estimate with either correlated or uncorrelated variants.