Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2011;40:740–752
doi:10.1093/ije/dyq151

THEORY AND METHODS

Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants

Brandon L Pierce,[1]* Habibul Ahsan[1] and Tyler J VanderWeele[1,2,3]

[1]Department of Health Studies, University of Chicago, Chicago, IL, USA, [2]Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA and [3]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

*Corresponding author. Center for Cancer Epidemiology and Prevention, Department of Health Studies, University of Chicago, 5841 S. Maryland Ave., MC2007 Chicago, IL 60637, USA. E-mail: bpierce@health.bsd.uchicago.edu

Accepted 26 July 2010

Background Mendelian Randomization (MR) studies assess the causality of an exposure–disease association using genetic determinants [i.e. instrumental variables (IVs)] of the exposure. Power and IV strength requirements for MR studies using multiple genetic variants have not been explored.

Methods We simulated cohort data sets consisting of a normally distributed disease trait, a normally distributed exposure, which affects this trait and a biallelic genetic variant that affects the exposure. We estimated power to detect an effect of exposure on disease for varying allele frequencies, effect sizes and samples sizes (using two-stage least squares regression on 10 000 data sets—Stage 1 is a regression of exposure on the variant. Stage 2 is a regression of disease on the fitted exposure). Similar analyses were conducted using multiple genetic variants (5, 10, 20) as independent or combined IVs. We assessed IV strength using the first-stage F statistic.

Results Simulations of realistic scenarios indicate that MR studies will require large ($n > 1000$), often very large ($n > 10\,000$), sample sizes. In many cases, so-called 'weak IV' problems arise when using multiple variants as independent IVs (even with as few as five), resulting in biased effect estimates. Combining genetic factors into fewer IVs results in modest power decreases, but alleviates weak IV problems. Ideal methods for combining genetic factors depend upon knowledge of the genetic architecture underlying the exposure.

Conclusions The feasibility of well-powered, unbiased MR studies will depend upon the amount of variance in the exposure that can be explained by known genetic factors and the 'strength' of the IV set derived from these genetic factors.

Keywords Mendelian randomization, instrumental variable analysis, power, weak instrument, causal inference, two-stage least squares regression

Introduction

Mendelian Randomization (MR) is a method used to test or estimate a causal effect of an exposure on a disease outcome when the exposure has a known genetic determinant.[1,2] Data on such a genetic determinant, also known as an instrumental variable (IV), can be analysed jointly with exposure and outcome data to determine if an observed exposure–disease association is causal. Typically, MR is utilized when an observed exposure–outcome association is potentially attributable, at least in part, to confounding or reverse causation. In the absence of these phenomena, simple measures of association provide more precise effect estimates.

A valid MR analysis requires that the IV is (i) associated with the exposure, (ii) independent of the outcome given the exposure and confounders of the exposure–outcome association and (iii) independent of factors that confound the exposure–outcome relationship. Because genetic variation is randomly assigned prior to conception, it is not expected to be affected by any confounding factors other than ancestry, which can be accurately measured and accounted for using genetic data.[3] Consequently, if the IV affects the outcome only through the exposure, the unconfounded effect of the exposure on the outcome can be captured by comparing the effect of the IV on exposure to the indirect effect of the IV on the outcome. IV analysis is common in the econometrics literature,[4] although MR is a specific type of IV analysis in which is the IVs are genetic variants.

MR studies are becoming more feasible, as recent genome wide association (GWA) studies have identified genetic determinants (typically single nucleotide polymorphisms) for many health-related biomarkers. For example, GWA studies have linked high-density lipoprotein to 11 loci, low-density lipoprotein to 14 loci and triglycerides to 11 loci.[5–7] Genetic determinants have also been identified for C-reactive protein;[8,9] plasma levels of vitamins B,[10,11] A[12] and E;[12] mean platelet volume;[13,14] blood pressure[15,16] and fasting plasma glucose.[17–20] Several recent publications highlight the potential for GWA studies of many biomarkers simultaneously.[21–24] Such parallel GWA studies could expand our knowledge of biomarker-related polymorphisms very quickly, and as our understanding of these polymorphisms' functions improve, the broad application of MR methodology to epidemiological and biomarker research may become feasible.

In light of these developments, we have conducted a simulation study evaluating power and IV strength requirements for MR analyses using two-stage least squares (2SLS) regression, the most common statistical method used in MR studies of continuous exposures and continuous outcomes.[25,26] We provide power calculations for IVs of varying strength and exposures of varying effect, under realistic scenarios given current knowledge. We explore the use of multiple genetic variants in MR studies, employing different strategies to combine information across variants and evaluating the consequences of these strategies on power and overall IV strength, as measured by the first-stage F statistic in 2SLS. Weak IVs lead to biased effect estimates in the presence of confounding of the exposure–outcome relationship.[27–30] We intend for this work to guide researchers in their design of future MR studies, in selecting appropriate genetic variants, constructing strong IVs and obtaining adequate sample sizes.

Methods

Power estimates were generated using simulated data sets of samples drawn from a genetically homogenous population. Each simulation consisted of 10 000 data sets containing one or more biallelic loci (G) in Hardy–Weinberg equilibrium, a continuous exposure (X) affected by G and a continuous outcome (Y) affected positively by X. Each power estimate was obtained by applying 2SLS to all 10 000 simulated data sets and determining the percentage of data sets in which a positive effect of the fitted X on Y was observed using a two-sided significance test ($\alpha = 0.05$). The effect estimate from each simulated data set was also retained. Stage 1 of the 2SLS is a regression of the X on the IV(s) (G). Stage 2 is a regression of Y on the fitted X-values from Stage 1. In other words, only the variation in X that is explained by the IV(s) is used in Stage 2.

We extract two key parameters from each Stage 1 regression: R^2 and the F statistic. R^2 is the proportion of variability in the X that is explained by G, an indicator of power for MR studies. F reflects the 'strength' of an IV or a set of IVs. In the presence of X–Y confounding, 2SLS effect estimates will be biased towards the confounded X–Y association, but the size of the relative bias is inversely related to F ('relative bias' is defined as the ratio of the 2SLS bias to the bias of the confounded X–Y association).[29,30] In the MR literature, a threshold of $F < 10$ has typically been used to define a 'weak IV' (the Staiger–Stock rule[31]). This rule of thumb is based on the observation that an F value greater than ~11 ensures that relative bias will be <10% at least 95% of the time, regardless of the number of IVs used in the analysis.[29] The F statistic is defined as the ratio of the mean square of the model to the mean square of the error. However, F can be expressed as a function of the first-stage R^2, the sample size (n) and the number of IVs (k):

$$F = \frac{R^2(n-1-k)}{(1-R^2)k} \quad (1)$$

Thus, F increases as R^2 and n increase, but F decreases as k increases. We also extract the 'adjusted R^2' from each 2SLS regression, a modification of R^2 that adjusts for inflation due to large numbers of predictors

and small sample sizes.[32] The adjusted R^2 is defined as:

$$1-(1-R^2)\frac{n-1}{n-k-1} \quad (2)$$

Thus, for a given R^2, the adjusted R^2 decreases as the number of IVs (k) increases and approaches R^2 as the sample size increases. For simulations in which the R^2 and/or the adjusted R^2 were held constant, the appropriate effect sizes were found using an iterative search process.

Data simulation 1: Power estimates for MR studies using one genetic variant

Each simulated data set consisted of Y, X and a single G as an IV, both with and without an unobserved confounding variable, U (Figure 1a). The genotype at G was randomly generated, assuming a minor allele frequency (MAF) of 0.3 and an effect size (β_{gx}) that resulted in a specific R^2 value (either 0.005, 0.01, 0.05 or 0.10). For a fixed R^2, varying the MAF and the effect size does not affect power or F. G was coded as −1, 0 or 1, representing the presence of 0, 1 or 2X-increaser alleles, respectively. X was modelled as a random number drawn from a standard normal distribution plus an additive effect of alleles at G:

$$x_i = \beta_{gx} g_i + \varepsilon_i \text{ with } \varepsilon_i \sim N(0,1) \quad (3)$$

where β_{gx} represents the effect of G on X. Similarly, Y was modelled as a random number drawn from a standard normal distribution plus the linear effect of X:

$$y_i = \beta_{xy} g_i + \varepsilon_i \text{ with } \varepsilon_i \sim N(0,1) \quad (4)$$

with β_{xy} set to 0.1, 0.3 or 0.5. To explore the effects of unmeasured confounding on the power and the effect estimates, we generated similar data sets, but introduced a confounding variable (U), which affects both X and Y. U is a random number drawn from a standard normal distribution. X was modelled as

$$x_i = \beta_{gx} g_i + 0.5u_i + \varepsilon_i \text{ with } \varepsilon_i \sim N(0,1). \quad (5)$$

Similarly, Y was modelled as

$$y_i = \beta_{xy} g_i + 0.5u_i + \varepsilon_i \text{ with } \varepsilon_i \sim N(0,1). \quad (6)$$

Power estimates and median 2SLS estimates were obtained for each unique combination of the parameters R^2 and β_{xy} over a range of sample sizes ($n = 500$, 1000, 5000 and 10 000). For the scenarios where confounding was introduced, U was not used in the 2SLS analysis, as it is assumed that this is an unmeasured confounder.

Data simulation 2: Power estimates for single and multiple variant scenarios

To compare power estimates and F values between MR analyses using single and multiple genetic variants as independent IVs, we first generated power estimates assuming a single variant, G, with an MAF of 0.3 and effect sizes corresponding to specific adjusted R^2 values: 0.01, 0.05 and 0.10. For these adjusted R^2 values, we chose samples sizes of 5000, 1000 and 500, respectively, resulting in a broad range of power estimates for each adjusted R^2 value.

We then generated power estimates for scenarios with 5, 10 or 20 variants as independent IVs (Figure 1b), assuming equal effects for each variant on X and holding either R^2 or adjusted R^2 equal to the single-IV scenarios. For the five-variant model, G is a matrix containing five variants with X-increaser allele frequencies of 0.1, 0.3, 0.5, 0.7 and 0.9. These same frequencies (0.1, 0.3, 0.5, 0.7 and 0.9) were used for

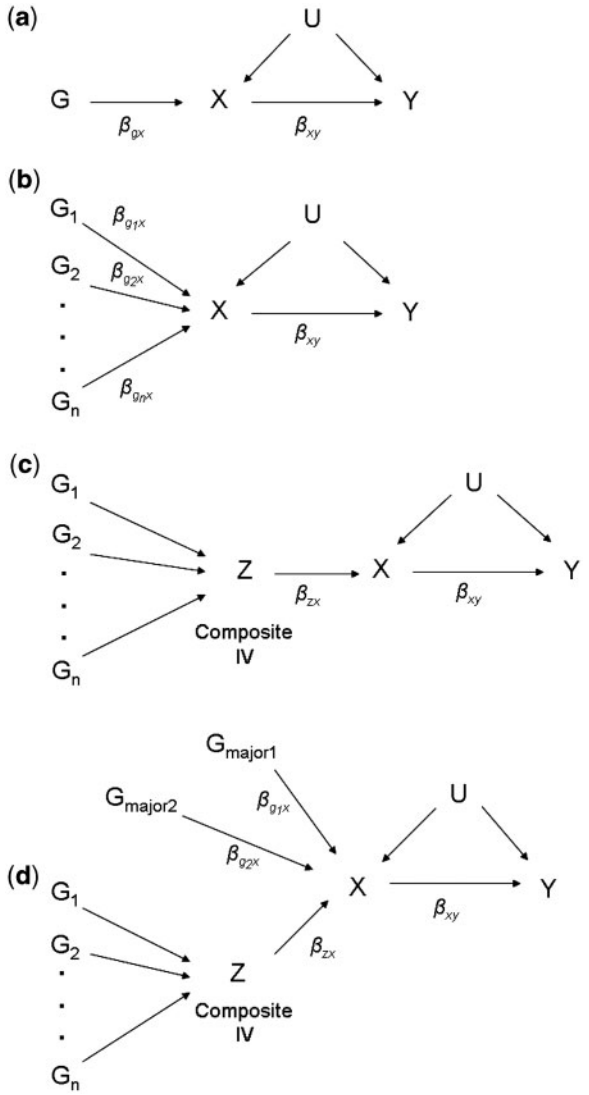

Figure 1 Causal diagram for a Mendelian randomization study using (a) a single instrumental variable, (b) multiple, independent instrumental variables (c) a single combined instrumental variable and (d) a major gene and polygene instrumental variables

the 10- and 20-variant models, with each variant represented two and four times, respectively. For all scenarios, power estimates and mean F statistics were generated both in the presence and absence of the unobserved confounding variable introduced in Data simulation 1.

These analyses allowed us to compare the multivariant scenarios to the single-variant scenarios in two ways: holding R^2-adjusted constant and holding R^2 constant. This was done to demonstrate which R^2 measure was more closely related to overall power across all single- and multivariant scenarios.

Data simulation 3: Power estimates for multivariant scenarios where variants are combined to reduce the number of IVs

To investigate the consequences of combining multiple variants into a composite IV, we simulated data sets using three distinct models regarding the genetic architecture of X: (i) equal effects on X for each variant in matrix G, (ii) a continuum of effects on X for the variants contained in G and (iii) major-gene and polygene effects on X for the variants contained in G. All data sets were generated both with and without the unobserved confounding variable U introduced in Data simulation 1.

For the 'equal effects' model, we simulated three types of data sets containing either 5, 10 or 20 variants of equal effect. The effect size, β_{gx}, was chosen to produce an adjusted R^2 of 0.05 when using all the variants as independent IVs in one 2SLS regression. Allele frequencies were identical to Data simulation 2.

For the 'continuum of effects' model, we again simulated three types of data sets, containing either 5, 10 or 20 variants that affect X. However, in this model, a range of effect sizes were assigned to the variants in the matrix G. These effects of these variants were weighted to be 0.5, 0.75, 1.0, 1.25 or 1.5 times as large as some reference allelic β_{gx} value, which was chosen to result in an adjusted R^2 of 0.05 when using all the variants as an independent IV in one 2SLS regression. Each of the five weights were represented once, twice or four times in the set of 5, 10 or 20 variants, respectively.

The 'major-gene/polygene' model consisted of 10 variants, two of which were assigned strong effects on X (i.e. major genes); the remaining eight variants were assigned weaker effects (i.e. polygenes), each of equal magnitude, defined as the reference β_{gx}. The two major-gene effects were five and three times as strong as the reference β_{gx}. The reference β_{gx} was chosen to result in an adjusted R^2 of 0.10 when using all 10 variants as independent IVs in one 2SLS regression.

For the equal effects model, 2SLS power analyses were conducted using each variant as an independent IV in one regression (Figure 1b) and using a single composite IV—an allele count (i.e. total number of X-increaser alleles present for all variants; Figure 1c). Identical analyses were conducted for the 'continuum of effects' model, but with an additional composite IV—a weighted allele count—where each X-increaser allele present is weighted by its true effect on X before the alleles are summed. Applications of the weighted allele count method would require previous knowledge of the magnitudes of the associations between the Gs and X. For the major-gene/polygene model, we conducted analyses identical to those for the 'continuum of effects' model, but included an additional analysis using three IVs in a single 2SLS regression: two independent IVs for the major-gene effects and a single allele count IV for the polygenic effects (Figure 1d).

Data simulation 4: Power for MR under different types of unmeasured confounding

We conducted additional simulations under both positive and negative confounding, varying the strength of the effect of the unobserved confounding variable U on X and Y. We explored the effects of these various confounding scenarios using the equal effects model from Data simulation 3, using a sample size of 1000 and an adjusted R^2 of 0.05.

Results

Power estimates for detecting significant causal effects (two-sided $\alpha = 0.05$) in single-variant MR studies are presented in Table 1. The estimated power increases as the asymptotic R^2, β_{xy} and n increase. The first-stage F increases as the asymptotic R^2 and n increase. Weak IV scenarios are shaded in grey, with darker grey corresponding to lower F values and more bias away from the null when positive X–Y confounding is present. Adequately powered scenarios (power >80%) are shown in bold, none of which have a weak IV problem (Table 1).

Power for MR studies using multiple variants as independent IVs is shown in Table 2. When the adjusted R^2 is held constant, power and unadjusted R^2 increase as the number of variants (i.e. IVs) increases, while F decreases. When R^2 is held constant under no X–Y confounding, there is no impact on power when the number of variants is increased; however, the adjusted R^2 and F decrease as the number of IVs increases. Each multi-IV scenario has a weak IV problem (shown in grey), resulting in upward bias in the effect estimates in the presence of confounding, especially for very low F values (dark grey). This bias results in power estimates that are artificially inflated compared with the corresponding unconfounded (and unbiased) scenario.

Comparisons of multivariant MR studies using independent IVs vs a composite IV are shown in Table 3. For the unconfounded scenarios, combining variants of equal effect into an X-increaser 'allele count' IV

Table 1 Power estimates, median effect estimates and instrument strength *(F)* for Mendelian randomization studies using one genetic variant

Sample size Asymptotic $R^{2,a}$	F^b	Unadj. mean R^2	No *X–Y* confounding				*X–Y* confounding			
			Power estimate[c] (Median effect estimate[c])				Power estimate[c] (Median effect estimate[c])			
			β_{xy}				β_{xy}			
			0.0	0.1	0.3	0.5	0.0	0.1	0.3	0.5
$n = 500$										
0.005	3	0.007	0.00 (0.00)	0.00 (0.10)	0.03 (0.30)	0.09 (0.50)	0.01 (0.02)	0.02 (0.13)	0.06 (0.33)	0.14 (0.52)
0.01	6	0.012	0.00 (0.00)	0.01 (0.10)	0.06 (0.30)	0.18 (0.50)	0.01 (0.00)	0.03 (0.11)	0.11 (0.32)	0.23 (0.51)
0.05	27	0.052	0.01 (0.00)	0.07 (0.10)	0.35 (0.30)	0.70 (0.50)	0.03 (0.00)	0.08 (0.10)	0.35 (0.30)	0.68 (0.50)
0.10	56	0.102	0.02 (0.00)	0.11 (0.10)	0.60 (0.30)	0.94 (0.50)	0.03 (0.00)	0.12 (0.10)	0.59 (0.30)	0.92 (0.50)
$n = 1000$										
0.005	6	0.006	0.00 (0.00)	0.01 (0.10)	0.06 (0.30)	0.18 (0.50)	0.01 (0.01)	0.03 (0.11)	0.11 (0.31)	0.23 (0.51)
0.01	11	0.011	0.01 (0.00)	0.03 (0.10)	0.14 (0.30)	0.35 (0.50)	0.02 (0.00)	0.05 (0.10)	0.18 (0.30)	0.37 (0.50)
0.05	54	0.051	0.02 (0.00)	0.11 (0.10)	0.58 (0.30)	**0.92 (0.50)**	0.03 (0.00)	0.12 (0.10)	0.58 (0.30)	**0.90 (0.50)**
0.10	113	0.101	0.02 (0.00)	0.18 (0.10)	**0.87 (0.30)**	**1.00 (0.50)**	0.03 (0.00)	0.19 (0.10)	**0.86 (0.30)**	**1.00 (0.50)**
$n = 5000$										
0.005	27	0.005	0.02 (0.00)	0.07 (0.10)	0.32 (0.30)	0.68 (0.50)	0.03 (0.00)	0.08 (0.10)	0.34 (0.30)	0.67 (0.50)
0.01	52	0.010	0.02 (0.00)	0.10 (0.10)	0.58 (0.30)	**0.92 (0.50)**	0.03 (0.00)	0.12 (0.10)	0.57 (0.30)	**0.90 (0.50)**
0.05	263	0.050	0.02 (0.00)	0.36 (0.10)	**1.00 (0.30)**	**1.00 (0.50)**	0.03 (0.00)	0.37 (0.10)	**1.00 (0.30)**	**1.00 (0.50)**
0.10	559	0.100	0.03 (0.00)	0.65 (0.10)	**1.00 (0.30)**	**1.00 (0.50)**	0.03 (0.00)	0.65 (0.10)	**1.00 (0.30)**	**1.00 (0.50)**
$n = 10\,000$										
0.005	51	0.005	0.02 (0.00)	0.10 (0.10)	0.56 (0.30)	**0.92 (0.50)**	0.03 (0.00)	0.11 (0.10)	0.56 (0.30)	**0.90 (0.50)**
0.01	105	0.010	0.02 (0.00)	0.17 (0.10)	**0.84 (0.30)**	**1.00 (0.50)**	0.03 (0.00)	0.18 (0.10)	**0.83 (0.30)**	**1.00 (0.50)**
0.05	528	0.050	0.02 (0.00)	0.63 (0.10)	**1.00 (0.30)**	**1.00 (0.50)**	0.03 (0.00)	0.62 (0.10)	**1.00 (0.30)**	**1.00 (0.50)**
0.10	>999	0.100	0.03 (0.00)	**0.92 (0.10)**	**1.00 (0.30)**	**1.00 (0.50)**	0.02 (0.00)	**0.91 (0.10)**	**1.00 (0.30)**	**1.00 (0.50)**

[a]This can also be interpreted as the 'adjusted R^{2}' for the regression of the exposure (X) on the genetic variant (G).
[b]Scenarios with weak instrument problems (i.e. low F values) are shown in grey, with darker shading corresponding to lower F values and more bias in the presence of *X–Y* confounding.
[c]Power and median effect estimates were obtained using 10 000 simulations. Scenarios with power >80% are shown in bold. Median effect estimates are shown in parentheses below the power estimates.

resulted in slight decreases in R^2 and power, when compared with using each G as an independent IV; these decreases becomes more substantial as the number of variants increases. In contrast, combining the variants in to a single IV has no impact on the adjusted R^2 values, which are identical to the values from the multi-IV scenario. Each multi-IV scenario had a weak IV problem (shaded in grey), again resulting in bias-inflated power estimates under confounding. Using an allele count IV resulted in large increases in F, eliminating the 'weak-IV' bias observed in the presence of *X–Y* confounding.

When combining variants with a continuum of effects into a single IV, using an allele count IV increased F, while resulting in decreases for R^2, adjusted R^2 and power, compared with the multi-IV scenario (Table 3). Using the 'weighted allele count' IV also resulted in large increases in F and decreases in R^2 and power (compared with the multi-IV scenario), although R^2 and power estimates greater than those obtained for the allele count IV. The weighted allele count method produced an adjusted R^2 similar to that obtained from the multi-IV scenario.

When constructing an IV using two variants with 'major-gene' effects and eight 'polygenic' variants of smaller effect, using an allele count resulted in a substantial decrease in R^2 and power and a large increase in F (compared with the 10-IV scenario). Using the weighted allele count IV, power was only slightly lower than when using all 10 IVs and the F statistic was much larger. If the variants with large effects were treated as independent IVs and the remaining polygenic variants were combined into a single allele count IV, power was similar to the weighted allele count IV scenario. The F statistic was smaller in this case (mean $F = 20$), but still >11, resulting in acceptable levels of relative bias (<10%).

Under each model presented in Table 3, introducing positive *X–Y* confounding results in effect estimates that are biased away from the null and bias-inflated power estimates when compared with the identical model with no *X–Y* confounding. Table 4 shows the

Table 2 Power, effect estimate and instrument strength (F) comparisons for single- vs multivariant Mendelian randomization studies holding either R^2 or adjusted R^2 constant (in bold)

No. of variants[a]	Mean Adj. $R^{2,b}$	Mean $R^{2,b}$	F^c	No X–Y confounding					X–Y confounding				
				Allelic β_{gx}	Power estimate[b] (median effect estimate[b])				Allelic β_{gx}	Power estimate[b] (median effect estimate[b])			
					β_{xy}					β_{xy}			
					0.0	0.1	0.3	0.5		0.0	0.1	0.3	0.5
Low R^2 ($n = 5000$)													
1	0.010	0.010	51	0.155	0.02 (0.00)	0.10 (0.10)	0.56 (0.30)	0.93 (0.50)	0.172	0.03 (0.00)	0.12 (0.10)	0.55 (0.30)	0.89 (0.50)
5	**0.010**	0.011	11	0.077	0.02 (0.00)	0.11 (0.10)	0.61 (0.30)	0.94 (0.50)	0.0865	0.03 (0.02)	0.15 (0.11)	0.63 (0.31)	0.93 (0.51)
10	**0.010**	0.012	6	0.055	0.02 (0.00)	0.12 (0.10)	0.64 (0.30)	0.96 (0.50)	0.062	0.05 (0.04)	0.19 (0.13)	0.72 (0.33)	0.96 (0.53)
20	**0.010**	0.014	4	0.039	0.03 (0.00)	0.13 (0.10)	0.70 (0.30)	0.98 (0.50)	0.043	0.07 (0.06)	0.26 (0.16)	0.82 (0.35)	0.99 (0.56)
5	0.009	**0.010**	10	0.073	0.02 (0.00)	0.10 (0.10)	0.56 (0.30)	0.91 (0.50)	0.082	0.04 (0.02)	0.14 (0.12)	0.60 (0.32)	0.91 (0.52)
10	0.008	**0.010**	5	0.049	0.02 (0.00)	0.10 (0.10)	0.57 (0.30)	0.93 (0.50)	0.055	0.05 (0.04)	0.18 (0.14)	0.64 (0.34)	0.93 (0.54)
20	0.006	**0.010**	2	0.030	0.02 (0.00)	0.10 (0.10)	0.56 (0.30)	0.92 (0.50)	0.0335	0.08 (0.08)	0.26 (0.18)	0.73 (0.37)	0.96 (0.58)
Moderate R^2 ($n = 1000$)													
1	0.050	0.051	54	0.354	0.02 (0.00)	0.10 (0.10)	0.58 (0.30)	0.93 (0.50)	0.396	0.03 (0.00)	0.12 (0.10)	0.58 (0.30)	0.90 (0.50)
5	**0.050**	0.055	12	0.176	0.02 (0.00)	0.12 (0.10)	0.61 (0.30)	0.94 (0.50)	0.197	0.03 (0.02)	0.15 (0.11)	0.64 (0.31)	0.93 (0.51)
10	**0.050**	0.059	6	0.1245	0.02 (0.00)	0.12 (0.10)	0.64 (0.30)	0.96 (0.50)	0.139	0.04 (0.03)	0.19 (0.13)	0.72 (0.33)	0.96 (0.53)
20	**0.050**	0.069	4	0.088	0.02 (0.00)	0.13 (0.10)	0.70 (0.30)	0.98 (0.50)	0.098	0.07 (0.05)	0.26 (0.15)	0.82 (0.35)	0.99 (0.55)
5	0.046	**0.051**	11	0.169	0.02 (0.00)	0.11 (0.10)	0.59 (0.30)	0.93 (0.50)	0.189	0.03 (0.02)	0.14 (0.12)	0.61 (0.31)	0.92 (0.52)
10	0.041	**0.051**	5	0.113	0.02 (0.00)	0.10 (0.10)	0.59 (0.30)	0.93 (0.50)	0.1265	0.05 (0.03)	0.18 (0.13)	0.68 (0.34)	0.94 (0.54)
20	0.032	**0.051**	3	0.0695	0.02 (0.00)	0.10 (0.10)	0.58 (0.30)	0.93 (0.50)	0.0775	0.08 (0.07)	0.26 (0.18)	0.76 (0.37)	0.97 (0.58)
High R^2 ($n = 500$)													
1	0.100	0.102	57	0.515	0.03 (0.00)	0.10 (0.10)	0.60 (0.30)	0.94 (0.50)	0.577	0.03 (0.00)	0.12 (0.10)	0.59 (0.30)	0.92 (0.50)
5	**0.100**	0.109	12	0.256	0.02 (0.01)	0.11 (0.10)	0.63 (0.30)	0.95 (0.50)	0.2865	0.03 (0.02)	0.15 (0.11)	0.66 (0.31)	0.95 (0.52)
10	**0.100**	0.118	7	0.181	0.02 (0.00)	0.12 (0.10)	0.67 (0.30)	0.96 (0.50)	0.2025	0.04 (0.03)	0.18 (0.13)	0.74 (0.33)	0.97 (0.53)
20	**0.100**	0.136	4	0.128	0.03 (0.00)	0.14 (0.10)	0.73 (0.30)	0.98 (0.50)	0.143	0.07 (0.05)	0.27 (0.15)	0.83 (0.35)	0.99 (0.55)
5	0.093	**0.102**	11	0.246	0.03 (0.01)	0.11 (0.10)	0.60 (0.30)	0.94 (0.50)	0.275	0.03 (0.02)	0.15 (0.12)	0.64 (0.32)	0.93 (0.51)
10	0.084	**0.102**	6	0.164	0.02 (0.01)	0.11 (0.10)	0.60 (0.30)	0.94 (0.50)	0.1835	0.04 (0.03)	0.19 (0.13)	0.68 (0.33)	0.95 (0.53)
20	0.065	**0.102**	3	0.101	0.02 (0.00)	0.11 (0.10)	0.59 (0.30)	0.93 (0.50)	0.113	0.08 (0.07)	0.26 (0.17)	0.76 (0.37)	0.97 (0.57)

[a]The one-variant models have one variant with an X-increaser allele frequency of 0.3. The five-variant models have X-increaser allele frequencies of 0.1, 0.3, 0.5, 0.7 and 0.9. The 10-variant models have two of each of the following X-increaser allele frequencies 0.1, 0.3, 0.5, 0.7 and 0.9. The 20-variant models have four of each of the following X-increaser allele frequencies 0.1, 0.3, 0.5, 0.7 and 0.9.

[b]Values were obtained using 10 000 simulations. R^2 values correspond to the regression of the exposure (X) on the genetic variant(s) (G) and are held constant to the single-variant scenario when shown in bold. Median effect estimates are shown in parentheses below the power estimates.

[c]Scenarios with weak instrument problems (i.e. low F values) are shown in grey, with darker shading corresponding to lower F values and more bias in the presence of X–Y confounding.

Table 3 Power, effect estimate and instrument strength (F) comparisons for Mendelian randomization studies using independent vs combined instruments, holding the per-allele effect size of G on X constant (in bold)

No. of variants IV(s)	Mean Adj. $R^{2,a}$	Mean $R^{2,a}$	F^b	No X–Y Confounding					X–Y Confounding				
				Allelic β_{gx}[c]	Power estimate[a] (median effect estimate [a])				Allelic β_{gx}	Power estimate[a] (median effect estimate[a])			
					β_{xy}					β_{xy}			
					0.0	0.1	0.3	0.5		0.0	0.1	0.3	0.5
Equal effects (n = 1000)													
5 variants[d]													
5 IVs	0.050	0.055	12	**0.1760**	0.02 (0.00)	0.12 (0.10)	0.61 (0.30)	0.94 (0.50)	**0.197**	0.03 (0.02)	0.15 (0.11)	0.64 (0.31)	0.93 (0.51)
Allele count	0.050	0.051	54	**0.1760**	0.02 (0.00)	0.10 (0.10)	0.59 (0.30)	0.91 (0.50)	**0.197**	0.03 (0.00)	0.12 (0.10)	0.57 (0.30)	0.90 (0.50)
10 variants[e]													
10 IVs	0.050	0.059	6	**0.1245**	0.02 (0.00)	0.12 (0.10)	0.64 (0.30)	0.96 (0.50)	**0.139**	0.04 (0.03)	0.19 (0.13)	0.72 (0.33)	0.96 (0.53)
Allele count	0.050	0.051	54	**0.1245**	0.03 (0.00)	0.09 (0.10)	0.60 (0.30)	0.93 (0.50)	**0.139**	0.03 (0.00)	0.11 (0.10)	0.57 (0.30)	0.91 (0.50)
20 variants[f]													
20 IVs	0.050	0.069	4	**0.0880**	0.02 (0.01)	0.13 (0.10)	0.70 (0.30)	0.98 (0.50)	**0.098**	0.07 (0.05)	0.26 (0.15)	0.82 (0.35)	0.99 (0.55)
Allele count	0.050	0.051	54	**0.0880**	0.02 (0.00)	0.11 (0.10)	0.57 (0.30)	0.92 (0.50)	**0.098**	0.03 (0.00)	0.12 (0.10)	0.57 (0.30)	0.91 (0.50)
A continuum of effects (n = 1000)													
5 variants[d]													
5 IVs	0.050	0.055	12	**0.1895**	0.02 (0.00)	0.10 (0.10)	0.61 (0.30)	0.93 (0.50)	**0.212**	0.03 (0.01)	0.14 (0.11)	0.63 (0.31)	0.93 (0.51)
Allele count	0.044	0.045	47	**0.1895**	0.02 (0.00)	0.10 (0.10)	0.55 (0.30)	0.89 (0.50)	**0.212**	0.03 (0.00)	0.11 (0.10)	0.53 (0.30)	0.87 (0.50)
Weighted count	0.050	0.051	54	**0.1895**	0.03 (0.00)	0.11 (0.10)	0.58 (0.30)	0.92 (0.50)	**0.212**	0.03 (0.01)	0.12 (0.10)	0.57 (0.30)	0.91 (0.50)
10 variants[e]													
10 IVs	0.050	0.060	6	**0.1340**	0.02 (0.00)	0.11 (0.10)	0.64 (0.30)	0.94 (0.50)	**0.150**	0.04 (0.03)	0.18 (0.13)	0.72 (0.33)	0.96 (0.53)
Allele count	0.044	0.045	47	**0.1340**	0.02 (0.00)	0.09 (0.10)	0.53 (0.30)	0.89 (0.50)	**0.150**	0.03 (0.00)	0.11 (0.10)	0.53 (0.30)	0.87 (0.50)
Weighted count	0.050	0.051	54	**0.1340**	0.02 (0.00)	0.11 (0.10)	0.57 (0.30)	0.93 (0.50)	**0.150**	0.03 (0.00)	0.12 (0.10)	0.58 (0.30)	0.91 (0.50)
20 variants[f]													
20 IVs	0.050	0.069	4	**0.0945**	0.03 (0.00)	0.14 (0.10)	0.71 (0.30)	0.98 (0.50)	**0.106**	0.07 (0.05)	0.28 (0.15)	0.83 (0.35)	0.99 (0.55)
Allele count	0.044	0.045	47	**0.0945**	0.02 (0.01)	0.10 (0.10)	0.53 (0.30)	0.88 (0.50)	**0.106**	0.03 (0.00)	0.11 (0.10)	0.53 (0.30)	0.87 (0.50)
Weighted count	0.050	0.051	54	**0.0945**	0.03 (0.00)	0.11 (0.10)	0.58 (0.30)	0.91 (0.50)	**0.106**	0.03 (0.00)	0.12 (0.10)	0.58 (0.30)	0.90 (0.50)
2 major genes + 8 polygenes (n = 500)													
10 variants[e]													
10 IVs	0.100	0.118	7	**0.0815**	0.02 (0.00)	0.13 (0.10)	0.67 (0.30)	0.96 (0.50)	**0.091**	0.05 (0.03)	0.19 (0.12)	0.73 (0.33)	0.97 (0.52)
Allele count	0.061	0.063	34	**0.0815**	0.02 (0.00)	0.07 (0.10)	0.42 (0.30)	0.79 (0.50)	**0.091**	0.03 (0.00)	0.10 (0.10)	0.43 (0.30)	0.77 (0.50)
Weighted count	0.100	0.102	57	**0.0815**	0.02 (0.00)	0.10 (0.10)	0.60 (0.30)	0.94 (0.50)	**0.091**	0.03 (0.00)	0.12 (0.10)	0.60 (0.30)	0.93 (0.50)
2 major IVs + allele count	0.100	0.105	20	**0.0815**	0.02 (0.00)	0.11 (0.10)	0.62 (0.30)	0.94 (0.50)	**0.091**	0.03 (0.00)	0.13 (0.10)	0.63 (0.30)	0.94 (0.50)

[a]Values were obtained using 10 000 simulations. R^2 values correspond to the regression of the exposure (X) on the IV(s). Median effect estimates are shown in parentheses below the power estimates.
[b]Scenarios with weak instrument problems (i.e. low F values) are shown in grey, with darker shading corresponding to lower F values and more bias in the presence of X–Y confounding.
[c]The allelic β_{gx} is held constant for all analyses with a specific genetic model (in bold). See the methods section for an interpretation of the allelic β_{gx} in the 'continuum of effects' and the 'main effects/polygenes' models.
[d]Five variants with the frequencies are 0.1, 0.3, 0.5, 0.7 and 0.9.
[e]Ten variants total, two with each of the following frequencies: 0.1, 0.3, 0.5, 0.7 and 0.9.
[f]Twenty variants total, four with each of the following frequencies: 0.1, 0.3, 0.5, 0.7 and 0.9.

Table 4 Power[a] and effect estimate comparisons for Mendelian randomization studies, varying the strength of *X–Y* confounding and holding adjusted R^2 constant (in bold)

No. of variants IV(s)	Mean Adj. $R^{2,a}$	Mean $R^{2,a}$	F[b]	Power[a] under the 'equal-effects' model with $\beta_{xy}=0.3$ and $n=1000$ (median effect estimate[a])					
				No Confounding	Positive *X–Y* confounding		Negative *X–Y* confounding		
				$\beta_{ux}=0$ $\beta_{uy}=0$	$\beta_{ux}=0.25$ $\beta_{uy}=0.25$	$\beta_{ux}=0.50$ $\beta_{uy}=0.50$	$\beta_{ux}=0.25$ $\beta_{uy}=-0.25$	$\beta_{ux}=0.50$ $\beta_{uy}=-0.50$	$\beta_{ux}=0.75$ $\beta_{uy}=-0.75$
5 variants[c]									
5 IVs	**0.050**	0.055	12	0.61 (0.30)	0.63 (0.31)	0.64 (0.31)	0.60 (0.30)	0.58 (0.29)	0.55 (0.27)
Allele count	**0.050**	0.051	54	0.58 (0.30)	0.58 (0.30)	0.57 (0.30)	0.58 (0.30)	0.59 (0.30)	0.60 (0.30)
10 variants[d]									
10 IVs	**0.050**	0.059	6	0.65 (0.30)	0.67 (0.31)	0.70 (0.32)	0.62 (0.29)	0.57 (0.27)	0.52 (0.25)
Allele count	**0.050**	0.051	54	0.59 (0.30)	0.58 (0.30)	0.57 (0.30)	0.58 (0.30)	0.58 (0.30)	0.59 (0.30)
20 variants[e]									
20 IVs	**0.050**	0.069	4	0.70 (0.30)	0.75 (0.32)	0.82 (0.35)	0.68 (0.29)	0.56 (0.25)	0.41 (0.20)
Allele count	**0.050**	0.051	54	0.58 (0.30)	0.59 (0.30)	0.57 (0.30)	0.59 (0.30)	0.59 (0.30)	0.60 (0.30)

[a]Values are derived using 10 000 simulations. R^2 values correspond to the regression of the exposure (X) on the IV(s). Median effect estimates are shown in parentheses below the power estimates.
[b]Scenarios with weak instrument problems (i.e. low F values) are shown in grey, with darker shading corresponding to lower F values and more bias in the presence of *X–Y* confounding.
[c]Five variants with the frequencies 0.1, 0.3, 0.5, 0.7 and 0.9.
[d]Ten variants total, two with each of the following frequencies: 0.1, 0.3, 0.5, 0.7 and 0.9.
[e]Twenty variants total, four with each of the following frequencies: 0.1, 0.3, 0.5, 0.7 and 0.9.

effects of introducing various types of unmeasured *X–Y* confounding on the estimates from the 'equal-effects model'. Increasing the strength of the confounding effects increases the magnitude of the bias for the weak IV scenarios. The direction of the bias is always towards the confounded association estimate—away from the null under positive *X–Y* confounding and towards the null under negative *X–Y* confounding.

Discussion

In this simulation study of power and IV strength requirements for MR studies using 2SLS regression, we have evaluated power estimates for single-variant studies over a range of allele frequencies, effect sizes and sample sizes. Our results indicate that well-powered single-IV MR studies are not prone to weak IV problems, and therefore provide unbiased effect estimates in the presence of unmeasured confounding.

For MR studies using multiple variants, we have described the relationships among numerous key study variables, including the number of variants, variant effect sizes, exposure effect sizes, R^2, adjusted R^2, F and power. Our results suggest that power to detect a causal effect depends strongly on the R^2 value of the first-stage regression (not the adjusted R^2) and is not influenced by allele frequencies or the number of IVs included in a regression. However, for a fixed R^2, F decreases as the number of IVs increases, potentially creating weak IV problems (i.e. low F values). The weak IV problem has been extensively described in the econometrics literature.[27–30,33] In short, if an *X–Y* association is confounded, bias increases as F decreases. This bias is substantial for small F values (~4), but typically becomes negligible for F values >11 (Tables 1–4). This work shows that using each variant as an independent IV results in maximal power, but this strategy is often undesirable because of low F values. The weak IV problem can be overcome by combining the IVs, with modest reductions in power.

We have demonstrated several methods for combining IVs and explored their effects on power. The allele count IV is appropriate when each G has a similar effect, but suboptimal otherwise, as the effect sizes of the variants will be inherently mis-specified. If effect sizes are known, one can calculate a weighted allele count that has only slightly less power than using each variant as an independent IV and avoids weak IV problems. However, this method requires accurate effect estimates derived from previous research on independent samples. An alternative option is to use knowledge regarding 'major-gene' and 'polygenic' effects to create multiple IVs: one for each variant of large effect, and one representing the collective effects of the variants of small effect. For some circulating proteins, this model could potentially represent

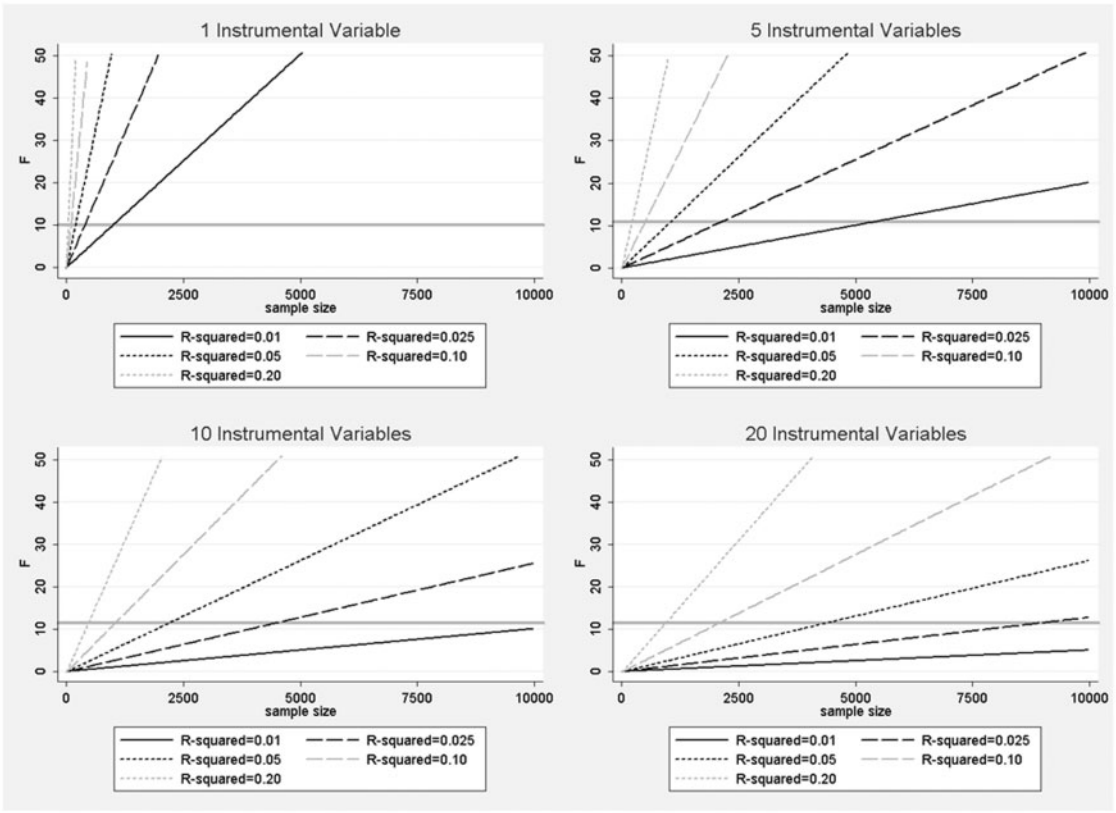

Figure 2 Relationship between F and R^2 for varying sample sizes and number of instrumental variables. The F thresholds from Stock *et al*.[29,30] are shown as horizontal lines

cis-effects (which tend to be strong) and *trans*-effects (which tend to be weak) on gene transcription.[34] This model requires no specific information on effect sizes, just knowledge of specific major genes and polygenes.

This work can be better understood in the context of Figure 2, which shows the relationship between R^2, F, sample size (n) and number of IVs (k) (from Equation 1), in the context of the F thresholds provided by Stock *et al*.[29] A set of IVs with an F below the threshold of ~11 is considered weak and will lead to relative bias >10% in the presence of X–Y confounding, with increasing bias as F decreases. This threshold can be increased or decreased to allow for lower or higher levels of relative bias, respectively.[29] Stock and Yogo[29] also provide thresholds for F which keep the actual size of the nominal 5% significance test <15% in the presence of 'maximal' confounding (i.e. X–Y correlation of 1). Unlike the 'relative-bias' threshold used in this work, the 'actual-size' threshold increases substantially as the number of IVs increases; however, because such 'maximal' confounding is unrealistic for epidemiological applications, we focus on the 'relative-bias' threshold in this work.

GWA studies typically identify several weak, independent genetic effects for biomarkers of interest. MR studies utilizing this information will require careful treatment of the weak IV problem, as information on genetic variants of weak effect will often need to be combined to produce strong IVs and an adequately powered MR analysis. The ideal method for translating genetic information into a reasonable number of IVs depends on several factors, including the total number of variants and their relative effect sizes. If using a small number of variants (<5), it may be possible to treat them as independent IVs, depending upon the value of F and the number of IVs.[30] This method will maximize power, while making no assumptions regarding the effect sizes of each G. If fewer IVs are needed, the model used to construct the IVs should consider the relative effect size of each G, while attempting to maximize the first-stage R^2 and F values and minimize the risk of effect-size mis-specification (resulting in loss of power). Models should be based on existing epidemiological and biological evidence.

This work is timely, as GWA studies have recently identified genetic determinants for a wide array of circulating biomarkers with known and suspected roles in a wide array of diseases (e.g. C-reactive protein,[8,9] urate,[35–39] lipids and triglycerides,[5–7] fasting plasma glucose[17–20] and B vitamins[10,11]). Clinically relevant physical measures (e.g. blood pressure[15,16] and body mass index[40–42]) and life-course traits

(e.g. age at menarche and menopause[43–48]) have also been shown to have genetic determinants. These variants are often not causal, but in linkage disequilibrium with a causal variant. The proportion of total phenotypic variance in these traits explained by the collective effects of known, common variants (R^2) is rarely >0.10 but often >0.01, while some studies fail to report the overall R^2 for genetic factors. Thus, we have conducted our power analyses using reasonable R^2 values, in light of current knowledge.

Our simulated data sets were created according to the three key criteria for MR studies: the IV is (i) associated with *X*, (ii) independent of *Y* given *X* and *X*–*Y* confounding factors and (iii) independent of factors that confound the *X*–*Y* association. Assumption 1 should be well established when conducting an MR study, but Assumptions 2 and 3 may require careful attention. Assumption 2 (known as the 'exclusion restriction' in the econometrics literature) requires that *G* affect *Y* only through *X*. This assumption could be violated in the presence of pleiotropy,[49] where a variant has independent effects on multiple traits (i.e. both *X* and *Y*). A violation could also occur if *G* is in linkage disequilibrium with a nearby variant that affects *Y*, thereby inducing a correlation between *G* and *Y*. It is not possible to test Assumption 2 by testing for *G*–*Y* independence while adjusting for *X*, because this adjustment will induce an association between *G* and *Y* in the presence of *X*–*Y* confounding (i.e. *X* is a 'collider'), even when Assumption 2 is valid; however, this test could be used to assess the absence of *X*–*Y* confounding. When using multiple IVs, it is possible to assess Assumption 2 by ensuring the estimates for each IV are similar (the 'over-identification test'[50]), but this itself requires additional assumptions. Assumption 3 could be violated as a consequence of population stratification, if the distributions of *G*, *X* and *Y* differ between sub-groups of the study sample;[51] however, these differences can be measured and adjusted for using genetic data.[3]

Our analysis makes several additional simplifying assumptions, namely, additive allelic effects for each variant, no gene–gene interactions and a linear effect for *X* (on *Y*). However, if these assumptions were known to be invalid, such knowledge could be incorporated into the model that relates *G* to the IV(s). For example, for variants with known non-additive effects on *X* (i.e. dominant or recessive), a binary *G* variable representing a dominant or recessive effect could be used as an independent IV or as an additional factor included in an allele count. IVs can also be constructed using data on haplotypes[25,26] rather than single variants. Gene–gene interactions could be modelled in a number of ways, including creating independent IVs representing the presence of effects due to interaction or including interaction terms when generating weighted allele counts.

In this analysis, we use the 2SLS regression on simulated cohort data sets to obtain an estimate of the causal effect of *X* on *Y*, when both are continuous variables. If there is heterogeneity in the effect of *X* on *Y* between individuals then although the 2SLS estimator does give an estimate of the causal effect, its precise interpretation is somewhat complicated.[52,53] The Wald estimator can be used for continuous *X* and *Y* variables and produces point estimates identical to the 2SLS method; however, this method cannot accommodate multiple IVs.[1] When the *X* variable is binary rather than continuous, it is possible to estimate so-called local average treatment or (with additional assumptions) the population average treatment effect.[54,55] When *Y* is binary, several other methods are available[56] and a description of some of these for epidemiologists is given in Rassen *et al*.[57] These methods include probit structural equation models, two-stage logistic models and generalized method of moments estimators. The biases that can arise with a continuous *X* and a binary *Y* are difficult to completely account for using standard statistical methods, although bias can be reduced using a residual-based adjustment in a two-stage logistic model and quantified under varying degrees of hypothesized *X*–*Y* confounding using sensitivity analyses.[58] The causal inference literature has developed methods to handle such settings,[59–61] but these impose an assumption of homogeneity or give only local causal effects. The power calculations derived here are valid only for cohort studies, as case–control studies require analysis techniques that integrate data on disease incidence into the analysis.[62] For example, if we let π denote the outcome prevalence in the population and we let p denote the ratio of cases to the sum of cases and controls in the study, then to conduct MR analyses with case–control data, we could use standard IV techniques but weight each case by π/p and each control subject by $(1-\pi)/(1-p)$ to obtain valid results.[63]

GWA studies are providing new tools for exploring causation using MR studies, but these tools must be applied carefully. The feasibility of MR studies will depend heavily upon the amount of variance in *X* that can be explained by known genetic factors and our understanding of those genetic effects. Given current knowledge of the genetic determinants of exposures of interest, samples sizes for MR studies will need to be quite large (>1000, sometimes >10 000). As more is learned regarding the genetics of health-related biomarkers, MR methods may become more efficient and broadly applicable.

Funding

National Institutes of Health (grant numbers UO1 CA122171, RO1 CA107431, RO1 CA102484, P30 CA014599 and P42 ES10349 to H.A.).

Conflict of interest: None declared.

KEY MESSAGES

- In Mendelian randomization studies, genetic factors that influence an exposure of interest can be used as 'instrumental variables' to assess the causality of an exposure–disease association using a two-stage least squares regression.
- Well-powered Mendelian randomization studies will require large ($n > 1000$), often very large ($n > 10\,000$), sample sizes.
- Using multiple genetic variants as instrumental variables can lead to 'weak instrument' scenarios, in which effect estimates may be substantially biased.
- Combining genetic factors into fewer instrumental variable results in modest power decreases, but reduces the 'weak instrument' bias.
- Ideal methods for combining genetic factors into fewer instrumental variables depend upon knowledge of the genetic architecture underlying the exposure.

References

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

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

[3] Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet* 2008;**17:**R143–50.

[4] Wehby GL, Ohsfeldt RL, Murray JC. 'Mendelian randomization' equals instrumental variable analysis with genetic instruments. *Stat Med* 2008;**27:**2745–9.

[5] Aulchenko YS, Ripatti S, Lindqvist I *et al*. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* 2009;**41:**47–55.

[6] Hegele RA. Plasma lipoproteins: genetic influences and clinical implications. *Nat Rev Genet* 2009;**10:**109–21.

[7] Kathiresan S, Willer CJ, Peloso GM *et al*. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* 2009;**41:**56–65.

[8] Reiner AP, Barber MJ, Guan Y *et al*. Polymorphisms of the HNF1A gene encoding hepatocyte nuclear factor-1 alpha are associated with C-reactive protein. *Am J Hum Genet* 2008;**82:**1193–201.

[9] Ridker PM, Pare G, Parker A *et al*. Loci related to metabolic-syndrome pathways including LEPR, HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's Genome Health Study. *Am J Hum Genet* 2008;**82:**1185–92.

[10] Hazra A, Kraft P, Selhub J *et al*. Common variants of FUT2 are associated with plasma vitamin B_{12} levels. *Nat Genet* 2008;**40:**1160–62.

[11] Tanaka T, Scheet P, Giusti B *et al*. Genome-wide association study of vitamin B_6, vitamin B_{12}, folate, and homocysteine blood concentrations. *Am J Hum Genet* 2009;**84:**477–82.

[12] Ferrucci L, Perry JR, Matteini A *et al*. Common variation in the beta-carotene 15,15′-monooxygenase 1 gene affects circulating levels of carotenoids: a genome-wide association study. *Am J Hum Genet* 2009;**84:**123–33.

[13] Meisinger C, Prokisch H, Gieger C *et al*. A genome-wide association study identifies three loci associated with mean platelet volume. *Am J Hum Genet* 2009;**84:**66–71.

[14] Soranzo N, Rendon A, Gieger C *et al*. A novel variant on chromosome 7q22.3 associated with mean platelet volume, counts, and function. *Blood* 2009;**113:**3831–37.

[15] Levy D, Ehret GB, Rice K *et al*. Genome-wide association study of blood pressure and hypertension. *Nat Genet* 2009;**41:**677–87.

[16] Newton-Cheh C, Johnson T, Gateva V *et al*. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* 2009;**41:**666–76.

[17] Bouatia-Naji N, Bonnefond A, Cavalcanti-Proenca C *et al*. A variant near MTNR1B is associated with increased fasting plasma glucose levels and type 2 diabetes risk. *Nat Genet* 2009;**41:**89–94.

[18] Bouatia-Naji N, Rocheleau G, Van Lommel L *et al*. A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels. *Science* 2008;**320:**1085–88.

[19] Chen WM, Erdos MR, Jackson AU *et al*. Variations in the G6PC2/ABCB11 genomic region are associated with fasting glucose levels. *J Clin Invest* 2008;**118:**2620–28.

[20] Prokopenko I, Langenberg C, Florez JC *et al*. Variants in MTNR1B influence fasting glucose levels. *Nat Genet* 2009;**41:**77–81.

[21] Cho YS, Go MJ, Kim YJ *et al*. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* 2009;**41:**527–34.

[22] Melzer D, Perry JR, Hernandez D *et al*. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet* 2008;**4:**e1000072.

[23] Sabatti C, Service SK, Hartikainen AL *et al*. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* 2009; **41:**35–46.

[24] Yuan X, Waterworth D, Perry JR *et al*. Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am J Hum Genet* 2008;**83:**520–28.

[25] Brunner EJ, Kivimaki M, Witte DR *et al*. Inflammation, insulin resistance, and diabetes–Mendelian randomization using CRP haplotypes points upstream. *PLoS Med* 2008;**5:**e155.

[26] Timpson NJ, Lawlor DA, Harbord RM *et al*. C-reactive protein and its role in metabolic syndrome: Mendelian randomisation study. *Lancet* 2005;**366:**1954–59.

[27] Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *J Am Statist Assoc* 1995;**90:**443–50.

[28] Nelson CR, Startz R. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *J Bus* 1990;**63:**S125–40.

[29] Stock JH, Yogo M. *Testing for Weak Instruments in Linear IV Regression*. Cambridge, MA: National Bureau of Economic Research, Inc, 2002.

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

[31] Staiger D, Stock JH. *Instrumental Variables Regression with Weak Instruments*. Cambridge, MA: National Bureau of Economic Research, Inc., 1994.

[32] Draper NR, Smith H. *Applied Regression Analysis*. New York: Wiley-Interscience, 1998.

[33] Donald SG, Newey WK. Choosing the number of instruments. *Econometrica* 2001;**69:**1161–91.

[34] Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet* 2008;**24:**408–15.

[35] Dehghan A, Kottgen A, Yang Q *et al*. Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study. *Lancet* 2008;**372:** 1953–61.

[36] Doring A, Gieger C, Mehta D *et al*. SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. *Nat Genet* 2008;**40:**430–36.

[37] Vitart V, Rudan I, Hayward C *et al*. SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat Genet* 2008;**40:** 437–42.

[38] Wallace C, Newhouse SJ, Braund P *et al*. Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am J Hum Genet* 2008;**82:**139–49.

[39] Li S, Sanna S, Maschio A *et al*. The GLUT9 gene is associated with serum uric acid levels in Sardinia and Chianti cohorts. *PLoS Genet* 2007;**3:**e194.

[40] Meyre D, Delplanque J, Chevre JC *et al*. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat Genet* 2009;**41:**157–59.

[41] Thorleifsson G, Walters GB, Gudbjartsson DF *et al*. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat Genet* 2009;**41:**18–24.

[42] Willer CJ, Speliotes EK, Loos RJ *et al*. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* 2009;**41:**25–34.

[43] He C, Kraft P, Chen C *et al*. Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. *Nat Genet* 2009;**41:**724–8.

[44] Liu YZ, Guo YF, Wang L *et al*. Genome-wide association analyses identify SPOCK as a key novel gene underlying age at menarche. *PLoS Genet* 2009;**5:**e1000420.

[45] Ong KK, Elks CE, Li S *et al*. Genetic variation in LIN28B is associated with the timing of puberty. *Nat Genet* 2009;**41:** 729–33.

[46] Perry JR, Stolk L, Franceschini N *et al*. Meta-analysis of genome-wide association data identifies two loci influencing age at menarche. *Nat Genet* 2009;**41:**648–50.

[47] Sulem P, Gudbjartsson DF, Rafnar T *et al*. Genome-wide association study identifies sequence variants on 6q21 associated with age at menarche. *Nat Genet* 2009;**41:** 634–8.

[48] Stolk L, Zhai G, van Meurs JB *et al*. Loci at chromosomes 13, 19 and 20 influence age at natural menopause. *Nat Genet* 2009;**41:**645–7.

[49] Thomas DC, Conti DV. Commentary: the concept of 'Mendelian Randomization'. *Int J Epidemiol* 2004;**33:** 21–25.

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

[51] Little J, Khoury MJ. Mendelian randomisation: a new spin or real progress? *Lancet* 2003;**362:**930–31.

[52] Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J Am Statist Assoc* 1995;**90:**431–42.

[53] Heckman JJ, Vytlacil EJ. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 2005;**73:**669–738.

[54] Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. *Econometrica* 1994;**62:** 467–75.

[55] Tan Z. Regression and weighting methods for causal inference using instrumental variables. *J Am Statist Assoc* 2006;**101:**1607–18.

[56] Greene WH. *Econometric Analysis*. 5th edn. Upper Saddle River, NJ: Prentice Hall, 2003.

[57] Rassen JA, Schneeweiss S, Glynn RJ, Mittleman MA, Brookhart MA. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. *Am J Epidemiol* 2009;**169:**273–84.

[58] Palmer TM, Thompson JR, Tobin MD, Sheehan NA, Burton PR. Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses. *Int J Epidemiol* 2008;**37:**1161–68.

[59] Robins JM, Rotnitzky A. Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika* 2004;**91:**763–83.

[60] van der Laan MJ, Hubbard A, Jewell NP. Estimation of treatment effects in randomized trials with noncompliance and a dichotomous outcome. *J R Stat Soc B* 2007; **69:**442–82.

[61] Vansteelandt S, Goetghebeur E. Causal inference with generalized structural mean models. *J R Stat Soc B* 2003; **65:**817–35.

[62] Shinohara RT, Frangakis CE, Platz E, Tsilidis K. *Estimating Effects by Combining Instrumental Variables with Case-control Designs: the Role of Principal Stratification*. Johns Hopkins University, Dept of Biostatistics Working Papers, 2008. (Working Paper 198): http://www.bepress.com/jhubiostat/paper198 (1 November 2009, date last accessed).

[63] Van der Laan MJ. Estimation based on case-control designs with known prevalence probability. *Int J Biostat* 2008;**4:**Article 17.

Published by Oxford University Press on behalf of the International Epidemiological Association

International Journal of Epidemiology 2011;**40**:752–754
doi:10.1093/ije/dyr017

Commentary: Can 'many weak' instruments ever be 'strong'?

Nuala A Sheehan[1]* and Vanessa Didelez[2]

[1]Department of Health Sciences, University of Leicester and [2]Department of Mathematics, University of Bristol, Bristol, UK

*Corresponding author. Department of Health Sciences, University of Leicester, Leicester LE1 7RH, UK.
E-mail: nas11@leicester.ac.uk

Accepted 12 January 2011

Investigations into the aetiology of common complex diseases based on observational data should make use of any opportunity to reduce bias due to unobserved confounding. In this context, it has become popular to exploit instrumental variable (IV) methods via Mendelian randomization but the key to success lies in finding suitable genetic instruments. Genome-wide association studies are increasingly yielding large numbers of biomarkers and the understanding of the functionality of these variants is continually improving. However, genetic instruments typically explain only a small proportion of the overall variation in a given exposure and are therefore loosely regarded as 'weak' instruments. Combining several instruments intuitively seems like a plausible approach to improving overall instrument strength. Given the likely availability of ever more genetic instruments in the foreseeable future, an investigation into the power and instrument strength requirements of Mendelian randomization analyses with multiple instruments, as proposed by Pierce *et al.*[1], is both relevant and timely.

In a Mendelian randomization study, the typical target of inference is the effect of an exposure X on a disease outcome Y in the presence of unmeasured confounding factors, U, using one or a combination of several genetic variant(s), G, as an IV. It is often assumed that X and Y are continuous and that all relationships are linear with no interactions, as in Pierce *et al.*[1] (Note that the linear models in Equations (4) and (6) in Pierce *et al.*[1] are not correct, as stated: g_i should be replaced by x_i, as implied in the surrounding text, and not as written.) The causal parameter of interest is the effect that manipulating X, to change it by one unit, has on Y—the so-called average causal effect (ACE)—and happens to coincide with the coefficient of X in the regression of Y on X and U under the above model assumptions. The two-stage least squares (2SLS) IV estimator is commonly used in this context, as it is asymptotically unbiased for the ACE under these model assumptions, but, crucially, this is not necessarily the case in finite samples.

In the work of Pierce *et al.*,[1] simulation studies were carried out where different strategies for combining multiple genetic variants into instruments were considered, and their impact on power to detect a causal effect of X on Y, based on 2SLS, assessed. The authors focus on the case of 'weak' instruments because of their relevance to Mendelian randomization applications. The problem with weak instruments is 2-fold: not only is there limited power to detect any effect at all but there can also be 'weak instrument bias'. Bound *et al.*[2] noted that any correlation between G and U, however small, can lead to large inconsistencies in the IV estimate if the true relationship between G and X is weak and the sample size insufficiently large to compensate. Even when G is a legitimate instrument and no such correlation with U exists on a population level, sampling variation can induce an empirical correlation and hence bias in the IV estimate. The bias is in the direction of