Int. J. Epidemiol. Advance Access published February 9, 2011

Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2011;1–3
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 the bias of the ordinary least squares (OLS) estimate obtained from a regression of Y on X, which is confounded by U. The more parameters in the first-stage regression (of X on G), the greater the opportunity for any accidental correlation between G and U to affect the 2SLS estimator—we could call this 'over-fitting' of this first-stage regression, and it can easily arise when using multiple instruments. The effects of over-fitting the first-stage regression can be seen clearly in the reported results of Table 2 of Pierce *et al.*,[1] e.g. where the true allele effect sizes were all equal but the fitted models allowed them to be distinct. We can also see that when there is no confounding between X and Y, over-fitting is actually desirable as it increases the power, but in this case an IV analysis is not actually required. In the more relevant case, when there is unobserved confounding, the IV analysis is prone to bias, which is larger when there are many free parameters (i.e. more instruments) in the first-stage regression.

These comments highlight the fact that our primary concern when dealing with 'weak' instruments should be the 'bias' of the IV estimator and not the power of an IV-based test, and indeed this turns out to be of central importance when interpreting the results in Pierce *et al.*[1] If the IV estimator is biased, the actual level of a corresponding statistical test will be larger than its nominal level and so any comparison of power between situations with different amounts of bias is uninformative. This must be kept in mind when drawing any conclusions about power from the observations in Pierce *et al.*[1]

The above is not a formal definition of weak IVs, nor do the authors of Pierce *et al.*[1] provide us with such a definition. It is plausible to relate instrument strength somehow to the first-stage regression R^2 and F statistics,[2,3] as R^2 is an empirical measure for the variation explained in the first-stage regression based on a given data set (adjusted R^2 uses the correct degrees of freedom), and F is a test statistic for the null hypothesis that the IVs G do not predict the exposure X at all. Although we certainly agree with Bound *et al.*[2] that these should always be reported in any given analysis, we want to caution against using these quantities to define instrument strength as they are affected by sampling variation and do not in themselves represent a population quantity. Pierce *et al.*[1] implicitly use this definition in terms of sample quantities, as they 'keep R^2 fixed' and vary the sample size and number of parameters in the first-stage regression to obtain different F-values. This involves adapting the allelic coefficients in every one of the 10 000 repetitions for their simulations, which is slightly unusual as we tend to regard model parameters as fixed, whereas R^2 is a sampling statistic and not a parameter. (Hence, the reported allelic coefficients in Tables 2 and 3 are presumably averages over the 10 000 data sets.)

In contrast, Staiger and Yogo[4] propose two different ways of defining weak instruments in terms of their population behaviour. Accordingly, IVs are weak (i) if the relative bias of the IV 2SLS estimator in relation to the ordinary observational estimator exceeds a certain threshold e.g. 10% or (ii) if the actual level of a statistical test, using the IVs, with nominal α-level exceeds a certain threshold, e.g. 15% when $\alpha = 5\%$. These definitions are both expressed in terms of how wrong, 'on average', inference based on the given IVs is. This depends obviously on the amount of confounding, and also on the sample size, the (joint) distribution of the instruments, the true coefficients in the first-stage regression and the true residual variance of the first-stage regression. For a univariate exposure, the first-stage F-statistic then happens to provide a significance test for the null hypothesis that the IV is 'too weak', but the critical value depends on which of the above two definitions of 'weak IV' we use. With the first definition, i.e. that the induced worst case relative bias is more than 10% compared with OLS, the check $F \geqslant 10$ for up to four IVs in the first-stage regression (and $F > 11$ for more IVs) rejects the 'weak IV' hypothesis roughly at a 5% level, giving rise to the much-quoted rule of thumb. However, for the alternative definition of an IV being 'too weak' in the sense that an IV-based test of the effect of X on Y with nominal $\alpha = 5\%$ has an actual level of more than 15%, then the check of $F \geqslant 10$ is only valid for a single IV and, crucially, the critical F-value increases dramatically with the number of IVs in the first-stage regression and is about 26 when there are 15 IVs.[3] This latter phenomenon seems particularly relevant to the power considerations in Pierce *et al.*[1] Like every statistical test, these tests for weak IVs can lead to erroneous conclusions, so any prior knowledge or external evidence that can confirm the strength of the IVs is always useful. We hence emphasize that the F-statistic is only a 'test' for weak instruments and that a small F-value is hence not a 'definition' of a weak instrument as is often implied. In particular, we would suspect that any data-driven choice of the first-stage regression model if systematically based on R^2 and F, which tend to over-estimate the true explained variation especially when over-fitted, will re-introduce bias into the analysis.

The most interesting aspect of the investigations of Pierce *et al.*[1] is that it appears that the weak instrument bias can be alleviated by combining instruments and hence reducing the number of parameters to be estimated in the first-stage regression, without much loss in power. This is particularly suited to Mendelian randomization applications, where prior knowledge of the genetic mechanisms determining how the instruments affect the exposure will often be available. Methods for combining IVs have a natural interpretation in these settings. For example, a total 'X-increaser' score obtained from adding up the values of each of the separate IVs is simply an allele score, which assumes that the genetic variants have equal and additive effects with no interactions on the exposure. If the equal effects model is implausible, a weighted score can be used, but this now requires either prior knowledge about appropriate weights or estimates obtained from a separate data set. Again, we emphasize that the weights should not, of course, be estimated from the same data on which the IV analysis is carried out as this would be entirely equivalent to fitting a first-stage regression model with many parameters and would re-introduce weak instrument bias via over-fitting. Another option is to incorporate prior information about major gene vs polygenic effects by fitting individual coefficients for the major genes and combining the polygenes into an allele score assuming equal weights, as is typically done in genetic analyses. If appropriate, this has the added attraction of not requiring any specification of the true weighting factors: it only requires prior knowledge about which genes have major effects.

The flip side of the above is that in combining the genetic instruments in a chosen way, we might make a mistake, e.g. we might choose the wrong major genes, or we might assume the weights are equal when they are not. In principle, such mis-specification of the first-stage regression model might itself induce bias regardless of the strength of the IV. From those four such scenarios that were investigated by Pierce *et al.*[1] (e.g Table 3 using 'allele counts' models for 'continuum of effects' and '2 major genes +8 polygenes' models), it seems that if this model is not grossly wrong, the bias is still negligible while maintaining competitive levels of power. However, it would be interesting to see whether a seriously incorrect first-stage regression model could actually induce more bias than it aims to prevent. Until further analyses targeting this specific question are performed, we feel it is premature to draw general conclusions.

To summarize, Mendelian randomization studies provide a unique opportunity to deal with suspected unobserved confounding, as we often have good reasons to believe in the validity of proposed genetic instruments. However, we will often have to deal with the case of many weak instruments, potentially running the risk of biased estimates. The findings of Pierce *et al.*[1] suggest that, provided we have good biological evidence or other prior knowledge informing a more parsimonious modelling of the first-stage regression, we should be able to reduce the risk of weak instrument bias and still retain reasonable power. In our view, strategies for choosing the first-stage regression model still require more investigation, especially with regard to the questions of: (i) to what extent these can be data driven without re-introducing bias in the absence of reliable prior knowledge and (ii) how problems induced by mis-specifications of the first-stage regression model are balanced out by the resulting model simplifications. Also, the results so far are all specific to the linear/no-interactions model. An investigation into the usefulness of the proposed first-stage strategies to combine multiple genetic instruments for binary outcome models is arguably more relevant to Mendelian randomization studies. Finally, we must not forget that the main source of bias is due to violations of model assumptions and these should therefore be checked wherever possible.[5]

Funding

Medical Research Council (Project Grant G0601625); Leverhulme Trust (Research Fellowship RF/9/RFG/2009/0062).

Conflict of interest: None declared.

References

[1] Pierce BL, Ahsan H, VanderWeele TJ. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int J Epidemiol* 2010. doi:10.1093/ije/dyq151 [Epub 2 September 2010].

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

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

[4] Staiger JH, Yogo M. Testing for weak instruments in linear IV regressions. Chapter 5. In: Andrews DWK (ed.). *Identification and Inference for Econometric Models*. New York: Cambridge University Press, 2005, pp. 80–108.

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