Statistics in Medicine

Research Article

Received 14 April 2014, Accepted 20 October 2014 Published online 10 November 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6358

The many weak instruments problem and Mendelian randomization

Neil M. Davies,[a,b*†] Stephanie von Hinke Kessler Scholder,[c] Helmut Farbmacher,[d] Stephen Burgess,[e] Frank Windmeijer[a,c] and George Davey Smith[a,b]

Instrumental variable estimates of causal effects can be biased when using many instruments that are only weakly associated with the exposure. We describe several techniques to reduce this bias and estimate corrected standard errors. We present our findings using a simulation study and an empirical application. For the latter, we estimate the effect of height on lung function, using genetic variants as instruments for height. Our simulation study demonstrates that, using many weak individual variants, two-stage least squares (2SLS) is biased, whereas the limited information maximum likelihood (LIML) and the continuously updating estimator (CUE) are unbiased and have accurate rejection frequencies when standard errors are corrected for the presence of many weak instruments. Our illustrative empirical example uses data on 3631 children from England. We used 180 genetic variants as instruments and compared conventional ordinary least squares estimates with results for the 2SLS, LIML, and CUE instrumental variable estimators using the individual height variants. We further compare these with instrumental variable estimates using an unweighted or weighted allele score as single instruments. In conclusion, the allele scores and CUE gave consistent estimates of the causal effect. In our empirical example, estimates using the allele score were more efficient. CUE with corrected standard errors, however, provides a useful additional statistical tool in applications with many weak instruments. The CUE may be preferred over an allele score if the population weights for the allele score are unknown or when the causal effects of multiple risk factors are estimated jointly. © 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

Keywords: Mendelian randomization; many weak instruments; continuously updating estimator; allele scores; height; ALSPAC

1. Introduction

Mendelian randomization uses genetic variants as instrumental variables to investigate the effects of modifiable risk factors for disease such as weight, blood pressure, cholesterol, alcohol, and tobacco consumption on different outcomes of interest [1–9]. An instrument is a variable external to the model of interest that is robustly associated with the modifiable risk factor, but is not associated with the outcome variable, other than through its effect on the risk factor. One challenge in using genetic instrumental variables is that many genetic variants are only modestly associated with the risk factor of interest, which limits the power and precision of a study. One approach commonly used in econometric studies to increase the power and precision is to include multiple instruments for the exposure of interest. However, studies

[a]*Medical Research Council Integrative Epidemiology Unit, University of Bristol, Barley House, Oakfield Grove, Bristol, BS8 2BN, U.K.*
[b]*School of Social and Community Medicine, University of Bristol, Barley House, Oakfield Grove, Bristol, BS8 2BN, U.K.*
[c]*Department of Economics, University of Bristol, 8 Woodland Road, Bristol, BS8 1TN, U.K.*
[d]*Munich Center for the Economics of Aging, Max Planck Institute for Social Law and Social Policy, Amalienstr 33, 80799 Munich, Germany*
[e]*Department of Public Health and Primary Care, School of Clinical Medicine, University of Cambridge, Cambridge CB1 8RN, U.K.*
[*]*Correspondence to: Neil Davies, School of Social and Community Medicine, University of Bristol, Barley House, Oakfield Grove, Bristol, BS8 2BN, U.K.*
[†]*E-mail: neil.davies@bristol.ac.uk*

using multiple instruments that are only modestly associated with the risk factor of interest can suffer from many weak instruments bias [10].

A weak instrumental variable is one that explains only a small proportion of the variation in the exposure. Studies using a single weak instrument will have low power to reject the null hypothesis. In addition, any minor violation in the 'exclusion restriction' (i.e. the assumption that the instrument is unrelated to any confounders and does not directly affect the outcome of interest) can cause large biases [11].

Studies using many instruments that are each only weakly associated with the exposure can suffer from a further distinct bias: a many weak instruments bias. This implies that, even if a set of multiple instruments is valid, (i.e. they are not associated with confounding factors, have no direct effect on the outcome, and are at least weakly associated with the exposure) the two-stage least squares estimator can still be biased towards the conventional regression estimate [10, 12]. This is a concern in applications using multiple genetic variants as instruments, because each variant is typically weakly associated with the exposure of interest [13–16]. We demonstrate that, in agreement with previous theoretical and simulation results [17–19], weak instrument robust methods such as combining the variants into a single allele score, using the limited information maximum likelihood (LIML) estimator, or the continuously updating estimator (CUE) do not suffer from the many weak instruments bias. Whilst there are a number of techniques for using multiple variants, such as meta-analysis and structural mean models [20, 21], the aim of this paper is to illustrate the use of CUE and LIML compared with 2SLS and allele scores. In addition, we show that the usual asymptotic standard errors for LIML and CUE are downward biased with many weak instruments. We therefore calculate corrected standard errors [18, 22] and show that these can be used for correct inference. Finally, we illustrate these issues in the context of an empirical application, evaluating the effect of height on lung function.

The original example of the many weak instruments problem occurred in the work of Angrist and Krueger [23] who investigated the effects of length of schooling on wages and used quarter of birth as an instrument for time spent in school. In their basic specification, they used three variables indicating quarter of birth as instruments for length of schooling. However, these instruments explained relatively little of the variation in their exposure (partial $R^2 = 0.00012$). In order to increase the precision of their estimates, the authors included the interaction of length of schooling and year of birth, which resulted in 180 instruments. This increased the partial R^2 to 0.00043 and reduced the standard errors on the effect of schooling by over half. However, Bound *et al.* [10] and Hansen *et al.* [17] showed that, whilst the results were more precise, the point estimates from the instrumental variable regression were biased towards the ordinary least squares (OLS) estimates.

2. Methodology

In this section, we describe the different instrumental variables estimators applied here using the generalized method of moments (GMM) framework. OLS, two-stage least squares, and the CUE are all special cases of GMM. Consider a structural model with one continuous outcome, l continuous exposures, and k instrumental variables

$$y_i = x_i'\beta + u_i, \tag{1}$$

for individual observations $i = 1, \ldots, n$, with n as the sample size, and where outcome y_i is a scalar, x_i is a $l \times 1$ vector of the exposures, and u_i is a scalar indicating the error term. β is a $l \times 1$ vector of parameters. Unobserved confounding implies that u_i is correlated with the exposure x_i and hence the OLS estimator is inconsistent. An estimator is inconsistent if it is a biased estimator of the causal effect even in large samples, whereas an estimator is consistent if it is an unbiased estimator of the causal effect in large samples: in other words, it is asymptotically unbiased. The $J \times 1$ vector of instruments z_i, with $J \geqslant l$, satisfy the population moment conditions

$$E\left[g_i(\beta)\right] = E\left[z_i\left(y_i - x_i'\beta\right)\right] = 0. \tag{2}$$

This results in one equation (moment condition) for each instrument, and the objective of the estimator is to find the value of β that minimizes a criterion function $Q(\beta)$. The general form of a GMM criterion function is given by

$$\hat{\beta} = \arg\min_{\beta} Q(\beta); Q(\beta) = \bar{g}(\beta)' W \bar{g}(\beta),$$

where $\bar{g}(\beta) = n^{-1}\sum_{i=1}^{n} g_i(\beta) = n^{-1}\sum_{i=1}^{n} z_i\left(y_i - x_i'\beta\right)$ is the sample analog of the population moment conditions (2) and where W is a weight matrix giving potentially differential weight to each moment condition, which affects the efficiency of the GMM estimator. This weight matrix affects the properties of the estimator only if the model is overidentified; that is, when there are more instruments than exposures; $J > l$. Note that the OLS estimator is equivalent to a GMM estimator with z_i set equal to x_i

2.1. *Two-stage least squares*

A widely used weighting matrix is $W_z = \left(n^{-1}\sum_{i=1}^{n} z_i z_i'\right)^{-1}$, which, combined with the moment conditions earlier (2), results in the two-stage least squares estimator

$$\hat{\beta}_{2SLS} = \arg\min_{\beta}\ \dot{Q}(\beta); \dot{Q}(\beta) = \bar{g}(\beta)' W_z \bar{g}(\beta). \tag{3}$$

Setting the first derivative of $\dot{Q}(\beta)$ to zero and rearranging terms give the well-known formula for the two-stage least squares estimator

$$\hat{\beta}_{2SLS} = (D' W_z D)^{-1} D W_z s, \tag{4}$$

where $D = n^{-1}\sum_{i=1}^{n} z_i x_i'$ and $s = n^{-1}\sum_{i=1}^{n} z_i y_i$. The two-stage least squares estimator is efficient, given moment conditions (2), when the variance of u_i is homoskedastic; that is, $E\left[u_i | z_i\right] = \sigma_u^2$. When the conditional variance of u_i is heteroskedastic, that is, a function of the instruments $E\left[u_i | z_i\right] = \sigma^2\left(z_i\right)$, two-stage least squares is no longer efficient, but the two-step GMM estimator, as described later, is.

2.2. *Two-step generalized method of moments (GMM)*

Hansen described the two-step GMM estimator [24]. The two-step estimator uses a preliminary consistent estimate of β, denoted $\dot{\beta}$ (e.g. $\dot{\beta} = \hat{\beta}_{2SLS}$), to compute the efficient weight matrix

$$\tilde{W}\left(\dot{\beta}\right) = \left(n^{-1}\sum\nolimits_{i=1}^{n} g_i\left(\dot{\beta}\right) g_i\left(\dot{\beta}\right)'\right)^{-1},$$

where $g_i\left(\dot{\beta}\right) = z_i\left(y_i - x_i'\dot{\beta}\right)$. In the second step, the weight matrix is substituted into the objective function to obtain the two-step GMM estimator $\ddot{\beta}$,

$$\ddot{\beta} = \arg\min_{\beta} \ddot{Q}(\beta); \ddot{Q}(\beta) = \bar{g}(\beta)' \tilde{W}\left(\dot{\beta}\right) \bar{g}(\beta). \tag{5}$$

2.3. *The continuously updating estimator (CUE)*

The CUE proposed by Hansen *et al.* simultaneously minimizes over β in the moment conditions and the weighting matrix [25]. The CUE $\hat{\beta}_C$ is defined as

$$\hat{\beta}_C = \arg\min_{\beta} \hat{Q}(\beta); \hat{Q}(\beta) = \bar{g}(\beta)' \tilde{W}(\beta) \bar{g}(\beta), \tag{6}$$

where $\tilde{W}(\beta) = \left(n^{-1}\sum_{i=1}^{n} g_i(\beta) g_i(\beta)'\right)^{-1}$. Note that the weight matrix now depends on β. Imposing conditional homoskedasticity, the CUE objective function simplifies to

$$\hat{Q}(\beta) = \frac{\bar{g}(\beta)' W_z \bar{g}(\beta)}{(\sigma(\beta))^2}, \tag{7}$$

where $(\sigma(\beta))^2 = n^{-1}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$. The estimator that minimizes this function is the same as the LIML estimator, and therefore, the CUE is identical to LIML when imposing conditional homoskedasticity. CUE has the same limiting distribution as the two-step GMM estimator under standard asymptotics with a fixed number of instruments and is hence efficient under conditional heteroskedasticity.

2.4. *The bias with many weak instruments*

Theoretical and simulation studies such as Newey and Windmeijer and Burgess and Thompson have demonstrated that LIML and CUE are less biased than two-stage least squares (2SLS) when using many weak instruments [17–19]. Under conventional, strong instrument, asymptotics, the bias of the 2SLS estimator when there are more than three instruments can be approximated by [10, 26–28]

$$E\left[\hat{\beta}_{2SLS} - \beta\right] \approx \frac{\sigma_{uv}}{\sigma_v^2}\left(\frac{J-2}{\mu^2}\right), \tag{8}$$

where the concentration parameter, $\mu^2 = \frac{\pi' Z' Z \pi}{\sigma_v^2}$, is the amount of the variation in the exposure that is jointly explained by the instruments, σ_v^2 is the variance of the error term in the first stage equation for x, $x_i = z_i'\pi + v_i$, and σ_{uv} is the covariance of the error terms u_i and v_i. In the sample, the concentration parameter evaluated at the OLS estimates for π and σ_v^2 and divided by J is equal to the F-statistic for testing the null hypothesis that $\pi = 0$. Thus, the 2SLS bias shown in Equation (8) is inversely proportional to the concentration parameter and proportional to the number of instruments and covariance term. If a researcher adds an extra instrument, which increases J, the bias will only fall if it explains a sufficient additional amount of the variance in the exposure (i.e. it decreases the $\left(\frac{J-2}{\mu^2}\right)$ term). If the additional instrument only explains a small proportion of the variation in the exposure, the 2SLS bias will increase. This means that, when using 2SLS, researchers need to consider not only the strength of the association between each instrument and the exposure but also the number of instruments included.

The analogous bias approximation for LIML is given by [27, 28]

$$E\left[\hat{\beta}_{LIML} - \beta\right] \approx -\frac{\sigma_{uv}}{\sigma_v^2}\frac{1}{\mu^2}. \tag{9}$$

The bias of LIML is inversely proportional to the concentration parameter and negatively proportional to the covariance term but does not depend on the number of instruments. This latter feature is also valid for the CUE [29, 30]. In other words, additional instruments can be added when using LIML/CUE as long as they increase the concentration parameter.

Limited information maximum likelihood and CUE are consistent and asymptotically normally distributed under many weak instrument asymptotics, where the number of instruments is allowed to grow with the number of observations [18, 31]. 2SLS and two-step GMM are inconsistent estimators of the causal effect under these asymptotics. For both LIML and CUE, the usual asymptotic standard errors are too small when there are many weak instruments. Bekker [31] and Newey and Windmeijer [18] provide corrected standard errors for the LIML estimator and CUE respectively resulting in reliable inference when there are many weak instruments. We will base our inference later on these corrected standard errors. See Baum *et al.*, for a more detailed review [32]. Hausman *et al.* [33] further show that LIML can be biased with many weak instruments when the errors u_i are conditionally heteroskedastic. As CUE remains efficient under these conditions, we argue that CUE is preferable to LIML for most applications.

3. Simulation study

We use a simulation study to illustrate the many weak instrument bias and the performance of the corrected standard errors for the LIML and CUE in Mendelian randomization. The data generating process has been described in detail previously [19]. In short, there is a single continuous exposure x, a continuous outcome y, and J instruments $\{z_j\} = \{z_1, \ldots, z_J\}$ that are discrete with values $k = \{0, 1, 2\}$, mimicking the allele frequencies of genetic markers. We simulated data for three examples using different numbers of genetic variants as instrumental variables. Specifically, we set $J = 9$, 25, and 100 variants.

The data are generated from

$$x_i = \pi_z \sum_{j=1}^{J} z_{ij} + v_i, \tag{10}$$

$$y_i = \beta x_i + u_i, \tag{11}$$

$$v_i = w_i + e_{xi},$$

$$u_i = w_i + e_{yi},$$

$$w_i, e_{xi}, e_{yi} \sim N(0, 1). \tag{12}$$

We set $\pi_z = 0.1$ for 9 variants, $\pi_z = 0.06$ for 25 variants, and $\pi_z = 0.03$ for 100 variants with a minor allele frequency of 0.3 for each variant. The exposure x_i has no causal effect on y_i (i.e. $\beta = 0$). There is, however, a positive correlation between u_i and v_i because of the presence of the unobserved confounder w_i in the models for both the risk factor and the outcome. We used this data generating process to simulate 10,000 datasets with 3000 observations each and report median bias and rejection proportions of Wald tests for the null hypothesis H_0: $\beta = 0$. To avoid local optima, we estimated the CUE by taking the minimum of $\hat{Q}(\beta)$ obtained from five different starting values. Specifically, we used $\{-2, -1, 0, 1, 2\}$ times the two-step GMM estimates as starting values.

Using the bias shown in Equations (8) and (9), we can estimate the theoretical bias.[‡] As the instruments are independently and identically distributed with variance

$$\sigma_z^2 = \sum_{k=0}^{2} \Pr(z = k)k^2 - \left(\sum_{k=0}^{2} \Pr(z = k)k \right)^2 = 0.42, \tag{13}$$

and $\sigma_v^2 = 2$, the concentration parameter simplifies to

$$\mu^2 = \frac{nJ\pi_z^2\sigma_z^2}{\sigma_v^2} = 0.21nJ\pi_z^2. \tag{14}$$

According to Stock *et al.* [34], the expectation of the F-statistic minus one approximately equals the concentration parameter divided by the number of instruments or $E(F) \cong \mu^2/J + 1$. Thus, solely based on the parameters of the data generating process, we know the concentration parameter, the approximate expectation of the F-statistic, and the approximate bias of 2SLS and LIML estimators. We report these theoretical results together with the corresponding results from the simulation study in Table I. Furthermore, we report the rejection frequencies of the Wald tests for 2SLS, LIML, CUE, and the allele score instrumental variable estimators, calculated on the basis of the asymptotic and the corrected standard errors. The 2SLS estimator, LIML estimator, and CUE use the full set of instruments, not imposing any restrictions on the parameter vector π. The allele score is constructed as the unweighted sum of the instruments, which in this case is the sum of the risk alleles (10).

Table I presents medians and interquartile ranges (IQR) of the estimators. As the LIML estimator and CUE have occasionally very large outliers, this affects the mean and variance of these estimators, but not the median and IQR. The results show that, as expected, 2SLS suffers from the many weak instruments bias, whereas LIML and CUE are approximately unbiased in all simulations. The allele score instrumental variable estimator is also unbiased. Keeping the number of instruments fixed, all estimators have the same limiting distribution when the sample size increases. However, in the generated samples, only the allele score instrumental variable estimator displays a finite sample behavior close to the asymptotic distribution. Both LIML and CUE show larger finite sample dispersions than the allele score instrumental variable estimator as the number of instruments increases. The usual asymptotic standard error for LIML underestimates its true variability in this design, in line with previous work [18, 19]. We also find this for the uncorrected asymptotic standard errors of the CUE. In contrast, the Wald tests based on the Bekker standard errors for LIML and the Newey–Windmeijer standard errors for the CUE show rejection frequencies close to the nominal size of 0.05 in all three simulations. Thus, the corrected standard errors make valid inference possible in situations with many weak instruments.

[‡] *We calculate the bias for the 2SLS estimator from the improved formula as given by Bun and Windmeijer* [26]. *The results are very similar to those of (8).*

Table I. Simulation results comparing the properties of two-stage least squares, LIML, CUE, and allele score instrumental variable estimators when using 9, 25, and 100 variants as instrumental variables.

		Theoretical results	Empirical results		
k	Estimator	Bias	Median	IQR	RF
9	2SLS	0.057	0.065	0.158	0.101
	LIML	−0.009	0.002	0.186	0.057
	Corrected LIML				0.047
	CUE		0.002	0.187	0.079
	Corrected CUE				0.045
	Allele score IV		0.001	0.178	0.047
25	2SLS	0.147	0.150	0.139	0.321
	LIML	−0.009	0.001	0.204	0.079
	Corrected LIML				0.049
	CUE		0.002	0.207	0.159
	Corrected CUE				0.046
	Allele score IV		0.000	0.181	0.044
100	2SLS	0.318	0.318	0.095	0.988
	LIML	−0.009	0.002	0.277	0.173
	Corrected LIML				0.046
	CUE		0.001	0.297	0.481
	Corrected CUE				0.045
	Allele score IV		0.000	0.176	0.042

Note that $\mu^2 = 56.70$ in all designs. $E[F] \approx 7.30, 3.27, 1.57$ for $k = 9, 25, 100$, respectively. The means of the empirical F-statistics are 7.40, 3.32, and 1.58. The rejection frequency is the proportion of replications for which the true parameter ($\beta = 0$) lies outside the 95% confidence intervals. Median, median estimate, $\beta = 0$; $n = 3000$; 10,000 replications. 2SLS rejection frequencies assume a homoskedastic error term. Rejection frequencies for $H_0 : \beta = 0$ using Wald tests.
LIML, limited information maximum likelihood; CUE, continuously updating estimator; 2SLS, two-stage least square; IQR, interquartile range; RF, rejection frequency.

4. Empirical example: height and lung function

We illustrate the many weak instrument bias and its different estimators in an empirical application, investigating the relationship between height and lung function. We chose this example, because we know that there is a mechanical relationship between the two variables, with lung volume increasing with height. Therefore, we can use this example to illustrate how height variants can estimate the causal effect of height independent of a confounding factor – gender. Gender confounds the raw association of lung function and height because boys are both taller and have higher lung function. We used data from Avon Longitudinal Study of Parents and Children (ALSPAC), a cohort study of 15,247 pregnancies delivered between 1 April 1991 and 31 December 1992 in the South West of England, which has been described in detail elsewhere [35, 36]. Our sample includes all 3631 individuals with genome-wide data, information on height and lung spirometry (forced vital capacity) when aged between 14 and 17 years. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. The study website contains details of all the data that are available through a fully searchable data dictionary [37].

4.1. Outcome – lung function

Lung function was measured using forced vital capacity derived from a lung spirometry test. This measure has been described in detail elsewhere [38, 39]. We standardized this measure to mean zero and standard deviation one.

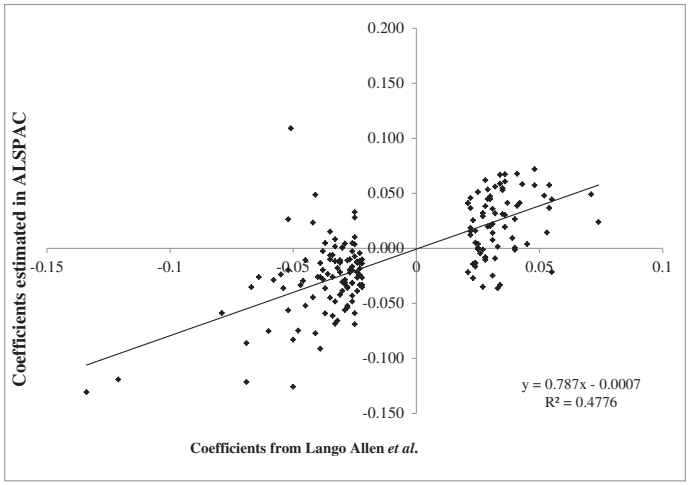

Figure 1. Plot of the association of 180 genetic variants and height in ALSPAC and Lango *et al.* [42].

4.2. Exposure – height

Individuals' height was measured in millimeters using the Harpenden stadiometer, when the individuals were aged 15 and a half [40]. We standardized height to mean zero and standard deviation one.

4.3. Covariates

We derived the following covariates: age at measurement (in months); birth weight; number of younger and older siblings; log weekly household income adjusted for inflation, household size, and composition; parents' education; and social class and employment status. Individuals whose fathers were in the military were coded as social class 3. The derivation of these covariates has been described elsewhere [41].

4.4. Instruments

In a genome-wide study of 183,727 people, Lango *et al.* reported 180 independent genetic variants associated with height [42]. Because the ALSPAC cohort was not used in Lango *et al.,* our results will not suffer from bias due to selecting variants based on their observed association within our data [19]. We use the same 180 variants listed in Table S1, which shows their effect on height as reported by Lango *et al.*, as well as in our ALSPAC sample. In addition, Figure 1 plots the association of the coefficients estimated in our data and the coefficients reported by Lango *et al.* The protocols for cleaning the genome-wide data have been described in detail elsewhere [43, 44].

We used the 180 height variants as 180 individual instruments. Each variant was coded 0, 1, or 2 depending on the combination of height-increasing alleles the individual had. To further explore alternative definitions, we also consider two additional versions of the instrument. First, a simple count of the number of height-increasing alleles

$$z_{1i} = \sum_{j=1}^{180} G_{ij}\,, \tag{15}$$

where G_{ij} indicates the jth genetic variant. Second, a weighted allele score, where the weights, $\hat{\pi}_j$, denote the strength of the association of the alleles of variant j and height, as reported by Lango *et al.*

$$z_{2i} = \sum_{j=1}^{180} \hat{\pi}_j G_{ij}\,. \tag{16}$$

To simplify the interpretation of the results, we normalized the unweighted and weighted allele scores to have mean zero and standard deviation one.

4.5. Descriptive statistics

We tested whether the height, the height variants, and the allele scores were associated with the covariates using robust linear regression. Furthermore, we used linear regression to estimate the associations of

153 pairwise correlations of the 18 nongenetic covariates, and the 3240 pairwise associations of the 180 genetic variants and the 18 nongenetic covariates. We compared the number of significant associations (at $\alpha = 0.05, 0.01$, and 0.001) that occurred to the number that would be expected if all the variables were uncorrelated. Finally, we tested the strength of the association of the instruments and height using linear regression and report the F-statistics for this association. We further report R-squared values as measures of the proportion of the variance in height explained by each definition of the instruments.

We compared the different instrumental variable methods for estimating the effect of the height on lung function with conventional unadjusted OLS regression of lung function on height and the same association adjusted for all the covariates mentioned earlier. We used the user-written `ice` command in Stata to impute missing values of the covariates for the adjusted OLS analysis [45, 46]. We do not adjust any of the instrumental variable results for covariates. We report five versions of instrumental variable estimation results using all individual variants as instruments: (i) 2SLS, (ii) LIML, (iii) LIML with corrected standard errors, (iv) CUE, and (v) CUE with corrected standard errors. We further present results for (vi) 2SLS using an unweighted allele score as a single instrument and (vii) 2SLS using a weighted allele score as a single instrument.

For each estimation method, we report – where possible – the weak identification F-statistic, Hausman endogeneity test, and the Sargan or Hansen J-test of overidentifying restrictions [24, 47–51]. The Hausman test evaluates whether there is any evidence that the instrumental variable estimate differs from the OLS estimate. The Sargan and Hansen J-tests of overidentification evaluate whether there is any evidence of difference between the instrumental variable results based on each variant. The Sargan test assumes conditional homoskedasticity, whereas the Hansen J-test is robust to conditional heteroskedasticity [24, 52]. Hence, the Sargan test can be more powerful but misleading when conditional heteroskedasticity is present. We ran all analyses using Stata 13.0. The instrumental variable results were estimated with the user-written commands `ivreg2` [53, 54]. We report the corrected standard errors for both CUE and LIML using the user-written commands `nwind` and `bekker` [18, 22, 31].

To illustrate how the many weak instrument bias affects the results in the analysis that includes the individual variants, we sorted the height variants in order of effect size and repeated the analysis adding one variant at a time, starting with the variant with the biggest effect and ending with the weakest variant. We report the estimates and standard errors for each of the regressions using 2SLS and the CUE and also report F-statistics for the hypothesis that the concentration parameter equals zero ($\mu^2 = 0$).

5. Results

After cleaning and imputation, there were 8365 individuals with genome-wide data. We excluded data from 4734 individuals, who had missing measures of either height or lung function at age 15 years, which resulted in a final analysis sample of 3631. The characteristics of the individuals are described in Table II. Table III shows that the individuals' actual height at age 15 years was strongly associated with 9 of the 17 covariates ($p < 0.05$). Taller individuals were more likely to be from a higher-income household, have more educated parents and grandparents, have been breast-fed, and have an older mother. In contrast, there was little evidence of associations between the allele scores and the covariates (Table IV). Assuming no associations between the nongenetic covariates (153 possible pairwise correlations), we would expect to detect eight associations at the 5% significance level, two at the 1% significance level, and less than one at the 0.01% significance level. We detected 82 (54%), 70 (46%), and 54 (35%) associations at the 5%, 1%, and 0.01% significance level, respectively. For the 3240 genetic pairwise associations with the nongenetic covariates, we would expect to detect 162, 32, and less than one association at the 5%, 1%, and 0.01% significance level, respectively. We detected 166, 35, and no associations, respectively. This is consistent with the previous studies [8, 55].

The associations between the individual genetic variants and height were similar to the weights reported by Lango *et al.* ($R^2 = 0.478$) (Table S1 and Figure 1). Table V shows that there is a strong association between individuals' height and the allele scores. The F-statistics for the unweighted and weighted allele scores were 164 and 190, respectively, suggesting that the results based on these instruments are unlikely to suffer from weak instrument bias [12]. A one-standard-deviation increase in the unweighted allele score was associated with a 0.21 (95% confidence interval (CI): 0.18, 0.24) standard deviation increase in height at 15. A one-standard-deviation increase in the weighted allele score was associated with a 0.22 (95% CI: 0.19, 0.25) standard deviation increase in height at 15. The unweighted and weighted allele scores explained 4.3% and 4.9% of the variation in height, respectively.

Table II. Baseline characteristics of included ALSPAC participants.

	Mean	Standard deviation	*N*
Height at age 15 years (in cm)	169.35	8.42	3631
Male	0.48	0.50	3631
Age in months at Teen Focus 3 Clinic	185.31	3.58	3631
Birth weight (g)	3444	525	3435
No. of older siblings in household	1.02	0.99	3319
No. of younger siblings in household	0.92	0.95	3319
Ln(income)	5.74	0.46	3141
Mother's education	3.39	1.18	3391
Father's education	3.39	1.37	3320
Mothers' mother's education	2.35	1.33	2627
Mothers' father's education	2.57	1.46	2471
Child not ever raised by natural father	0.11	0.32	3314
Father's social class at birth	2.78	1.27	3146
Mother works part time	0.43	0.50	3113
Mother works full time	0.09	0.28	3113
Partner works full time	0.93	0.25	1608
Mother drank during pregnancy	0.57	0.50	3419
Mother smoked during pregnancy	0.15	0.36	3418
Ever breast-fed	0.86	0.34	3226
Mother's age	29.82	4.51	3495
Participant had tried tobacco at age 8 years	0.02	0.15	3133

Table III. Association of height at age 15 years with baseline characteristics.

	Actual height		
	N	Coef	*p*-value
Age in months at Teen Focus 3 Clinic	3631	0.14	0.02
Male	3631	0.29	<0.001
Older siblings in household	3321	0.00	0.98
Younger siblings in household	3321	0.02	0.30
Ln(income)	3141	0.02	0.03
Mother's education	3391	0.06	0.002
Father's education	3320	0.08	0.001
Mothers' mother's education	2627	0.06	0.02
Mothers' father's education	2471	0.06	0.05
Child not ever raised by natural father	3314	0.00	0.62
Father's social class at birth	3146	−0.04	0.07
Mother works part time	3113	0.00	0.72
Mother works full time	3113	0.01	0.07
Partner works full time	1608	0.00	0.57
Mother drank during pregnancy	3419	0.00	0.85
Mother smoked during pregnancy	3418	−0.01	0.08
Ever breast-fed	3226	0.02	<0.001
Mother's age	3495	0.27	<0.001
Participant had tried tobacco at age 8 years	3133	0.01	0.03

Coef, coefficient from a robust ordinary least squares regression of covariate on normalized height.

Table IV. Association of variants and allele scores with covariates.

	Unweighted allele score		Weighted allele score	
	Coef	*p*-value	Coef	*p*-value
	(1)	(2)	(3)	(4)
Male	0.01	0.41	0.01	0.33
Birth weight (g)	31.47	<0.001	34.43	<0.001
Older siblings in household	0.01	0.60	0.01	0.45
Younger siblings in household	−0.02	0.19	−0.02	0.25
Ln(income)	0.00	0.62	−0.01	0.37
Mother's education	0.00	0.94	0.00	0.99
Father's education	0.00	0.94	0.00	0.84
Mothers' mother's education	0.02	0.50	0.02	0.42
Mothers' father's education	0.02	0.50	0.03	0.29
Child not ever raised by natural father	0.00	0.51	0.01	0.36
Father's social class at birth	0.00	0.83	0.00	0.99
Mother works part time	0.01	0.15	0.01	0.24
Mother works full time	0.00	0.66	0.00	0.50
Partner works full time	−0.01	0.21	−0.01	0.08
Mother drank during pregnancy	−0.01	0.52	−0.01	0.25
Mother smoked during pregnancy	−0.01	0.35	−0.01	0.18
Ever breast-fed	0.00	0.58	0.00	0.71
Mother's age	−0.09	0.21	−0.09	0.23
Participant had tried tobacco at age 8 years	0.00	0.58	0.00	0.32

Table V. Association of allele scores and normalized height at age 15 years ($N = 3631$).

	Unweighted score mean difference (95% confidence interval)	*p*-value	Weighted score mean difference (95% confidenceinterval)	*p*-value
Allele score	0.21 (0.18, 0.24)	<0.001	0.22 (0.19,0.25)	<0.001
R^2	0.043		0.049	
F-statistic	164		190	

Table VI. The relationship between height and lung function ($N = 3631$).

Method	Mean difference (95% confidence intervals)	Robust standard error	*p*-value	*F*-statistic	Sargan/Hansen *J*-test *p*-value	Hausman endogeneity tests	*p*-value
Ordinary least squares	0.67 (0.65, 0.70)	0.013	<0.001				
Adjusted OLS	0.53 (0.50, 0.56)	0.015	<0.001				
Two-stage least squares	0.60 (0.52, 0.68)	0.040	<0.001	2.03	0.02	4.08	0.04
LIML	0.47 (0.34, 0.60)	0.067	<0.001	2.03	0.01*	3.70	0.05
LIML corrected	0.47 (0.25, 0.68)	0.109	<0.001				
CUE	0.43 (0.35, 0.50)	0.039	<0.001	2.03	0.05		
CUE corrected	0.43 (0.21, 0.64)	0.110	<0.001				
Unweighted allele score	0.44 (0.32, 0.56)	0.062	<0.001	164.18		16.80	<0.001
Weighted allele score	0.42 (0.31, 0.53)	0.057	<0.001	190.15		22.71	<0.001

Robust confidence intervals. Adjusted OLS adjusts for covariates described in Table II.
OLS, ordinary least squares; LIML, limited information maximum likelihood; CUE, continuously updating estimator.
*We used the Sargan test for LIML and the Hansen *J*-test for the other estimators. The Hausman test assumes homoskedasticity; therefore, we do not include it for CUE.

Using conventional OLS regression, we found that a one-standard-deviation increase in height was associated with a 0.67 (95% CI: 0.65, 0.70) standard deviation increase in lung function (Table VI and Figure 2). After we adjusted for the observed covariates, this association was attenuated to 0.53 (95% CI: 0.50, 0.56). The 2SLS estimates using the individual alleles of all variants suggested that a one-standard-deviation increase in height caused a 0.60 (95% CI: 0.52,0.68) standard deviation increase in

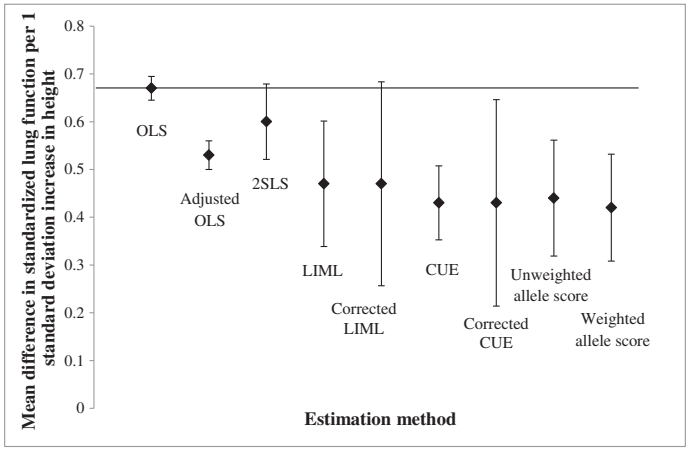

Figure 2. Estimated effect of standardized height on lung function. These results are presented in Table VI. OLS, ordinary least squares; 2SLS, two-stage least squares regression; LIML, limited information maximum likelihood; CUE, continuously updating estimator. The horizontal line indicates the OLS estimate. Lung function standardized to mean zero and standard deviation one.

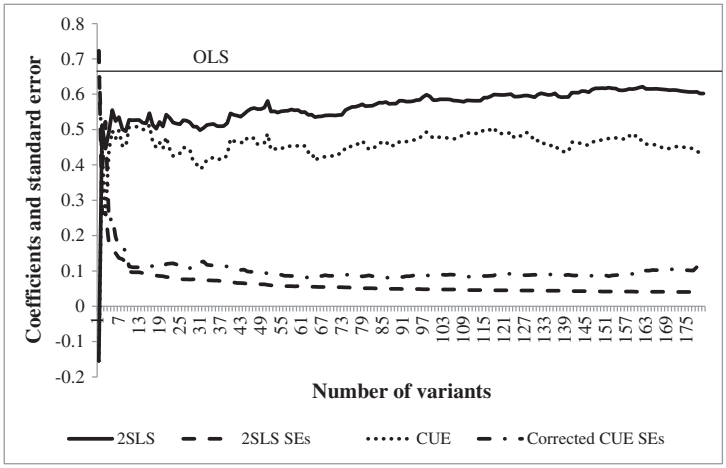

Figure 3. Coefficients and standard errors from two-stage least squares and continuously updating estimator by number of variants that are included as instruments. Variants included in order of association with height; thus, the analysis with one variant uses the strongest variant reported by Lango *et al*., and the analysis with two variants uses the two strongest variants and so on.

lung function (Table VI and Figure 2). The 2SLS estimates are relatively close to the OLS estimates. The LIML estimate suggested that a one-standard-deviation increase in height caused a 0.47 (95% CI: 0.34, 0.60) standard deviation increase in lung function, similar to the CUE estimate. The corrected standard error for the CUE is virtually identical to the Bekker standard error for LIML. The Sargan/Hansen tests for the 2SLS, LIML, and CUE results suggested some evidence of differences between the estimates based on each variant.

The instrumental variable results using an unweighted allele score were close to the LIML and CUE results, but the latter's uncorrected standard errors were 8.3% smaller and 36.2% smaller, respectively. The weighted allele score results were almost identical to the unweighted allele score, but the standard error was reduced by an additional of 7.9%. The standard errors of the weighted allele score estimates are 48% smaller than those from the corrected CUE or the corrected LIML analyses. As shown by the Hausman test, the allele score specifications found strong evidence of differences between the OLS and instrumental variable estimates.

As we have shown earlier, the 2SLS estimate using all variants as instruments is relatively close to the OLS estimate. This is likely to be because 2SLS is biased towards OLS when using many weak instruments [10]. We illustrate the consequences of the many weak instrument bias in 2SLS further in Figure 3

and Figure S1. When we use fewer than 15 variants, the 2SLS and CUE estimates are similar. However, as we move along the *x*-axis and add more variants, the 2SLS estimates trend upwards, becoming more biased towards the OLS estimate, whilst the CUE remains relatively stable.

6. Discussion

When using many weak instruments in Mendelian randomization studies, researchers should consider the fact that the routinely used 2SLS estimator is biased. So far, two methods have been proposed in the epidemiology and econometric literature to account for this problem. First, constructing a weighted or unweighted allele score and using 2SLS with only one instrument. Second, using the full set of instruments together with an estimation method that is robust to using many weak instruments. In the latter case, only LIML has been discussed in the literature [13]. However, the conventional standard errors for LIML are incorrect with many weak instruments, leading to inaccurate inference for testing hypotheses. We have developed a Stata routine to calculate corrected standard errors for LIML to account for the many weak instruments problem and show that these adequately reflect the decrease in precision of the LIML estimator for these cases.

The LIML has recently been shown to be biased with many weak instruments when the errors are conditionally heteroskedastic [33]. The CUE is consistent under these conditions, and therefore, the CUE would be the estimator of choice in most empirical applications; for example, when estimating risk differences for binary outcomes. However, its standard errors also need to be adjusted when using many weak instruments. We have developed a Stata routine also for the CUE to adjust the standard errors and, again, show that these adequately reflect the decrease in precision when using many weak instruments.

Using a single allele score as instrument can be a useful and appropriate way to get around the many weak instruments problem. In contrast, the CUE will be useful in situations when there is no clear way to aggregate a set of instruments. For example, consider a researcher who is interested in the relative contributions of body mass index (BMI), fat mass and lean mass to asthma [56]. Whilst 32 genetic variants have been shown to associate with BMI, there have been no genome-wide association studies specifically for fat and lean body mass. Granell *et al.* [56] used the same weighted allele score as an instrument for lean and fat mass as the one derived from a meta-analysis for the gene–BMI relationship. As the weights are unlikely to be the same for lean and fat mass, efficiency may be enhanced by using the CUE and the 32 BMI variants to accurately test whether BMI, lean or fat mass affects asthma.

The CUE may be particularly useful in cases where there are multiple risk factors of interest and many variants, as it may be difficult then to construct different allele scores for each of the risk factors. Indeed, in the example earlier, one could use the CUE to estimate the causal effects of lean and fat mass jointly. Outside the world of Mendelian randomization, recent pharmaco-epidemiological studies have found that including multiple physicians' previous prescriptions as instruments for their prescribing behavior increases the precision of estimates [57–59]. These results may be improved upon by using the CUE.

In this proof of principle study, we showed how to apply these methods using the illustrative example of the effect of height on lung function. We found that the height variants and allele scores were associated with height and that the genetic variants were no more associated with the observed covariates than would be expected by chance. Our results suggest that, compared with the causal effects identified by the genetic variants, the observational association may suffer from a slight upward bias. We demonstrated that the 2SLS estimates suffered from many weak instruments bias when we used all the individual variants. In contrast, the results using LIML, CUE, and the weighted or unweighted allele scores instrumental variable estimators were mutually in concordance, and all suggested a weaker underlying causal relationship between height and lung function. As with the simulated results, in the empirical example, the corrected standard errors for LIML were almost identical to the corrected CUE standard errors but considerably larger than those for the weighted allele score instrumental variable estimates. This is likely to reflect the relative efficiency of these approaches in many applied examples. Nevertheless, by holding the number of instruments constant, as the sample size increases, the difference in efficiency between these approaches will reduce.

In conclusion, when using many weak instruments, 2SLS estimates can be biased. We demonstrated that, under homoskedasticity, the allele score instrumental variable estimators, CUE, and LIML with corrected standard errors provide accurate inferences. If homoskedasticity holds, the LIML is likely to be slightly more efficient than the CUE. In small samples, both estimators are likely to be less efficient than the allele score. However, if there is conditional heteroskedasticity, LIML can be inconsistent when using many weak instruments and researchers should use the CUE instead. However, both methods are likely

to be less efficient than using allele scores when there are many instruments and modest sample sizes. Using multiple variants as instruments can increase precision of the results over using single variants, but researchers must report corrected standard errors.

Acknowledgements

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses.

This work was supported by the United Kingdom Medical Research Council (MRC), the Wellcome Trust, and the University of Bristol who provide core support for Avon Longitudinal Study of Parents and Children (ALSPAC). The Integrative Epidemiology Unit is supported by the Medical Research Council and the University of Bristol (G0600705, MC_UU_12013/1-9). Funding from the European Research Council grant DevHEALTH (269874) that supported the specific work presented in this paper is gratefully acknowledged. Helmut Farbmacher received funding from the Fritz Thyssen Stiftung. Stephanie von Hinke Kessler Scholder is funded by an MRC Early Career Fellowship (G1002345). Stephen Burgess is funded by the Wellcome Trust fellowship (100114). No funding body has influenced data collection, analysis, or its interpretations. This publication is the work of the authors, who serve as the guarantors for the contents of this paper.

References

1. C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC). Association between C reactive protein and coronary heart disease: Mendelian randomisation analysis based on individual participant data. *BMJ* 2011; **342**:d548–d548.
2. Davey Smith G, Ebrahim S. "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* 2003; **32**(1):1–22.
3. Timpson NJ, Lawlor DA, Harbord RM, Gaunt TR, Day INM, Palmer LJ, Hattersley AT, Ebrahim S, Lowe GDO, Rumley A, Davey Smith G. C-reactive protein and its role in metabolic syndrome: Mendelian randomisation study. *Lancet* 2005; **366**(9501):1954–1959.
4. Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, Jensen MK, Hindy G, Hólm H, Ding EL, Johnson T, Schunkert H, Samani NJ, Clarke R, Hopewell JC, Thompson JF, Li M, Thorleifsson G, Newton-Cheh C, Musunuru K, Pirruccello JP, Saleheen D, Chen L, Stewart AFR, Schillert A, Thorsteinsdottir U, Thorgeirsson G, Anand S, Engert JC, Morgan T, Spertus J, Stoll M, Berger K, Martinelli N, Girelli D, McKeown PP, Patterson CC, Epstein SE, Devaney J, Burnett M-S, Mooser V, Ripatti S, Surakka I, Nieminen MS, Sinisalo J, Lokki M-L, Perola M, Havulinna A, de Faire U, Gigante B, Ingelsson E, Zeller T, Wild P, de Bakker PIW, Klungel OH, Maitland-van der Zee A-H, Peters BJM, de Boer A, Grobbee DE, Kamphuisen PW, Deneer VHM, Elbers CC, Onland-Moret NC, Hofker MH, Wijmenga C, Verschuren WMM, Boer JMA, van der Schouw YT, Rasheed A, Frossard P, Demissie S, Willer C, Do R, Ordovas JM, Abecasis GR, Boehnke M, Mohlke KL, Daly MJ, Guiducci C, Burtt NP, Surti A, Gonzalez E, Purcell S, Gabriel S, Marrugat J, Peden J, Erdmann J, Diemert P, Willenborg C, König IR, Fischer M, Hengstenberg C, Ziegler A, Buysschaert I, Lambrechts D, Van de Werf F, Fox KA, El Mokhtari NE, Rubin D, Schrezenmeir J, Schreiber S, Schäfer A, Danesh J, Blankenberg S, Roberts R, McPherson R, Watkins H, Hall AS, Overvad K, Rimm E, Boerwinkle E, Tybjaerg-Hansen A, Cupples LA, Reilly MP, Melander O, Mannucci PM, Ardissino D, Siscovick D, Elosua R, Stefansson K, O'Donnell CJ, Salomaa V, Rader DJ, Peltonen L, Schwartz SM, Altshuler D, Kathiresan S. Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study. *Lancet* 2012; **380**(9841):572–580.
5. Nordestgaard BG, Palmer TM, Benn M, Zacho J, Tybjærg-Hansen A, Davey Smith G, Timpson NJ. The effect of elevated body mass index on ischemic heart disease risk: causal estimates from a Mendelian randomisation approach. *PLoS Medicine* 2012; **9**(5):e1001212.
6. Brunner EJ, Kivimäki M, Witte DR, Lawlor DA, Davey Smith G, Cooper JA, Miller M, O Lowe GD, Rumley A, Casas JP, Shah T, Humphries SE, Hingorani AD, Marmot MG, Timpson NJ, Kumari M. Inflammation, insulin resistance, and diabetes – Mendelian randomization using CRP haplotypes points upstream. *PLoS Medicine* 2008; **5**(8):e155.
7. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* 2008; **27**(8):1133–1363.
8. Davey Smith G, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Medicine* 2007; **4**(12):e352.
9. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics* 2014; **23**:R89–R98.
10. Bound J, Jaeger D, Baker R. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 1995; **90**(430):443–450.
11. Hernán MA, Robins J. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006; **17**(4):360–372.
12. Staiger D, Stock J. Instrumental variables regression with weak instruments. *Econometrica* 1997; **65**(3):557–586.
13. Burgess S, Thompson SG. Bias in causal estimates from Mendelian randomization studies with weak instruments. *Statistics in Medicine* 2011; **30**(11):1312–1323.
14. Pierce BL, Ahsan H, VanderWeele TJ. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *International Journal of Epidemiology* 2010; **40**(3):740–752.

15. Burgess S, Thompson SG, CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology* 2011; **40**(3):755–764.
16. Palmer TM, Lawlor DA, Harbord RM, Sheehan NA, Tobias JH, Timpson NJ, Davey Smith G, Sterne JA. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistics Methods Medicine Research* 2012; **21**(3):223–242.
17. Hansen C, Hausman J, Newey W. Estimation with many instrumental variables. *Journal of Business and Economic Statistics* 2008; **26**(4):398–422.
18. Newey W, Windmeijer F. GMM with many weak moment conditions. *Econometrica* 2009; **77**(3):687–719.
19. Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology* 2013; **42**(4):1134–1144.
20. Palmer TM, Lawlor DA, Harbord RM, Sheehan NA, Tobias JH, Timpson NJ, Davey Smith G, Sterne JAC. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods Medical Research* 2012; **21**(3):223–242.
21. Clarke PS, Palmer TM, Windmeijer F. Estimating structural mean models with multiple instrumental variables using the generalised method of moments, The Centre for Market and Public Organisation, 2011.
22. Farbmacher H. GMM with many weak moment conditions: replication and application of Newey and Windmeijer (2009). *Journal of Applied Econometrics* 2012; **27**(2):343–346.
23. Angrist JD, Krueger A. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics* 1991; **106**(4):979–1014.
24. Hansen LP. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* 1982; **50**(4):1029–1054.
25. Hansen LP, Heaton J, Yaron A. Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics* 1996; **14**(3):262–280.
26. Bun MJG, Windmeijer F. A comparison of bias approximations for the two-stage least squares (2SLS) estimator. *Economics Letters* 2011; **113**(1):76–79.
27. Nagar AL. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica* 1959; **27**(4):575–595.
28. Rothenberg TJ. *Chapter 15 Approximating the Distributions of Econometric Estimators and Test Statistics*, Handbook of Econometrics[Internet]. Elsevier: Amsterdam, 1984. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1573441284020079. DOI: 10.1016/S1573-4412(84)02007-9.
29. Newey W, Smith RJ. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 2004; **72**(1):219–255.
30. Newey WK, Smith RJ. *Asymptotic Bias and Equivalence of GMM and GEL Estimators [Internet]*. Citeseer, 2001. Available from: http://economics.mit.edu/files/1097.
31. Bekker PA. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* 1994; **62**(3):657–681.
32. Baum CF, Schaffer ME, Stillman S. Enhanced routines for instrumental variables/GMM estimation and testing. *Stata Journal* 2007; **7**(4):465–506.
33. Hausman JA, Newey WK, Woutersen T, Chao JC, Swanson NR. Instrumental variable estimation with heteroskedasticity and many instruments: instrumental variable estimation. *Quantitative Economics* 2012; **3**(2):211–255.
34. Stock J, Wright J, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* 2002; **20**(4):518–529.
35. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, Molloy L, Ness A, Ring S, Davey Smith G. Cohort profile: the "Children of the 90s" – the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology* 2012; **42**(1):111–127.
36. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, Henderson J, Macleod J, Molloy L, Ness A, Ring S, Nelson SM, Lawlor DA. Cohort profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International Journal of Epidemiology* 2012; **42**(1):97–110.
37. Bristol U of. Data dictionary, ALSPAC, Bristol University [Internet]. Available from: http://www.bristol.ac.uk/alspac/researchers/data-access/data-dictionary/ [Accessed on 13 May 2013].
38. Kotecha SJ, Watkins WJ, Paranjothy S, Dunstan FD, Henderson AJ, Kotecha S. Effect of late preterm birth on longitudinal lung spirometry in school age children and adolescents. *Thorax* 2011; **67**(1):54–61.
39. Kotecha SJ, Watkins WJ, Heron J, Henderson J, Dunstan FD, Kotecha S. Spirometric lung function in school-age children: effect of intrauterine growth retardation and catch-up growth. *American Journal of Respiratory and Critical Care Medicine* 2010; **181**(9):969–974.
40. Howe LD, Matijasevich A, Tilling K, Brion M-J, Leary SD, Davey Smith G, Lawlor DA. Maternal smoking during pregnancy and offspring trajectories of height and adiposity: comparing maternal and paternal associations. *International Journal of Epidemiology* 2012; **41**(3):722–732.
41. von Hinke Kessler Scholder S, Davey Smith G, Lawlor DA, Propper C, Windmeijer F. Child height, health and human capital: evidence using genetic markers. *European Economic Review* 2013; **57**:1–22.
42. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segrè AV, Speliotes EK, Wheeler E, Soranzo N, Park J-H, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Mägi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson Å, Carola Zillikens M, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga J-J, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Hua Zhao J, Nyholt DR, Pellikka N, Perola M, Perry JRB, Surakka I, Tammesoo M-L, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C,

Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, König IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Müller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De Grandi A, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer NB, Geus EJC, Glorioso N, Haiqing S, Hartikainen A-L, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpeläinen TO, Koiranen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki M-L, Marusic A, Maschio A, Meitinger T, Mulas A, Paré G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietiläinen KH, Pouta A, Ridderstråle M, Rotter JI, Sambrook JG, Sanders AR, Oliver Schmidt C, Sinisalo J, Smit JH, Stringham HM, Bragi Walters G, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P, Alavere H, Farrall M, McArdle WL, Nelis M, Peters MJ, Ripatti S, van Meurs JBJ, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den Heijer M, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kähönen M, Kaprio J, Kathiresan S, Kiemeney L, Kocher T, Launer LJ, Lehtimäki T, Melander O, Mosley Jr TH, Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tönjes A, Tuomilehto J, van Ommen G-J, Viikari J, Heath AC, Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Grönberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Mark Lathrop G, Salomaa V, Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Adrienne Cupples L, Erdmann J, Fox CS, Gudnason V, Gyllensten U, Harris TB, Hayes RB, Jarvelin M-R, Mooser V, Munroe PB, Ouwehand WH, Penninx BW, Pramstaller PP, Quertermous T, Rudan I, Samani NJ, Spector TD, Völzke H, Watkins H, Wilson JF, Groop LC, Haritunians T, Hu FB, Kaplan RC, Metspalu A, North KE, Schlessinger D, Wareham NJ, Hunter DJ, O'Connell JR, Strachan DP, Wichmann H-E, Borecki IB, van Duijn CM, Schadt EE, Thorsteinsdottir U, Peltonen L, Uitterlinden AG, Visscher PM, Chatterjee N, Loos RJF, Boehnke M, McCarthy MI, Ingelsson E, Lindgren CM, Abecasis GR, Stefansson K, Frayling TM, Hirschhorn JN. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 2010; **467**(7317):832–838.

43. Paternoster L, Zhurov AI, Toma AM, Kemp JP, St. Pourcain B, Timpson NJ, McMahon G, McArdle W, Ring SM, Davey Smith G, Richmond S, Evans DM. Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. *The American Journal of Human Genetics* 2012; **90**(3): 478–485.
44. Fatemifar G, Hoggart CJ, Paternoster L, Kemp JP, Prokopenko I, Horikoshi M, Wright VJ, Tobias JH, Richmond S, Zhurov AI, Toma AM, Pouta A, Taanila A, Sipila K, Lähdesmäki R, Pillas D, Geller F, Feenstra B, Melbye M, Nohr EA, Ring SM, St Pourcain B, Timpson NJ, Davey Smith G, Jarvelin M-R, Evans DM. Genome-wide association study of primary tooth eruption identifies pleiotropic loci associated with height and craniofacial distances. *Human Molecular Genetics* 2013; **22**(18):3807–3817.
45. Royston P. Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables. *Stata Journal* 2009; **9**(3):466–477.
46. Royston P. Multiple imputation of missing values. *Stata Journal* 2004; **4**:227–241.
47. Hausman JA. Specification tests in econometrics. *Econometrica* 1978; **46**(6):1251–1271.
48. Wu DM. Alternative tests of independence between stochastic regressors and disturbances. *Econometrica* 1973; **41**(4): 733–750.
49. Wooldridge J. *Econometric Analysis of Cross Section and Panel Data*. The MIT press: Cambridge, Massachusetts, 2002.
50. Angrist JD, Pischke JS. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton Univirsity Press: Princeton, 2009.
51. Breusch TS, Pagan AR. The Lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies* 1980; **47**(1):239–253.
52. Sargan J. The estimation of economic relationships using instrumental variables. *Econometrica* 1958; **26**(3):393–415.
53. Baum C, Schaffer M, Stillman S. IVREG2: Stata module for extended instrumental variables/2SLS and GMM estimation. *Statistical Software Components* 2002. Available from: http://ideas.repec.org/c/boc/bocode/s425401.html.
54. StataCorp. *Stata Statistical Software: Release 13*. StataCorp LP: College Station, TX, 2013. Available from: http://www.stata.com/support/faqs/resources/citing-software-documentation-faqs/.
55. Von Hinke Kessler Scholder S, Wehby GL, Lewis S, Zuccolo L. Alcohol exposure *in utero* and child academic achievement. *The Economic Journal* 2014; **124**(576):634–667.
56. Granell R, Henderson AJ, Evans DM, Davey Smith G, Ness AR, Lewis S, Palmer TM, Sterne JAC. Effects of BMI, fat mass, and lean mass on asthma in childhood: a Mendelian randomization study. *PLoS Medicine* 2014; **11**(7):e1001669.
57. Davies NM, Davey Smith G, Windmeijer F, Martin RM. COX-2 selective nonsteroidal anti-inflammatory drugs and risk of gastrointestinal tract complications and myocardial infarction: an instrumental variable analysis. *Epidemiology* 2013; **24**(3):352–362.
58. Davies NM, Gunnell D, Thomas KH, Metcalfe C, Windmeijer F, Martin RM. Physicians' prescribing preferences were a potential instrument for patients' actual prescriptions of antidepressants. *Journal of Clinical Epidemiology* 2013; **66**: 1386–1396.
59. Rassen JA, Brookhart MA, Glynn R, Mittleman M, Schneeweiss S. Instrumental variables II: instrumental variable application – in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. *Journal of Clinical Epidemiology* 2009; **62**(12):1233–1241.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

