American Journal of Epidemiology

Vol. 186, No. 9
DOI: 10.1093/aje/kwx175
Advance Access publication:
June 1, 2017

Practice of Epidemiology

Correcting the Standard Errors of 2-Stage Residual Inclusion Estimators for Mendelian Randomization Studies

Tom M. Palmer*, Michael V. Holmes, Brendan J. Keating, and Nuala A. Sheehan

* Correspondence to Dr. Tom M. Palmer, Department of Mathematics and Statistics, Fylde College, Bailrigg, Lancaster University, Lancaster LA1 4YF, United Kingdom (e-mail: t.palmer1@lancaster.ac.uk).

Initially submitted December 23, 2015; accepted for publication December 21, 2016.

Mendelian randomization studies use genotypes as instrumental variables to test for and estimate the causal effects of modifiable risk factors on outcomes. Two-stage residual inclusion (TSRI) estimators have been used when researchers are willing to make parametric assumptions. However, researchers are currently reporting uncorrected or heteroscedasticity-robust standard errors for these estimates. We compared several different forms of the standard error for linear and logistic TSRI estimates in simulations and in real-data examples. Among others, we consider standard errors modified from the approach of Newey (1987), Terza (2016), and bootstrapping. In our simulations Newey, Terza, bootstrap, and corrected 2-stage least squares (in the linear case) standard errors gave the best results in terms of coverage and type I error. In the real-data examples, the Newey standard errors were 0.5% and 2% larger than the unadjusted standard errors for the linear and logistic TSRI estimators, respectively. We show that TSRI estimators with modified standard errors have correct type I error under the null. Researchers should report TSRI estimates with modified standard errors instead of reporting unadjusted or heteroscedasticity-robust standard errors.

causal inference; instrumental variables; Mendelian randomization; 2-stage predictor substitution estimators; 2-stage residual inclusion estimators

Abbreviations: BMI, body mass index; BS1, bootstrap, second stage only; BS2, bootstrap, both stages; LSMM, logistic structural mean model; SBP, systolic blood pressure; TSLS, 2-stage least squares; TSPS, 2-stage predictor substitution; TSRI, 2-stage residual inclusion.

Mendelian randomization studies aim to use genotypes as instrumental variables to test and estimate the causal effect of modifiable exposures on disease-related outcomes (1–4). A variety of instrumental variable estimators have been described and evaluated for use with data in a single study (5–12). A class of semiparametric estimators known as structural mean models have been found to be most robust to distributional assumptions for binary outcomes but can have problems with identification (7, 13–16). Therefore, researchers may wish to fit models that make more distributional assumptions.

One frequently used instrumental variable estimator is 2-stage least squares (TSLS). This is a series of 2 linear models and is most commonly applied when both the exposure and outcome variables are continuous. The first stage is a linear regression of the exposure on the instrumental variables. The second stage is a linear regression of the outcome on the predicted values of the exposure from the first stage. TSLS is consistent for the causal effect when all relationships are linear and there are no interactions between the instrument and unmeasured confounders and between the exposure and unmeasured confounders. Palmer et al. (17) investigated 2 instrumental variable estimators of the causal odds ratio for a binary outcome: the "standard" and "adjusted" logistic instrumental variable estimators. The standard logistic instrumental variable estimator replaced the linear regression in the second stage of TSLS with a logistic regression. Such estimators have been referred to as 2-stage predictor substitution (TSPS) estimators, and are written as follows (18):

$$\text{Stage 1:} \quad X = \alpha_0 + \alpha_1 Z + \varepsilon_1, \quad \varepsilon_1 \sim N(0, \sigma_1^2) \tag{1}$$

$$\text{Stage 2:} \quad h(E[Y]) = \beta_0 + \beta_1 \widehat{X}, \tag{2}$$

where X represents the exposure variable, Y the outcome variable, Z the instrumental variable, $h()$ the link function for the appropriate generalized linear model (19), and ε_1 the stage-1 residuals with variance σ_1^2.

The adjusted logistic instrumental variable estimator included the first-stage residuals alongside the predicted values of the exposure in the second-stage logistic regression (17). In the econometrics literature it is more common to fit the second stage of such estimators using the original values of the exposure (9, 18). When the residuals are included as an additive covariate, these estimators have been referred to as 2-stage residual inclusion (TSRI) estimators (18, 20, 21). If a function of the residuals is included in the second-stage model, these estimators have been referred to as control-function estimators (22). Therefore, the second stage of TSRI estimators considered here can be written as follows:

$$\text{Stage 2:}\quad h(E[Y]) = \beta_0 + \beta_1 X + \beta_2 \hat{\varepsilon}_1. \tag{3}$$

In this paper we use "linear/logistic TSRI estimator" to refer to the estimator using linear/logistic regression at the second stage (with a linear first stage).

A recent review of Mendelian randomization studies showed that TSRI estimators are commonly used but are typically being reported with unadjusted or heteroscedasticity-robust standard errors (23–34). One indication that this may not be appropriate is that when TSLS is estimated by fitting the 2 stages sequentially, the standard errors of the second-stage parameter estimates are not correct (Web Appendix 1, Web Figures 1–3, available at https://academic.oup.com/aje) (35). Interestingly, for the linear TSRI estimator, the standard error of the coefficient on the first-stage residuals is correct (36). For a binary outcome, Newey (37) developed a correction to the standard errors of the second-stage intercept and causal effect of the probit TSRI estimator. More recently Terza (38) has suggested an alternative correction. The aim of this paper is to investigate these corrections adapted to the linear and logistic TSRI estimators.

This paper proceeds by describing the probit TSRI estimator and Newey's correction for its standard errors. We then perform 2 simulation studies using binary and continuous outcomes to investigate the performance of the corrected standard errors. We then apply these corrections to a real-data example investigating the causal effect of body mass index (BMI) on systolic blood pressure (SBP) and on a binary indicator of diabetes status.

METHODS

Background to TSRI estimators

Two reviews of TSRI estimators and their application have been conducted (18, 22). The rationale for TSRI estimators is that the first-stage residuals capture some of the variability in the confounders. Therefore, the first-stage residuals can be used to correct for confounding between the exposure and the outcome, known as endogeneity in econometrics (39–43). It is well known that the linear TSRI estimator produces an estimate of the causal effect equivalent to that from TSLS (36, 44). Hausman (45) showed that the test of the coefficient of the first-stage residuals is a test for the presence of unmeasured confounding (46, 47). That it is necessary to correct the standard errors of the second-stage estimate of the causal effect of TSRI estimators has been referred to as the problem of using "generated regressors" in the second-stage model (36, 48, 49).

For binary outcomes, the use of probit TSRI estimators has been discussed (36, 50–52). There are several estimation methods available, including maximum likelihood and sequential 2-stage methods. For 2-stage estimation, a correction to the second-stage standard errors was proposed by Newey (37) and is implemented in the ivprobit and ivtobit commands in Stata (StataCorp LP, College Station, Texas) (53).

It is also important to distinguish between different causal effects. We refer to a conditional causal effect as the value of the causal effect conditioning on the unmeasured confounding and to a marginal effect as the causal effect averaged over some proportion of the unmeasured confounding. The maximum-likelihood probit TSRI estimator estimates the conditional effect, whereas the 2-stage probit and logistic TSRI estimators estimate marginal effects (12, 17, 18, 21, 53).

Probit TSRI estimator and Newey standard errors

Two-stage estimation of the probit TSRI estimator follows equations 1 and 3, where the inverse normal cumulative distribution function is used as the link function. If there are measured confounders, as with TSLS, these can be included as covariates in both stages of estimation. Letting $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_0 \end{bmatrix}$ denote the vector of estimates of the causal effect and intercept yielded by the probit TSRI estimator, and defining the matrix $\hat{D}$ as $\begin{bmatrix} \hat{\alpha}_1 & 0 \\ \hat{\alpha}_0 & 1 \end{bmatrix}$ (37, 53),

$$\hat{\beta} = (\hat{D}'\hat{\Omega}^{-1}\hat{D})^{-1}\hat{D}'\hat{\Omega}^{-1}\hat{\gamma} \tag{4}$$

The variance of the probit TSRI estimator is as follows, where $\hat{\gamma}$, $\hat{\Omega}$ and its components are defined below (37):

$$\mathrm{var}(\hat{\beta}) = (\hat{D}'\hat{\Omega}^{-1}\hat{D})^{-1} \tag{5}$$

$$\text{where } \hat{\Omega} = J_1^{-1} + \Sigma_2. \tag{6}$$

To obtain $\hat{D}$, $\hat{\gamma}$, and $\hat{\Omega}$, we use the following algorithm as described by Newey (37):

1. Perform the first-stage linear regression of X on Z to compile $\hat{D}$ and $\hat{\varepsilon}_1$.
2. Perform a probit regression of Y on Z and $\hat{\varepsilon}_1$, from which:
   - $\hat{\gamma}$ is the coefficients of Z and the estimated intercept.
   - J_1^{-1} is the variance-covariance matrix of these coefficients.
   - $\hat{\lambda}$ is denoted as the coefficient on $\hat{\varepsilon}_1$.
3. Fit the second stage of the probit TSRI estimator by a probit regression of Y on X and $\hat{\varepsilon}_1$.
   - The coefficient on X is $\hat{\beta}_1$, the estimate of the causal effect of interest.
4. Generate a new variable equal to $X(\hat{\lambda} - \hat{\beta}_1)$.
   - Perform a linear regression of this new variable on Z (also including a constant).

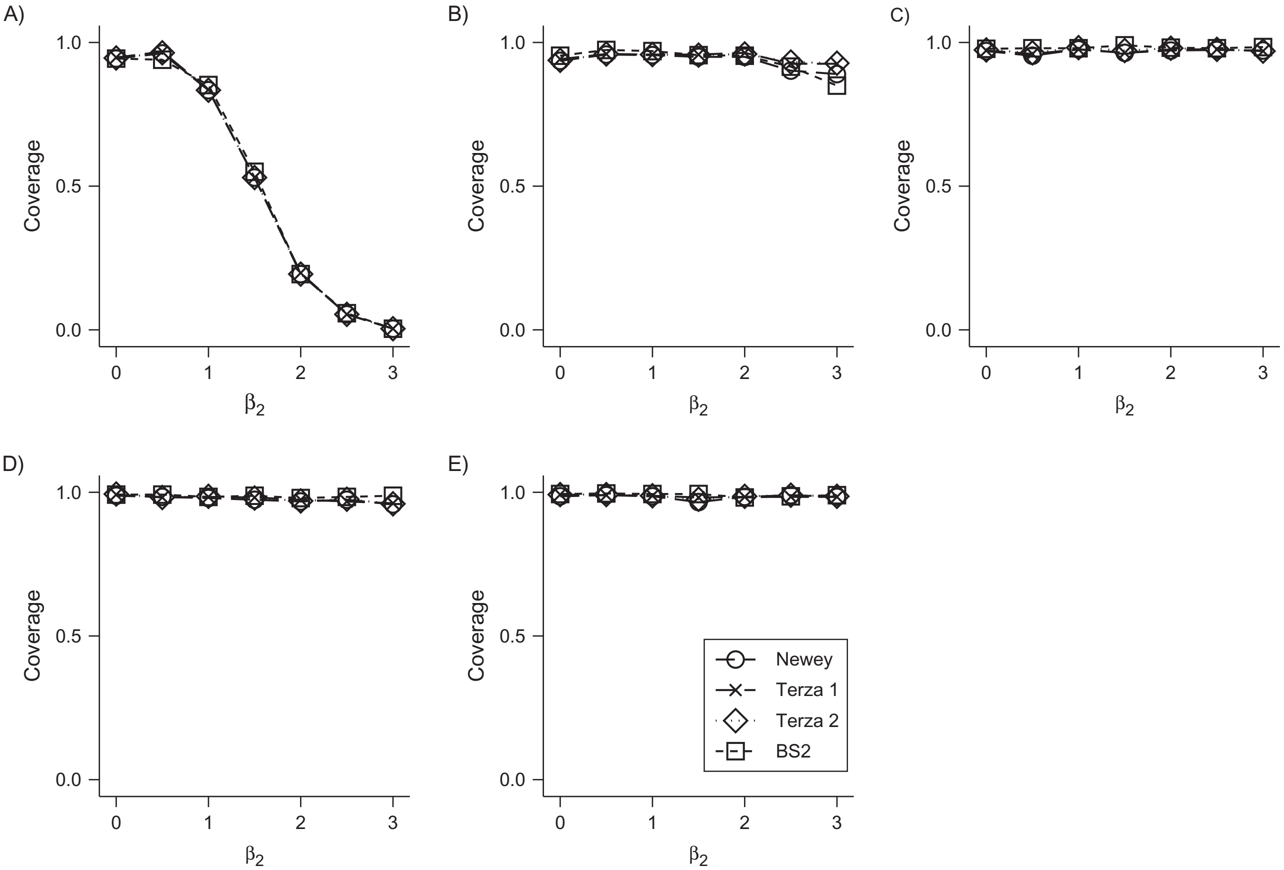

Figure 1. Coverage of the logistic 2-stage residual inclusion (TSRI) estimators for n = 1,000 with respect to the conditional parameter, $\beta_1 = 1$. The panels correspond to α_2 being set to the following values: 0 (A), 2 (B), 4 (C), 6 (D), and 8 (E). The labels Newey, Terza 1 and Terza 2, and BS2 refer to the TSRI estimator with those standard errors. BS2, bootstrapping, both stages.

- The covariance matrix from this model is the estimate of the second term in the expression for $\widehat{\Omega}$ (i.e., Σ_2).
- Add this covariance matrix to J_1^{-1}, giving $\widehat{\Omega}$.

5. Calculate $\widehat{\beta}$ and $\text{var}(\widehat{\beta})$. The standard errors of $\widehat{\beta}$ are simply the square root of the diagonal of $\text{var}(\widehat{\beta})$.

The rationale for this approach is that we obtain a standard error for our TSRI estimate that incorporates both the variability explained in the estimate by the instrumental variable Z and the predicted first-stage residuals $\widehat{\varepsilon}_1$.

To apply these standard errors to other TSRI estimators, we propose to replace the probit regressions in steps 2 and 3 with the second-stage models used by the specific TSRI estimator. Example code for Stata and R (R Foundation for Statistical Computing, Vienna, Austria) is provided in Web Appendix 2 (54, 55).

Terza (38) details an alternative algorithm for obtaining the standard error of TSRI estimators and provides example Stata code. We provide equivalent R code in Web Appendix 2. Terza uses heteroscedasticity-robust standard errors in both stages of the algorithm, which we refer to as Terza (SE) 1 (38). We additionally investigated using nonrobust standard errors, which we refer to as Terza (SE) 2. By following the code in Web Appendix 2, we can see that Terza's corrected variance-covariance matrix is the unadjusted TSRI covariance matrix plus some function of the first-stage covariance matrix.

We also investigated 2 types of nonparametric bootstrap standard errors. The first bootstraps only the second stage, which we refer to as BS1, and the second, which we refer to as BS2, bootstraps both the first and second stages. BS2 is implemented in the ivprobit and ivtobit Stata commands. For our binary outcome models we additionally investigated the probit TSRI estimator, whose estimates we converted to the odds ratio scale by dividing the estimate on the linear predictor scale by 0.6071 and taking the exponential (10). These probit estimates use Newey standard errors.

For all estimators, we calculated asymptotic normal 95% confidence interval limits as: estimate ± 1.96 × standard error.

SIMULATIONS

Logistic model simulations

Data were simulated using the basic model proposed in Palmer et al. (17) but modifying the parameter values. Specifically the data-generation model was as follows, where index i represents an observation and $\text{logit}(p_i) = \log(p_i/(1 - p_i))$:

$$
\begin{array}{ll}
g_i & \sim \text{Binomial}(2, 0.3) \\
u_i & \sim N(0, 1)\text{—representing the unmeasured confounding} \\
x_i & \sim \alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \varepsilon_{1i}, \quad \varepsilon_{1i} \sim N(0, 1) \\
\text{logit}(p_i) & = \beta_0 + \beta_1 x_i + \beta_2 u_i \\
y_i & \sim \text{Binomial}(1, p_i) \\
\multicolumn{2}{l}{\alpha_0 = 0, \alpha_1 = 1, \alpha_2 = \{0, 2, 4, 6, 8\}, \beta_0 = \log(0.05/0.95), \beta_1 = 1,} \\
\multicolumn{2}{l}{\beta_2 = [0, 3]}
\end{array}
\tag{7}
$$

Data were simulated for sample sizes of 1,000 and 5,000, and each scenario of values of α_2 and β_2, representing the effects of the unobserved confounding, was repeated 500 times. A number of different estimators were fitted to the data: the direct logistic regression of Y on X, the logistic TSPS, and the logistic TSRI with unadjusted, robust, Newey, Terza 1 and Terza 2, TSPS, and BS1 and BS2 standard errors. We also investigated the logistic structural mean model (LSMM) estimated via the generalized method of moments (GMM) (56) and the rescaled probit estimator with Newey standard errors.

In these simulations, with sample size 1,000 for the first-stage model, the average F statistics were 422, 85, 25, 12, and 7, and the average R^2 statistics were 0.30, 0.08, 0.02, 0.01, and 0.007 when α_2 was equal to 0, 2, 4, 6, and 8 respectively. With a sample size of 5,000, the average F statistics increased to 2, 104, 421, 125, 57, and 33, and the average R^2 statistics were approximately the same.

Type I error was assessed by generating the data with β_1 set to 0 (which corresponds to the null hypothesis of no causal effect) and counting the percentage of simulations for which the particular estimator gave a P value less than 0.05. Coverage was defined as a 95% confidence interval including the value of either the conditional or marginal value of β_1. Marginal values of β_1 for the estimators were obtained using the adjustments detailed in the Appendix of Palmer et al. (17) (and Web Appendix 3, Web Figure 4). Simulations were performed in Stata, version 14.1 (StataCorp LP) (54).

Linear model simulations

For a continuous outcome, the simulations were modified as follows:

$$
\begin{aligned}
y_i &\sim \beta_0 + \beta_1 x_i + \beta_2 u_i + \varepsilon_{2i}, \quad \varepsilon_{2i} \sim N(0, 1) \\
&\alpha_0 = 0, \alpha_1 = 1, \alpha_2 = \{0,2,4,6,8\}, \\
&\beta_0 = 0, \beta_1 = 1, \beta_2 = [0, 3]
\end{aligned}
\tag{8}
$$

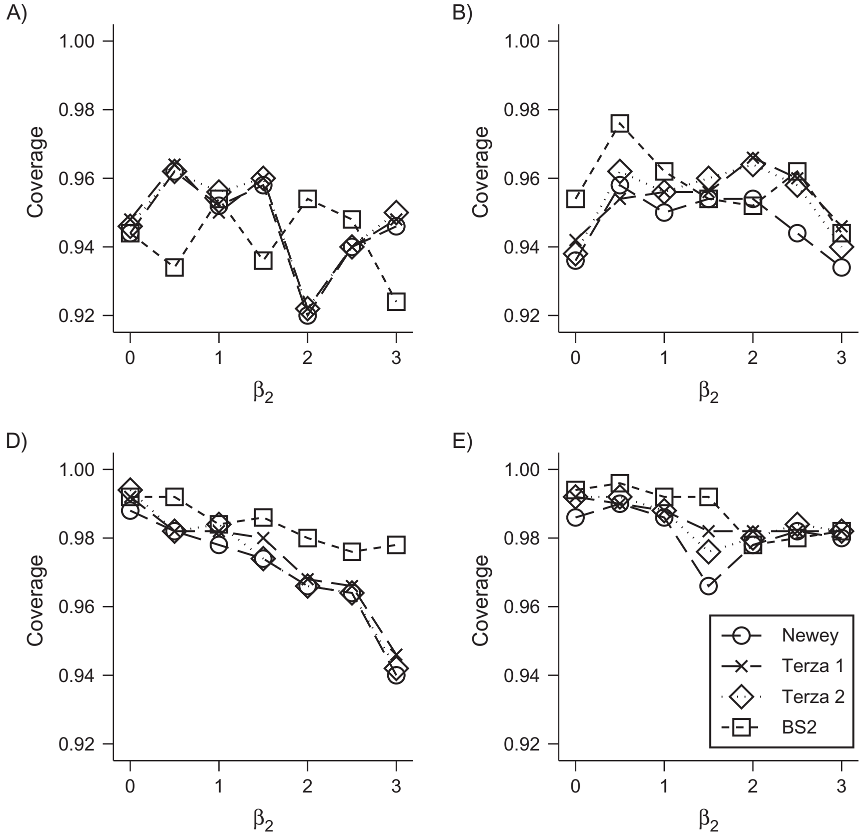

Figure 2. Coverage of the logistic 2-stage residual inclusion (TSRI) estimators for $n = 1{,}000$ with respect to the marginal parameter. The panels correspond to α_2 being set to the following values: 0 (A), 2 (B), 4 (C), 6 (D), and 8 (E). The labels Newey, Terza 1 and Terza 2, and BS2 refer to the TSRI estimator with those standard errors. BS2, bootstrapping, both stages.

For a linear second-stage model the conditional and marginal parameter values are the same. Type I error was assessed by setting β_1 to 0. A number of linear estimators were fitted to the data: the direct linear regression of Y on X, TSLS with adjusted and unadjusted (i.e., TSPS) standard errors, and the linear TSRI estimator with unadjusted, robust, Newey, Terza 1 and Terza 2, TSLS, and BS1 and BS2 standard errors.

RESULTS

Logistic model simulations

Figure 1 and Web Figure 5 show that, with respect to the conditional parameter ($\beta_1 = 1$), all estimators have low coverage at some point in the simulations. This is mainly because of the bias in the parameter estimates. The conditional coverage of several estimators (which were not expected to perform well because their standard errors do not account for the uncertainty in both stages of estimation (TSRI with unadjusted, robust, and BS1 standard errors)) was around the 95% level for some larger values of α_2. This occurred because their standard errors increased in proportion with their bias. The conditional coverage of TSRI using the Newey, Terza 1 and 2, and BS2 standard errors was the closest to 95% for the greatest proportion of simulated scenarios.

Figure 2 and Web Figure 6 show the coverage with respect to the marginal parameter values estimated by the TSRI estimator (the true marginal values are given in Figure 2 of Palmer et al. (17) and Web Appendix 3). The logistic TSPS estimator with unadjusted and robust standard errors had coverage values well below the target value of 95%. The coverage of the logistic TSRI estimator with unadjusted, robust, and BS1 standard errors was lower than the expected 95%. The coverage of the logistic TSRI estimator with TSPS standard errors was also too low and decreased as the confounding increased. However, the coverage of the logistic TSRI estimator using Newey, Terza 1 and 2, and BS2 standard errors, and the coverage of LSMM were approximately correct with values around 95%.

Figure 3 and Web Figure 7 show that the type I error of the logistic TSRI estimator with unadjusted, robust, and BS1 standard errors was too high, with values greater than the nominal level of 5%. Type I error was also too high for the logistic TSRI estimator with TSPS standard errors when there was confounding. For the LSMM estimates, the type I error was approximately correct, with values around 5%. The logistic TSPS estimator also had approximately correct type I error with unadjusted and robust standard errors with values around

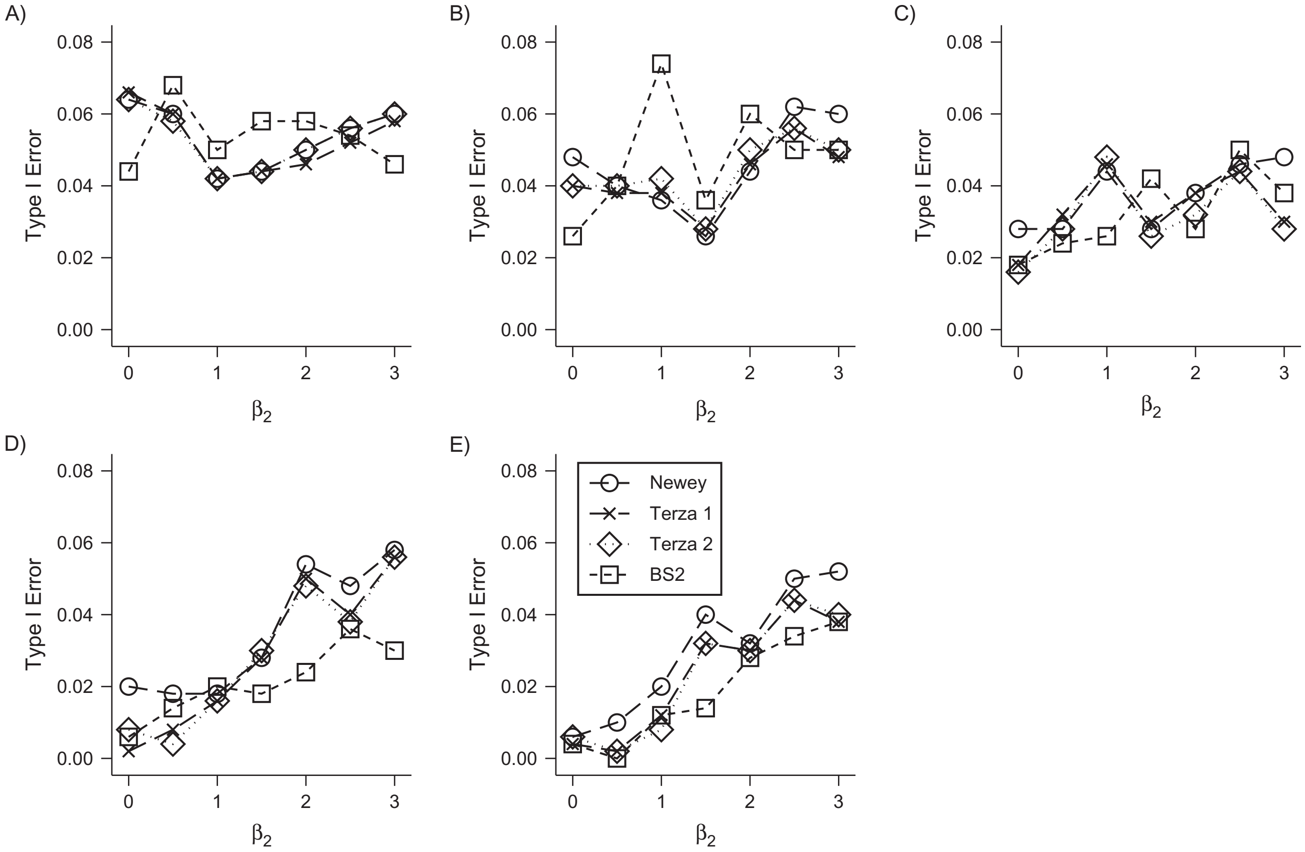

Figure 3. Type I error of the logistic 2-stage residual inclusion (TSRI) estimators for $n = 1{,}000$. The panels correspond to α_2 being set to the following values: 0 (A), 2 (B), 4 (C), 6 (D), and 8 (E). The labels Newey, Terza 1 and Terza 2, and BS2 refer to the TSRI estimator with those standard errors. BS2, bootstrapping, both stages.

5%. This is because under the null there is no substantial bias in the logistic TSPS estimates. The logistic TSRI estimator using Newey, Terza 1 and 2, and BS2 standard errors had approximately correct type I error, with values around 5%.

Similar trends can be seen in the results for the simulations using a sample size of 5,000 in Web Figures 8–10. Web Figures 2 and 3 show that the correction to the logistic TSRI standard error has the largest effect when the absolute value of the correlation between the confounders of the exposure and outcome is greater than about 0.5 or, more generally, when the effect of the confounder is stronger. The effect of the correction is also more pronounced when the outcome has a higher prevalence (up to 50%, beyond which the effect decreases).

Linear model simulations

The results in Figure 4 and Web Figure 11 show that the direct regression of Y and X has poor coverage when there is confounding. This is because of the bias in the point estimate. The TSRI estimator with unadjusted standard errors had poor coverage because the standard error does not account for the uncertainty from the first-stage estimation. TSRI using BS1 and robust standard errors also showed poor coverage for the same reason. Usually we want the standard errors for the TSRI estimate to be larger than the unadjusted standard error, but robust standard errors are often smaller. All other estimators demonstrated coverage values around 95%. The coverage values for TSRI using the Terza standard errors (1 and 2) were slightly above 95%. The coverage of TSRI using Newey and TSLS standard errors fell below 95% as the amount of confounding increased. TSRI using BS2 standard errors had the coverage values consistently closest to 95% over the range of the simulated scenarios.

The type I error results in Figure 5 and Web Figure 12 show essentially the same pattern as for the coverage results. The type I error of TSRI using unadjusted and BS1 standard errors is inflated, whereas it is approximately correct for the other TSRI standard errors. Again the type I error of the TSLS and Newey standard errors is inflated as the confounding increases. The type I error of the two Terza standard errors is slightly below 5%, and the results for BS2 are the closest to 5% over the range of the simulations.

Similar trends can also be seen in Web Figures 13 and 14. Web Figure 1 shows that the correction to the TSRI standard errors has the largest effect when the absolute value of the correlation between the confounders of the exposure and outcome residuals is greater than about 0.5 or, more generally, when the confounding is stronger.

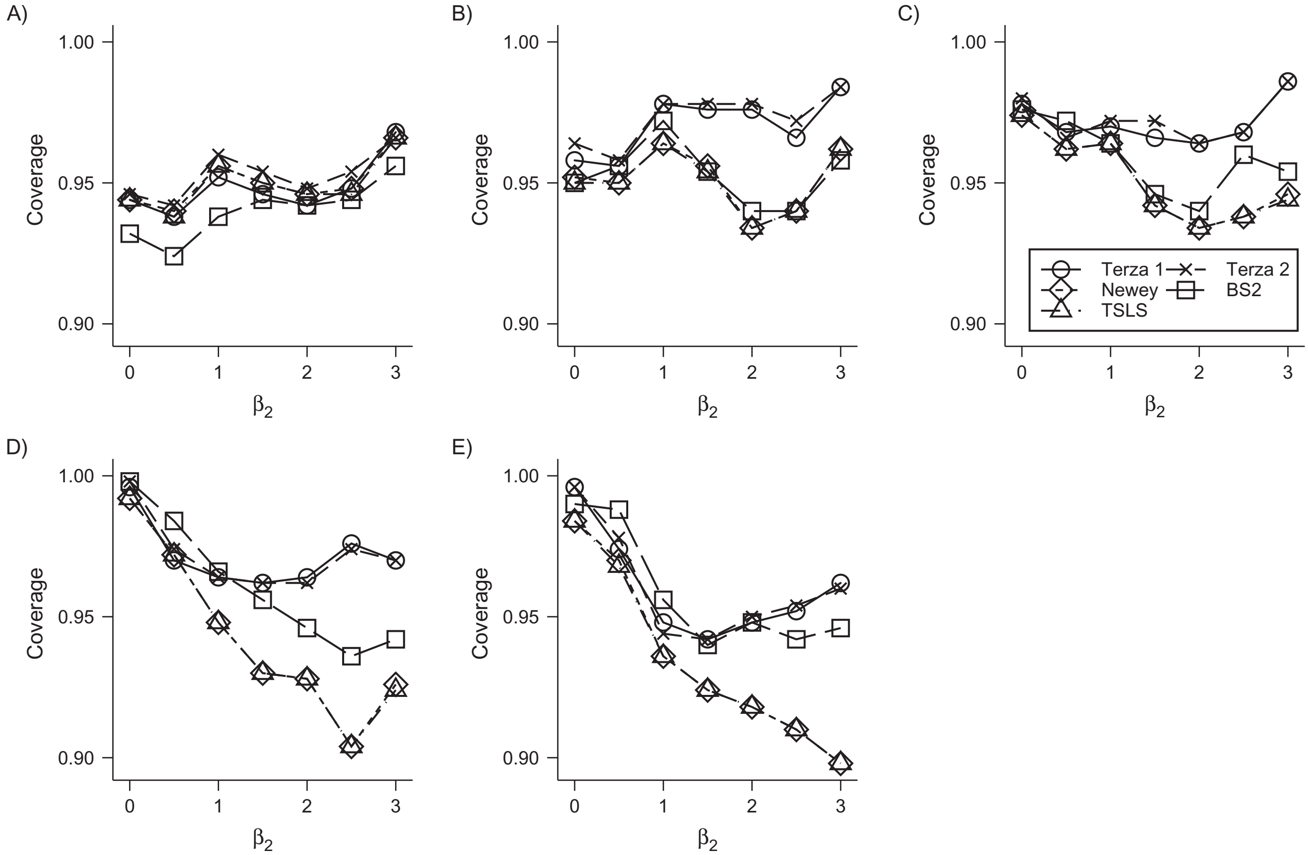

Figure 4. Coverage of the linear 2-stage residual inclusion (TSRI) estimators for $n = 1{,}000$. The panels correspond to α_2 being set to the following values: 0 (A), 2 (B), 4 (C), 6 (D), and 8 (E). The labels Newey, Terza 1 and Terza 2, and BS2 refer to the TSRI estimator with those standard errors. BS2, bootstrapping, both stages; TSLS, 2-stage least squares.

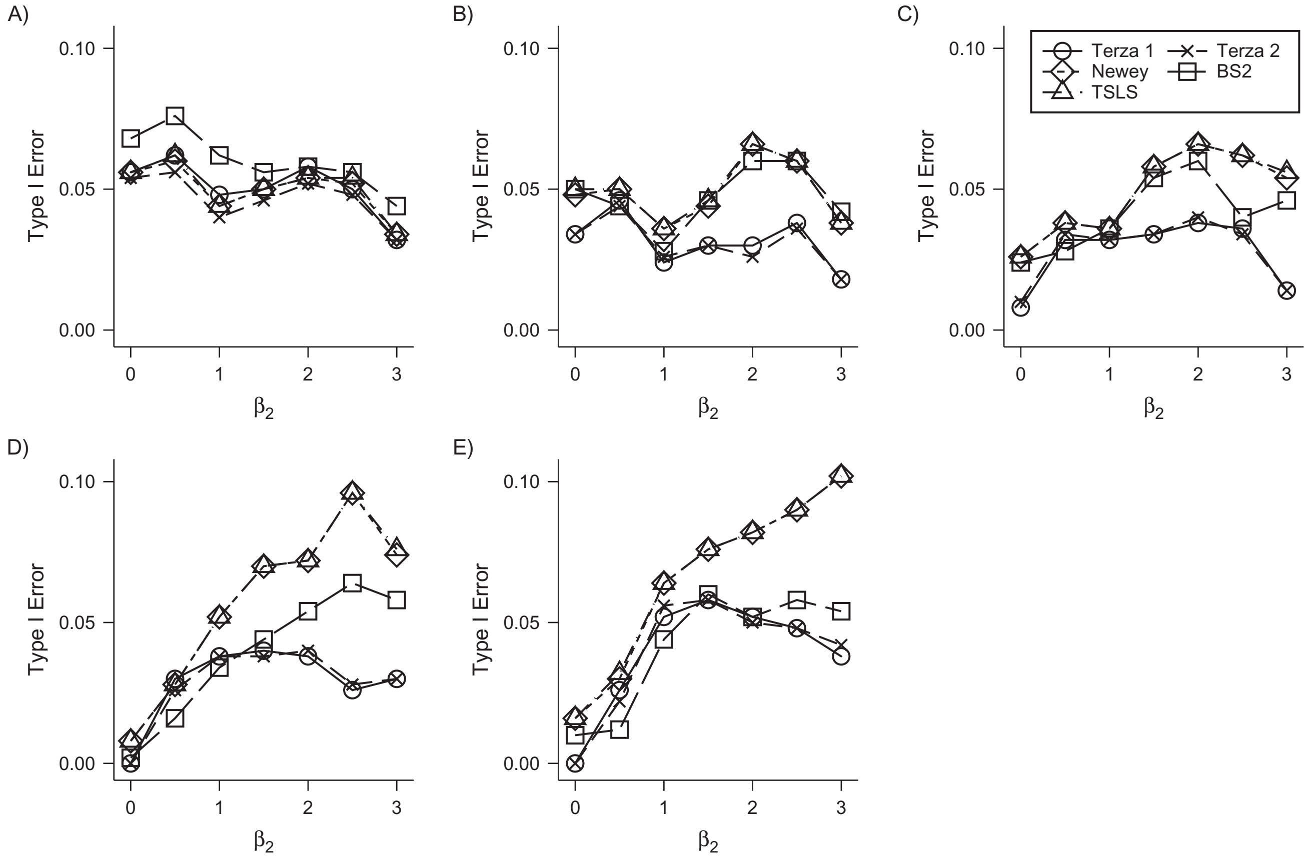

Figure 5. Type I error of the linear 2-stage residual inclusion (TSRI) estimators for n = 1,000. The panels correspond to α_2 being set to the following values: 0 (A), 2 (B), 4 (C), 6 (D), and 8 (E). The labels Newey, Terza 1 and Terza 2, and BS2 refer to the TSRI estimator with those standard errors. BS2, bootstrapping, both stages; TSLS, 2-stage least squares.

Example: causal effect of BMI on SBP and diabetes

Data were gathered on 17,057 participants from 6 prospective cohorts of European ancestry that had been genotyped with the HumanCVD BeadChip (Illumina Inc., San Diego, CA) (57). The 6 cohorts are Atherosclerosis Risk in Communities (ARIC) (58), the Cardiovascular Health Study (CHS) (59), Coronary Artery Risk Development in Young Adults (CARDIA) (60), the Framingham Heart Study (FHS) (61), Multinational Etoricoxib and Diclofenac Arthritis Long-term (MEDAL) (62), and the Multi-Ethnic Study of Atherosclerosis (MESA) (63).

Individuals had complete data on variables for BMI, SBP, and diabetes. An externally weighted allele score was constructed out of the genetic variants for BMI. Details of the genetic variants and the construction of the allele scores have been previously reported (64). In the first example, we estimate the causal effect of BMI on SBP using linear instrumental variable estimators. In the second example, we estimate the causal odds ratio for diabetes for a unit increase in BMI using binary outcome instrumental variable estimators. Analysis was performed using Stata, version 13.1 (StataCorp LP) (54).

The prevalence of the diabetes outcome was 13.7%. Table 1 shows the estimated causal odds ratios for diabetes for a 1-unit increase in BMI. The direct estimate of the odds ratio was 1.14 (95% CI: 1.13, 1.15). In the first stage of TSPS and TSRI estimation, the instrument gave a first-stage F statistic of 119, greater than the usual cutoff for a weak instrument of 10, but a low R^2 of 0.7%. The logistic TSPS estimate was larger at 1.32 (95% CI: 1.19, 1.48) and also excluded a null effect. The logistic TSRI gave the same point estimate of the causal odds ratio. For the TSRI estimator, the unadjusted standard error was 0.058, whereas the Newey standard error was 2% larger at 0.059. The Terza standard errors were 0.057 and 0.059. For logistic TSRI, the BS1 standard error was the same as the robust standard error, whereas the BS2 standard error at 0.061 was larger than the Newey and Terza 2 standard errors. For the logistic TSRI with the Newey standard error, the z statistic was 4.71, whereas the probit TSRI gave a slightly larger z statistic of 4.74. The LSMM gave a larger point estimate of the causal odds ratio of 1.39 (95% CI: 1.19, 1.59) and also a larger standard error as shown by the smaller z statistic and wider confidence interval. We conclude that the observational estimate of the causal odds ratio has been attenuated by unmeasured confounding and that these data support a causal effect of BMI on the risk of diabetes.

Table 2 shows the estimates of the effect on SBP of a 1-unit increase in BMI. The direct estimate of this association was 0.76 mm Hg (95% CI: 0.70, 0.82). Using the same first stage as

Table 1. Estimates of the Causal Odds Ratios for Diabetes for a 1-Unit Increase in Body Mass Index Across 6 Cohorts[a] ($n = 17{,}057$)

Estimator	SE[b]	z	OR	95% CI
Direct logistic	0.004	29.6	1.14	1.13, 1.15
Logistic TSPS (stage 1: $F = 119$, $R^2 = 0.007$)	0.056	4.96	1.32	1.19, 1.48
Logistic TSRI (unadjusted SE)	0.058	4.79	1.32	1.18, 1.48
Logistic TSRI (robust SE)	0.057	4.86	1.32	1.18, 1.47
Logistic TSRI (TSPS unadjusted SE)	0.056	4.96	1.32	1.18, 1.47
Logistic TSRI (BS1)[c]	0.057	4.80	1.32	1.18, 1.48
Logistic TSRI (BS2)[c]	0.061	4.50	1.32	1.17, 1.49
Logistic TSRI (Newey)	0.059	4.71	1.32	1.17, 1.48
Logistic TSRI (Terza 1)	0.057	4.83	1.32	1.18, 1.47
Logistic TSRI (Terza 2)	0.059	4.77	1.32	1.18, 1.48
LSMM	0.101	3.26	1.39	1.19, 1.59
Probit TSRI (on OR scale)	0.090	4.74	1.28	1.15, 1.42

Abbreviations: BS1, bootstrap, second stage only; BS2, bootstrap, both stages; CI, confidence interval; LSMM, structural mean model; OR, odds ratio; SE, standard error; TSPS, 2-stage predictor substitution; TSRI, 2-stage residual inclusion.

[a] The 6 cohorts were Atherosclerosis Risk in Communities (ARIC) (58), the Cardiovascular Health Study (CHS) (59), Coronary Artery Risk Development in Young Adults (CARDIA) (60), the Framingham Heart Study (FHS) (61), Multinational Etoricoxib and Diclofenac Arthritis Long-term (MEDAL) (62), and the Multi-Ethnic Study of Atherosclerosis (MESA) (63).

[b] SEs given on log odds ratio scale.

[c] Bootstrapping using 500 replications.

the logistic TSPS and TSRI estimators, TSLS gave an estimate for a 1-unit increase of BMI of 0.36 mm Hg (95% CI: −0.37, 1.10) with a standard error of 0.374. The linear TSRI gave the same point estimate with a smaller unadjusted standard error of 0.372. The Newey standard error of 0.374 was equal to the TSLS standard error, and the Terza standard errors were slightly smaller at 0.370 and 0.372. In this example, the BS2 standard error was the largest at 0.384. The Newey correction increased the standard error by 0.5%. We conclude that the observational association is likely to be partly explained by unmeasured confounding and that the data do not support a causal effect of BMI on SBP.

Table 2. Estimates of the Causal Effect of a 1-Unit Increase in Body Mass Index on Systolic Blood Pressure (mm Hg) Across 6 Cohorts[a] ($n = 17{,}057$)

Estimator	SE	Estimate	95% CI
Direct linear	0.031	0.76	0.70, 0.82
TSLS (stage 1: $F = 119$, $R^2 = 0.007$)	0.374	0.36	−0.37, 1.10
TSPS (unadjusted SE)	0.378	0.36	−0.38, 1.11
Linear TSRI (unadjusted SE)	0.372	0.36	−0.37, 1.09
Linear TSRI (robust SE)	0.370	0.36	−0.36, 1.09
Linear TSRI (TSPS unadjusted SE)	0.378	0.36	−0.38, 1.11
Linear TSRI (BS1)[b]	0.376	0.36	−0.37, 1.10
Linear TSRI (BS2)[b]	0.384	0.36	−0.39, 1.12
Linear TSRI (Newey)	0.374	0.36	−0.37, 1.10
Linear TSRI (Terza 1)	0.370	0.36	−0.36, 1.09
Linear TSRI (Terza 2)	0.372	0.36	−0.37, 1.09

Abbreviations: BS1, bootstrap, second stage only; BS2, bootstrap, both stages; CI, confidence interval; SE, standard error; TSLS, 2-stage least squares; TSPS, 2-stage predictor substitution; TSRI, 2-stage residual inclusion.

[a] The 6 cohorts are Atherosclerosis Risk in Communities (ARIC) (58), the Cardiovascular Health Study (CHS) (59), Coronary Artery Risk Development in Young Adults (CARDIA) (60), the Framingham Heart Study (FHS) (61), Multinational Etoricoxib and Diclofenac Arthritis Long-term (MEDAL) (62), and the Multi-Ethnic Study of Atherosclerosis (MESA) (63).

[b] Bootstrapping using 500 replications.

In this example, the standard errors that do not take into account the uncertainty from both stages of estimation (unadjusted, robust, and BS1) are only slightly smaller than those that do (TSLS, Newey, Terza 1 and 2, BS2, LSMM, and probit) because of the combination of low first-stage R^2 and large sample size.

DISCUSSION

In the present analysis, we have adapted corrections to the standard errors of TSRI estimators, developed by Newey (37) and Terza (38), to linear and logistic TSRI estimators. The results of our simulations show that Newey, Terza, BS2, and corrected TSLS (for the linear case) standard errors have the best properties in terms of coverage and type I error.

The methods were illustrated in real-data examples investigating the effect of BMI on SBP and diabetes risk. In the examples, the Newey standard errors were 0.5% and 2% larger than the unadjusted standard errors for the linear and logistic TSRI estimators, respectively. In the supplementary material, we show that the corrections to the TSRI standard errors have the greatest effect when the unmeasured confounding is greater and when the outcome prevalence is higher (up to 50%, beyond which the effect decreases). In the binary-outcome example, the probit TSRI estimator gave a slightly larger z statistic than the logistic TSRI estimator. The standard error of the logistic TSRI estimator could be scaled to give the same z statistic. We do not prefer this approach because using the scaled standard error in equation 4 would not give the same value as sequential 2-step estimation.

Further work could investigate the application of Newey and Terza standard errors to TSRI estimators using other generalized linear models at the second stage. For example, Terza et al. (18) used a parametric Weibull model and an ordered logistic regression model in the second stage. And Tchetgen Tchetgen et al. (65) discussed TSRI estimators for survival models using an Aalen additive hazard model at the second stage. Our work has applicability beyond Mendelian randomization studies because TSRI estimators have been used in other areas: for example, using randomized treatment status in a clinical trial as an instrumental variable to correct for noncompliance and in health economics (18, 22, 66).

Newey's correction to the standard errors of the 2-step probit TSRI estimator relates to Murphy-Topel (67) standard errors in econometrics, which can be used for TSPS estimates. Murphy-Topel standard errors have been implemented in Stata (StataCorp LP) (68–71). It has been argued that researchers may want to fit the logistic TSPS estimator because it is consistent for the effect averaged over the population (72, 73), whereas it is less clear what effect is identified by the TSRI estimator (74). Also TSPS estimators have correct type I error under the null (17, 72, 75). However, because we have shown that using Newey, Terza, and BS2 standard errors for TSRI estimators also gives correct type I error under the null, we argue that TSRI estimators are attractive to researchers using noncollapsible models at the second stage. There is also scope to use TSRI estimates, with corrected standard errors, as part of the algorithms in the recently proposed Mendelian randomization–Egger and median estimators, which are robust to different proportions of invalid instruments (76, 77).

In conclusion, we recommend that researchers fitting TSRI estimators should not report unadjusted or heteroscedasticity-robust standard errors. Instead, they should report standard errors using the Newey or Terza corrections or from bootstrapping, including both stages of estimation.

ACKNOWLEDGMENTS

Author affiliations: Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom (Tom M. Palmer); Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom (Michael V. Holmes); Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Pennsylvania (Michael V. Holmes, Brendan J. Keating); Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania (Brendan J. Keating); Department of Health Sciences, University of Leicester, Leicester, United Kingdom (Nuala A. Sheehan); and Department of Genetics, University of Leicester, Leicester, United Kingdom (Nuala A. Sheehan).

We thank Prof. Joe Terza (Indiana University Purdue University, Indianapolis, Indiana) for prepublication access to his *Stata Journal* article.

Conflict of interest: none declared.

REFERENCES

1. Davey Smith G, Ebrahim S. "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease. *Int J Epidemiol*. 2003;32(1):1–22.
2. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res*. 2007;16(4):309–330.
3. Burgess S, Timpson NJ, Ebrahim S, et al. Mendelian randomization: where are we now and where are we going? *Int J Epidemiol*. 2015;44(2):379–388.
4. Burgess S, Butterworth AS, Thompson JR. Beyond Mendelian randomization: how to interpret evidence of shared genetic predictors. *J Clin Epidemiol*. 2016;69:208–216.
5. Vansteelandt S, Goetghebeur E. Causal inference with generalized structural mean models. *J R Stat Soc Series B Stat Methodol*. 2003;65(4):817–835.
6. Johnston KM, Gustafson P, Levy AR, et al. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Stat Med*. 2008;27(9):1539–1556.
7. Didelez V, Meng S, Sheehan NA. Assumptions of IV methods for observational epidemiology. *Stat Sci*. 2010;25(1):22–40.
8. Bowden J, Vansteelandt S. Mendelian randomization analysis of case-control data using structural mean models. *Stat Med*. 2011;30(6):678–694.
9. Palmer TM, Sterne JA, Harbord RM, et al. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. *Am J Epidemiol*. 2011;173(12):1392–1403.

10. Vansteelandt S, Bowden J, Babanezhad M, et al. On instrumental variables estimation of causal odds ratios. *Stat Sci*. 2011;26(3):403–422.
11. Burgess S, Thompson SG. Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. *Stat Med*. 2012;31(15): 1582–1600.
12. Harbord RM, Didelez V, Palmer TM, et al. Severity of bias of a simple estimator of the causal odds ratio in Mendelian randomization studies. *Stat Med*. 2013;32(7):1246–1258.
13. Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A, eds. *Health Services Research Methodology: A Focus on AIDS*. Washington, DC: US Public Health Service; 1989: 113–159.
14. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Commun Stat Theory Methods*. 1994;23(8):2379–2412.
15. Clarke PS, Windmeijer F. Identification of causal effects on binary outcomes using structural mean models. *Biostatistics*. 2010;11(4):756–770.
16. Burgess S, Granell R, Palmer TM, et al. Lack of identification in semiparametric instrumental variable models with binary outcomes. *Am J Epidemiol*. 2014;180(1):111–119.
17. Palmer TM, Thompson JR, Tobin MD, et al. Adjusting for bias and unmeasured confounding in the analysis of Mendelian randomization studies with binary responses. *Int J Epidemiol*. 2008;37(5):1161–1168.
18. Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *J Health Econ*. 2008;27(3):531–543.
19. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd ed. New York, NY: Chapman & Hall; 1989.
20. O'Malley AJ, Frank RG, Normand SL. Estimating cost-offsets of new medications: use of new antipsychotics and mental health costs for schizophrenia. *Stat Med*. 2011;30(16): 1971–1988.
21. Cai B, Small DS, Have TR. Two-stage instrumental variable methods for estimating the causal odds ratio: analysis of bias. *Stat Med*. 2011;30(15):1809–1824.
22. Garrido MM, Deb P, Burgess JF Jr, et al. Choosing models for health care cost analyses: issues of nonlinearity and endogeneity. *Health Serv Res*. 2012;47(6):2377–2397.
23. Boef AG, Dekkers OM, le Cessie S. Mendelian randomization studies: a review of the approaches used and the quality of reporting. *Int J Epidemiol*. 2015;44(2):496–511.
24. Benn M, Tybjærg-Hansen A, Stender S, et al. Low-density lipoprotein cholesterol and the risk of cancer: a Mendelian randomization study. *J Natl Cancer Inst*. 2011;103(6): 508–519.
25. Collin SM, Metcalfe C, Palmer TM, et al. The causal roles of vitamin B(12) and transcobalamin in prostate cancer: can Mendelian randomization analysis provide definitive answers? *Int J Mol Epidemiol Genet*. 2011;2(4):316–327.
26. De Silva NM, Freathy RM, Palmer TM, et al. Mendelian randomization studies do not support a role for raised circulating triglyceride levels influencing type 2 diabetes, glucose levels, or insulin resistance. *Diabetes*. 2011;60(3): 1008–1018.
27. Lawlor DA, Harbord RM, Tybjaerg-Hansen A, et al. Using genetic loci to understand the relationship between adiposity and psychological distress: a Mendelian randomization study in the Copenhagen General Population Study of 53221 adults. *J Intern Med*. 2011;269(5):525–537.
28. Islam M, Jafar TH, Wood AR, et al. Multiple genetic variants explain measurable variance in type 2 diabetes-related traits in pakistanis. *Diabetologia*. 2012;55(8):2193–2204.
29. Theodoratou E, Palmer T, Zgaga L, et al. Instrumental variable estimation of the causal effect of plasma 25-hydroxy-vitamin D on colorectal cancer risk: a Mendelian randomization analysis. *PLoS One*. 2012;7(6):e37662.
30. Lawlor DA, Nordestgaard BG, Benn M, et al. Exploring causal associations between alcohol and coronary heart disease risk factors: findings from a Mendelian randomization study in the Copenhagen General Population Study. *Eur Heart J*. 2013; 34(32):2519–2528.
31. Haring R, Teumer A, Völker U, et al. Mendelian randomization suggests non-causal associations of testosterone with cardiometabolic risk factors and mortality. *Andrology*. 2013;1(1):17–23.
32. Thrift AP, Shaheen NJ, Gammon MD, et al. Obesity and risk of esophageal adenocarcinoma and Barrett's esophagus: a Mendelian randomization study. *J Natl Cancer Inst*. 2014; 106(11):dju252.
33. Holmes MV, Asselbergs FW, Palmer TM, et al. Mendelian randomization of blood lipids for coronary heart disease. *Eur Heart J*. 2015;36(9):539–550.
34. Ye Z, Haycock PC, Gurdasani D, et al. The association between circulating lipoprotein(a) and type 2 diabetes: is it causal? *Diabetes*. 2014;63(1):332–342.
35. Davidson R, Mackinnon JG. *Estimation and Inference in Econometrics*. New York, New York: Oxford University Press; 1993.
36. Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press; 2002.
37. Newey WK. Efficient estimation of limited dependent variable models with endogenous explanatory variables. *J Econom*. 1987;36(3):231–250.
38. Terza JV. Simpler standard errors for two-stage optimization estimators. *Stata J*. 2016;16(2):368–385.
39. Garen J. The returns to schooling: a selectivity bias approach with a continuous choice variable. *Econometrica*. 1984;52(5): 1199–1218.
40. Heckman J, Robb R. Alternative methods for evaluating the impact of interventions: an overview. *J Econom*. 1985; 30(1–2):239–267.
41. Wooldridge JM. On two stage least squares estimation of the average treatment effect in a random coefficient model. *Econ Lett*. 1997;56(2):129–133.
42. Newey WK, Powell JL, Vella F. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*. 1999;67(3):565–603.
43. Blundell RW, Powell JL. Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont M, Hansen LP, Turnovsky SJ, eds. *Advances in Economics and Econometrics: Theory and Applications. 8th World Congress of the Econometric Society*. Cambridge, UK: University Press; 2003:312–357.
44. Dhrymes PJ. *Econometrics: Statistical Foundations and Applications*. New York, NY: Harper and Row; 1970.
45. Hausman JA. Specification tests in econometrics. *Econometrica*. 1978;46(6):1251–1271.
46. Durbin J. Errors in variables. *Rev Int Stat Inst*. 1954;22(1/3): 23–32.
47. Wu DM. Alternative tests of independence between stochastic regressors and disturbances: finite sample results. *Econometrica*. 1974;42(3):529–546.
48. Pagan A. Econometric issues in the analysis of regressions with generated regressors. *Int Econ Rev*. 1984;25(1):221–247.

49. Cameron AC, Trivedi PK. *Microeconometrics: Methods and Applications*. New York, NY: Cambridge University Press; 2005.
50. Smith RJ, Blundell RW. An exogeneity test for a simultaneous equation tobit model with an application to labor supply. *Econometrica*. 1986;54(3):679–685.
51. Rivers D, Vuong QH. Limited information estimators and exogeneity tests for simultaneous probit models. *J Econom*. 1988;39(3):347–366.
52. Blundell RW, Smith RJ. Estimation in a class of simultaneous equation limited dependent variable models. *Rev Econ Stat*. 1989;56(1):37–57.
53. Stata Corp LP. *Stata Base Reference Manual Release 13*. College Station, TX: Stata Press; 2013:910–922.
54. Stata Corp. *Stata Statistical Software, Version 14.1*. College Station, TX; Stata Press; 2015.
55. R Core Team. *R: A Language and Environment for Statistical Computing, Version 3.2.1*. Vienna, Austria: R Foundation for Statistical Computing; 2015.
56. Clarke PS, Palmer TM, Windmeijer F. Estimating structural mean models with multiple instrumental variables using the generalised method of moments. *Stat Sci*. 2015;30(1):96–117.
57. Keating BJ, Tischfield S, Murray SS, et al. Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PLoS One*. 2008;3(10):e3583.
58. The Atherosclerosis Risk in Communities (ARIC) study: design and objectives. The ARIC investigators. *Am J Epidemiol*. 1989;129(4):687–702.
59. Fried LP, Borhani NO, Enright P, et al. The cardiovascular health study: design and rationale. *Ann Epidemiol*. 1991;1(3): 263–276.
60. Friedman GD, Cutter GR, Donahue RP, et al. CARDIA: study design, recruitment, and some characteristics of the examined subjects. *J Clin Epidemiol*. 1988;41(11):1105–1116.
61. Feinleib M, Kannel WB, Garrison RJ, et al. The Framingham Offspring Study: design and preliminary data. *Prev Med*. 1975; 4(4):518–525.
62. Cannon CP, Curtis SP, FitzGerald GA, et al. Cardiovascular outcomes with etoricoxib and diclofenac in patients with osteoarthritis and rheumatoid arthritis in the Multinational Etoricoxib and Diclofenac Arthritis Long-term (MEDAL) programme: a randomised comparison. *Lancet*. 2006; 368(9549):1771–1781.
63. Bild DE, Bluemke DA, Burke GL, et al. Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am J Epidemiol*. 2002; 156(9):871–881.
64. Holmes MV, Lange LA, Palmer T, et al. Causal effects of body mass index on cardiometabolic traits and events: a Mendelian randomization analysis. *Am J Hum Genet*. 2014;94(2): 198–208.
65. Tchetgen Tchetgen EJ, Walter S, Vansteelandt S, et al. Instrumental variable estimation in a survival context. *Epidemiology*. 2015;26(3):402–410.
66. Nagelkerke N, Fidler V, Bernsen R, et al. Estimating treatment effects in randomized clinical trials in the presence of non-compliance [published erratum appears in *Stat Med*. 2001;20(6):982]. *Stat Med*. 2000;19(14): 1849–1864.
67. Murphy KM, Topel RH. Estimation and inference in two-step econometric models. *J Bus Econ Stat*. 1985;3(4):370–379.
68. Hardin JW. The robust variance estimator for two-stage models. *Stata J*. 2002;2(3):253–265.
69. Hole AR. Calculating Murphy-Topel variance estimates in Stata: a simplified procedure. *Stata J*. 2006;6(4):521–529.
70. Hardin JW, Carroll RJ. Variance estimation for the instrumental variables approach to measurement error in generalized linear models. *Stata J*. 2003;3(4):342–350.
71. Hardin JW, Schmiediche H, Carroll RJ. Instrumental variables, bootstrapping, and generalized linear models. *Stata J*. 2003; 3(4):351–360.
72. Burgess S; CRP CHD Genetics Collaboration. Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. *Stat Med*. 2013; 32(27):4726–4747.
73. Burgess S, Thompson SG. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. London, UK: Chapman and Hall/CRC; 2015.
74. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization [published online ahead of print August 17, 2015]. *Stat Methods Med Res*. (doi:10.1177/0962280215597579).
75. Burgess S. Commentary: consistency and collapsibility: are they crucial for instrumental variable analysis with a survival outcome in Mendelian randomization? *Epidemiology*. 2015; 26(3):411–413.
76. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*. 2015; 44(2):512–525.
77. Bowden J, Davey Smith G, Haycock PC, et al. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol*. 2016;40(4):304–314.