Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2008;37:1161–1168
doi:10.1093/ije/dyn080

Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses

Tom M Palmer,[1]* John R Thompson,[1] Martin D Tobin,[2] Nuala A Sheehan[2] and Paul R Burton[2]

Accepted 3 April 2008

Background Mendelian randomization uses a carefully selected gene as an instrumental-variable (IV) to test or estimate an association between a phenotype and a disease. Classical IV analysis assumes linear relationships between the variables, but disease status is often binary and modelled by a logistic regression. When the linearity assumption between the variables does not hold the IV estimates will be biased. The extent of this bias in the phenotype-disease log odds ratio of a Mendelian randomization study is investigated.

Methods Three estimators termed direct, standard IV and adjusted IV, of the phenotype-disease log odds ratio are compared through a simulation study which incorporates unmeasured confounding. The simulations are verified using formulae relating marginal and conditional estimates given in the Appendix.

Results The simulations show that the direct estimator is biased by unmeasured confounding factors and the standard IV estimator is attenuated towards the null. Under most circumstances the adjusted IV estimator has the smallest bias, although it has inflated type I error when the unmeasured confounders have a large effect.

Conclusions In a Mendelian randomization study with a binary disease outcome the bias associated with estimating the phenotype-disease log odds ratio may be of practical importance and so estimates should be subject to a sensitivity analysis against different amounts of hypothesized confounding.

Keywords Instrumental-variable analysis, Mendelian randomization, bias, unobserved confounding

Introduction

In traditional epidemiological studies the associations between biological phenotypes and diseases can be distorted by confounding or reverse causation. The aim of Mendelian randomization analysis is to test or estimate the association between a biological phenotype and a disease in the presence of unmeasured confounding.[1–3] This is achieved using a carefully selected gene as an instrumental-variable (IV).[4–7] When certain assumptions hold Mendelian randomization will remove the distorting effects and produce unconfounded estimates of the association between a phenotype and a disease.[3,8] Genes that influence the disease through their effect on the biological phenotype of interest can be used as instrumental-variables in the analysis because a subject's genotype is essentially

[1] Department of Health Sciences, University of Leicester, UK.
[2] Departments of Health Sciences and Genetics, University of Leicester, UK.
* Corresponding author. University of Leicester, Department of Health Sciences, 2nd Floor, Adrian Building, University Road, Leicester LE1 7RH, UK. E-mail: tmp8@le.ac.uk

randomly assigned before birth and thus should not be influenced by the many environmental and lifestyle factors that typically act as confounders in epidemiology.[9]

In this article, we show that, for binary outcomes, the observed bias towards the null in Mendelian randomization estimates is due to the impact of random effects that are not explicitly included in the linear predictor. This is analogous to the discrepancy between marginal and conditional parameter estimates in generalized linear mixed models with a logistic link.[10,11] Theoretical formulae for approximating this difference are provided for each of three different estimators and their accuracy is verified by simulation. In theory, knowledge of the difference between marginal and conditional estimates could provide a correction for the bias that pertains in Mendelian randomization analyses. However, the extent of this bias depends on the properties of the unmeasured confounders, which are always unknown. An adjusted instrumental-variable estimator is applied to Mendelian randomization analyses to produce an improved estimate of the phenotype-disease association. The adjusted IV estimator partially compensates for the unknown confounders by exploiting information from the residuals of the regression of the intermediate phenotype on the genotype.

Methods

Estimators for Mendelian randomization studies with binary responses

The key variables in describing the Mendelian randomization model are; the disease status (Y), intermediate phenotype (X), genotype (G) and confounder (U). The assumed relationship between these variables is shown in Figure 1. For the ith subject in a cohort, let y_i represent their binary disease status, p_i represent their probability of having the disease, x_i represent the level of the biological phenotype and g_i represent their genotype, which is coded 0, 1 and 2 to indicate the number of copies of the relevant risk allele. Typically there will be many unmeasured confounders, so it is assumed that they can be represented by a single variable, u_i, that captures their combined effect. This confounding variable is arbitrarily assumed to be standardized to have a mean of zero and a standard deviation of one. For simplicity, we assume an additive effect of genotype on the intermediate phenotype, although the argument would apply equally to any known mode of inheritance. It is also assumed that the confounder acts additively in the linear predictors of the associations between the genotype and phenotype and between the phenotype and the disease.

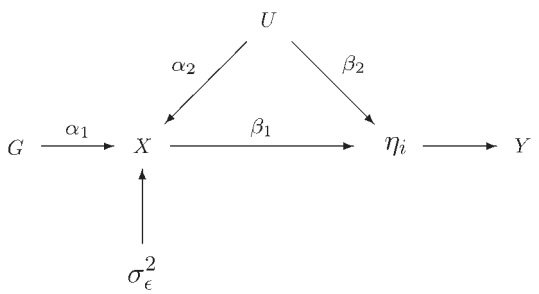

Figure 1 The relationship between the variables (η_i is the linear predictor of the logistic regression)

The coefficients in the regression of phenotype on genotype are denoted by α's so that,

$$x_i = \alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \epsilon_i, \text{ with } \epsilon_i \sim N(0, \sigma_\epsilon^2), \quad (1)$$

and ϵ represents the effects of measurement error and unmeasured factors that are not confounders because they do not influence disease. The coefficients in the linear predictor between phenotype and disease are denoted by β's, so that the disease status follows a Bernoulli distribution,

$$y_i \sim \text{Bern}(p_i), \text{ with } \log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i + \beta_2 u_i. \quad (2)$$

Implicit in the notation is the idea that ϵ_i and u_i are independent of one another. The primary interest in this paper is to recover β_1.

If both regressions were linear, ignoring the confounder in the instrumental-variable analysis would not bias the estimate of β_1, but this is not the case for a non-linear relationship between phenotype and disease.[12] Substituting the formula for x_i in Equation (1) into the logistic regression in Equation (2) gives,

$$\log\frac{p_i}{1-p_i} = \beta_0 + \beta_1(\alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \epsilon_i) + \beta_2 u_i. \quad (3)$$

The coefficient of g_i in this relationship is $\beta_1\alpha_1$ while the coefficient of g_i in the linear regression in Equation (1) is α_1. In principle the ratio of the estimates of these coefficients should give an estimate of β_1,[4] which is the effect of the phenotype on disease risk after adjusting for confounding. Unfortunately u_i and ϵ_i are unknown, so the estimate of $\beta_1\alpha_1$ is taken from the logistic regression without those terms, thus in effect replacing the true conditional model with a marginal model which averages over the unknown terms, u_i and ϵ_i.

An alternative to the ratio estimate of β_1 is obtained by taking the predicted values of the intermediate phenotype from the first regression ignoring the confounding,

$$\hat{x}_i = \hat{\alpha}_0 + \hat{\alpha}_1 g_i \approx \alpha_0 + \alpha_1 g_i \quad (4)$$

and substituting those into the logistic regression in Equation (2), in which case,

$$\log\frac{p_i}{1-p_i} \approx \beta_0 + \beta_1(\hat{x}_i + \alpha_2 u_i + \epsilon_i) + \beta_2 u_i. \quad (5)$$

In this two-stage approach, the estimate of interest is just the coefficient of the predicted phenotype $\hat{x}_i$,

but the biases will be similar to those that occur for the ratio estimator.

In an attempt to correct for this difference between marginal and conditional parameter estimates, and thus improve upon the standard instrumental-variable estimator an adjusted IV estimator is applied. The estimated residuals from the first stage linear regression in Equation (1) are,

$$r_i = x_i - \hat{x}_i. \tag{6}$$

These estimated residuals capture some of the variability contained in the unknown confounders and the phenotype error term, ϵ. This information can be used in the second regression by fitting,

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \hat{x}_i + \beta_r r_i. \tag{7}$$

The information about the confounding contained in the residuals should, in part, compensate for the missing terms in the marginal form of the logistic regression model and therefore reduce the difference between the conditional and marginal estimates of β_1.

This article considers three estimators of β_1. First, the direct estimator, that does not use Mendelian randomization but performs a logistic regression of disease status on the intermediate as in a traditional epidemiological study. The direct estimator of β_1 is derived from the linear predictor,

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i. \tag{8}$$

The standard IV estimator uses Mendelian randomization so that the linear predictor is,

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \hat{x}_i. \tag{9}$$

The third estimator is the adjusted IV estimator obtained from Equation (7). In the Appendix, formulae are given for calculating the size of the bias in β_1 under the three estimators.

Data simulation

A simulation study was performed to validate the formulae for the three estimators. In a cohort of size 10 000, subjects were each randomly assigned two alleles in Hardy-Weinberg equilibrium with the allele frequency of the risk allele set to 30%. The confounding variable was simulated to be normally distributed with mean zero and variance equal to one, $u_i \sim N(0,1)$. The phenotype, x_i, was generated as a Normal random variable with mean equal to, $\alpha_0 + \alpha_1 g_i + \alpha_2 u_i$ following Equation (1), and the standard deviation of the phenotype error term, σ_ϵ, was set to one. Each subject's probability of disease was simulated, following Equation (2) such that $\log p_i/(1-p_i) = \beta_0 + \beta_1 y_i + \beta_2 u_i$.

The baseline prevalence of disease was set to 5% by fixing β_0. Different amounts of confounding were considered by changing the values of α_2 and β_2. In particular, four confounding scenarios were considered by setting the confounding effect on the phenotype, α_2, to 0, 1, 2 and 3 whilst the confounding effect on the disease, β_2, was varied between zero and three for each scenario. The other parameters were fixed as follows; $\alpha_0 = 0$, $\alpha_1 = 1$ and $\beta_1 = 1$. For each set of parameter values 10 000 simulations were performed. Statistical analysis was performed using R (version 2.6.1).[13]

Results

The three estimators are assessed using the median parameter estimates, coverage probabilities and type I errors of the phenotype-disease log odds ratio, β_1. The coverage probability of β_1 was calculated as the proportion of simulations whose confidence interval included the true value of β_1. A set of simulations was performed with β_1 equal to 0 to represent the situation in which there is no association between phenotype and disease. For those simulations, the proportion of statistically significant estimates of β_1 is an estimate of the type I error of the Wald test of β_1.

Assessment of the bias of the estimators

Figure 2 shows the median of β_1 for the three estimators from the simulations, represented by the symbols, and the values of the estimators calculated from the formulae given in the Appendix represented by the lines.

Figure 2 shows that the median values from the simulations are in close agreement with the theoretical predictions, there is the same pattern to the estimates of β_1 for the different values of α_2 except when α_2 is equal to zero. When α_2 is equal to zero the direct and adjusted estimators are equivalent due to the assumptions underlying the relationship between the confounder and the phenotype. When α_2 is non-zero, allowing the confounder to take effect, the direct estimate of β_1 is greater than the set value of one. However, the effect the unmeasured confounding has on the standard IV estimates is to bias them towards zero, producing estimates that are always below the true value of one. The values of the adjusted IV estimator are between the other two sets of estimates and have the smallest bias of the three estimators. For the adjusted IV estimates the bias in β_1 reduces with largest values of α_2 because the estimated residuals are more informative.

Assessment of the coverage probabilities of the estimators

Figure 3 shows the coverage probabilities of the three estimators, when the nominal level was 95%. The direct estimator and the standard IV estimator demonstrate very low coverage for all four scenarios due to the bias in β_1. The adjusted IV estimator

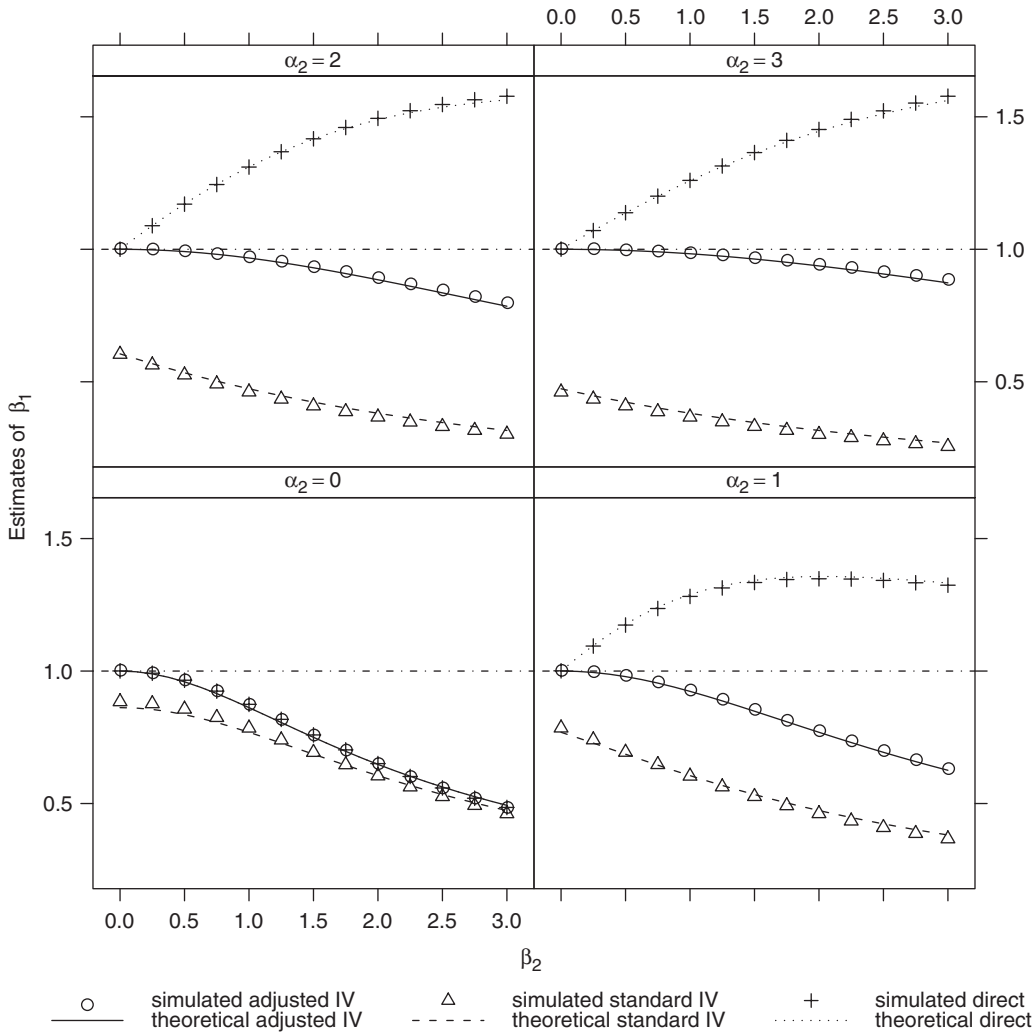

Figure 2 Simulated and theoretical values of β_1

demonstrates the best coverage properties with levels around 95% over the range of values of β_2 for which its estimate of β_1 was approximately equal to the set value of one in Figure 2.

Assessment of type I error

Figure 4 shows the type I error of the standard IV and adjusted IV estimators when the nominal rate is 5%. The type I error of the direct estimator is not shown on Figure 4 because the values were very large. Under the three scenarios with non-zero values of α_2 the adjusted IV estimator has a substantially higher type I error rate than the standard IV estimator because the inclusion of the estimated residuals in the adjusted IV estimator reduced its estimated standard error.

Discussion

This article considers the bias in the estimates from Mendelian randomization studies with binary outcomes. Three estimators of the phenotype-disease log odds ratio, termed; direct, standard IV and adjusted IV, have been evaluated through a simulation study. The simulations are in agreement with formulae relating conditional and marginal parameter estimates from logistic regression given in the Appendix. The adjusted IV estimator was the least biased, but it had high type I error when the effect of the unmeasured confounder was large. Further, unreported simulations show that the difference between marginal and conditional parameter estimates would also exist with probit regression and hence a similar but not identical adjustment between the conditional and marginal estimates of β_1 would be required if probit regressions were used in place of logistic regressions for the three estimators.[10]

The simulations investigated the performance of the estimators over a range of values of the confounder. Over the four panels in Figure 2, when $\alpha_2 = 0$, 1, 2 and 3, the confounder accounted for approximately 0%, 45%, 80% and 90% of the phenotype variance. For the log odds of disease the confounder accounted for between 0% and 90% of the variance in the linear

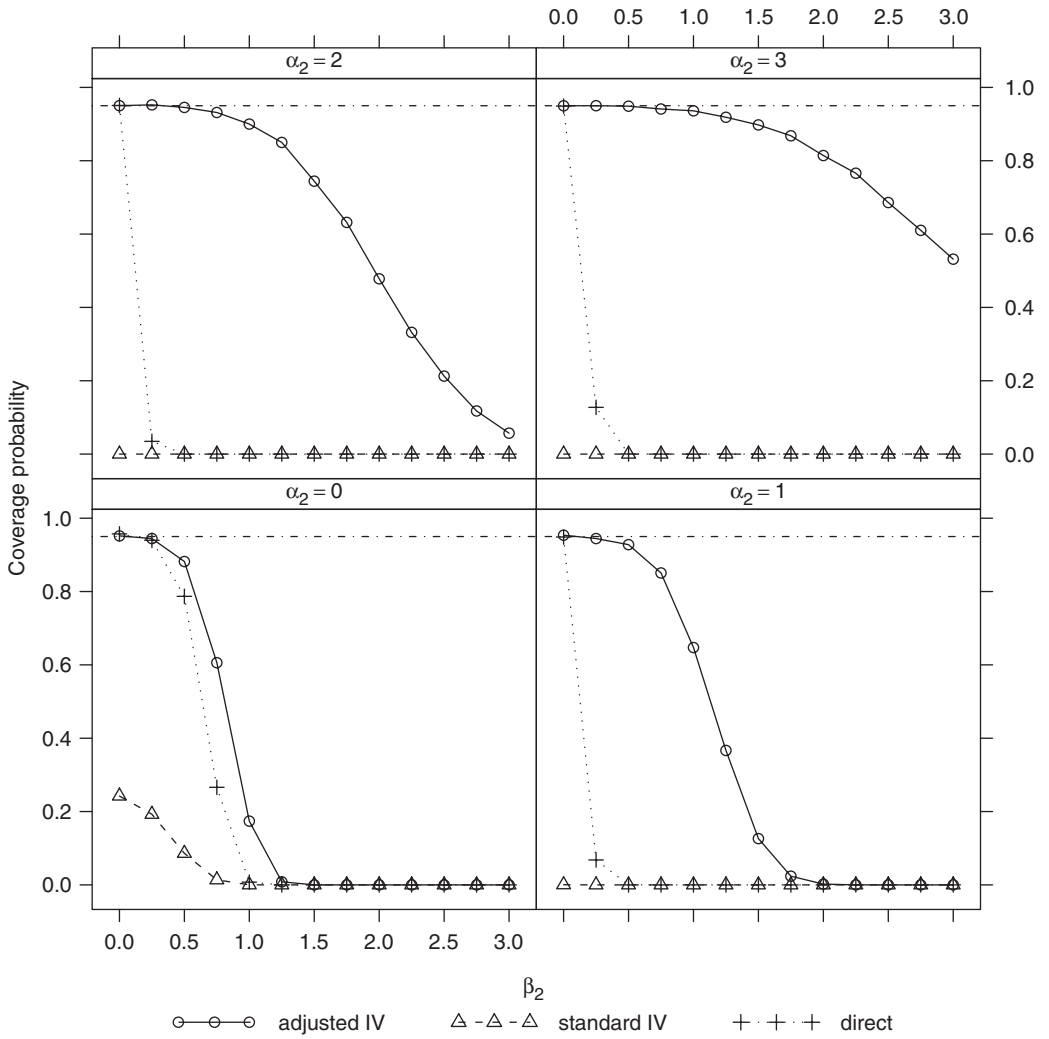

Figure 3 Coverage probabilities of the three estimators

predictor when $\alpha_2 = 0$ and β_2 varied from 0 to 3, between 45% and 90% when $\alpha_2 = 1$, between 80% and 90% when $\alpha_2 = 2$ and between 85% and 95% when $\alpha_2 = 3$. Typically the gene used in a Mendelian randomization study will only explain a small percentage of the variance in the phenotype, perhaps <10%. The impact of the confounders can therefore be large causing large bias. If it is possible to include measured confounders in the analysis this will reduce the importance of the unmeasured confounders and so reduce the bias in all of the estimators.

The adjusted IV estimator uses the estimated residuals as well as the predicted values from the first stage regression of the genotype on the phenotype as covariates in the second stage logistic regression between the phenotype and the disease outcome. A similar adjusted IV estimator was introduced in the context of clinical trials subject to non-compliance.[14] The first stage residuals contain some information about the unmeasured confounder since they capture the variance in the phenotype that is not explained by the genotype. The argument used in the clinical trials context was that these first stage residuals meet Pearl's back-door criterion and their inclusion in the model results in the adjusted IV estimate having a causal interpretation.[14]

Point estimates of causal effects from instrumental variable analyses require strong parametric and distributional assumptions, e.g. all relationships are linear without interactions.[6,15] Although the relationship between a gene and an intermediate phenotype might well be approximated by a linear regression, the final response variable in epidemiological studies is often a binary indicator of disease status and so the phenotype-disease relationship is typically non-linear. Instrumental variable theory has not been fully generalized to non-linear situations[6] so the practical implications of such a violation of the core assumptions have not yet been clearly defined. Most crucially, both the specification of the relevant causal parameter and identification of how it relates to what can be estimated in the observational regime are not

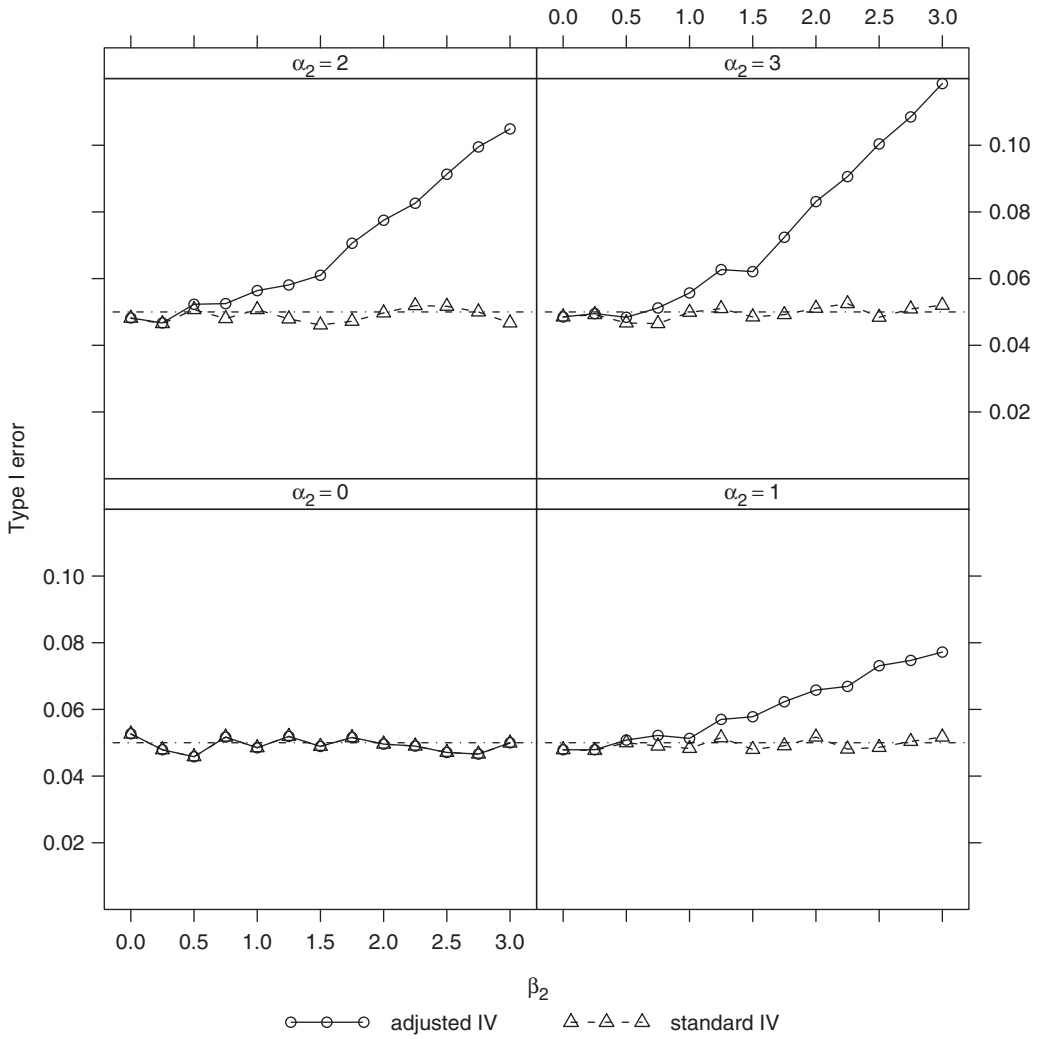

Figure 4 Type I error rate of the Wald test for the three estimators of β_1

generally straightforward.[12] There are many examples where causal estimates have been obtained for binary outcomes but the particular parameter that can be estimated depends on the situation being considered and the assumptions that can be made.[16–22] Whilst, this is an important issue, our focus here is simply on improving the estimates of the parameter for the effect of phenotype on disease in the relevant logistic regression equation when contemporary Mendelian randomization methods are applied to binary outcome data. For now, we ignore the issue of whether, and under what conditions, this parameter has a strictly causal interpretation.

The bias associated with binary outcomes in a Mendelian randomization study may be of practical importance, so more detailed sensitivity analyses should be performed in which the biasing effects of hypothesized amounts of confounding are investigated using the formulae given in the Appendix. The three estimators considered here give different values of the phenotype-disease log odds ratio under different scenarios of confounding. The differences between the estimates are greater when the effects of the unmeasured confounders are larger. There are now several published examples of Mendelian randomization analyses, and the collection of genotype, phenotype and disease status information is becoming increasingly common, especially with the creation of large-scale Biobanks such as the UK Biobank. Large-scale collaborative genetic epidemiological studies[23,24] will ensure that there will be many genes available for use as instrumental variables in future Mendelian randomization analyses.

Acknowledgements

TMP is funded by a Medical Research Council Capacity Building studentship in Genetic Epidemiology (G0501386). MDT is funded by a Medical Research

Council Clinician Scientist Fellowship (G0501942). The methodological research programme in Genetic Epidemiology at the University of Leicester forms one part of broader research programmes supported by: an MRC Program Grant (G0601625) addressing causal inference in Mendelian randomization; PHOEBE (Promoting Harmonization Of Epidemiological Biobanks in Europe) funded by the European Commission under Framework 6 (LSHG-CT-2006-518418); P^3G (Public Population Project in Genomics) funded under an International Consortium Initiative from Genome Canada and Genome Quebec; and an MRC Cooperative Grant (G9806740). The simulation study was performed using the University of Leicester Mathematical Modelling Centre's supercomputer which was purchased through the HEFCE Science Research Investment Fund. The authors would like to thank three anonymous referees whose comments helped improve the article.

References

[1] Katan MB. Apolipoprotein e isoforms, serum cholesterol, and cancer. *Lancet* 1986;**327:**507–8.

[2] Davey Smith G, Ebrahim S. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease. *Int J Epidemiol* 2003;**32:**1–22.

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

[4] Thomas DC, Conti DV. Commentary: The concept of 'mendelian randomization'. *Int J Epidemiol* 2004;**33:**21–25.

[5] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc* 1996;**91:**444–55.

[6] Pearl J. *Causality*. Cambridge: Cambridge University Press, 2000.

[7] Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000;**29:**722–29.

[8] Tobin MD, Minelli C, Burton PR, Thompson JR. Commentary: Development of mendelian randomization: from hypothesis test to 'mendelian deconfounding'. *Int J Epidemiol* 2004;**33:**26–29.

[9] Davey Smith G, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR. Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* 2005;**366:**1484–98.

[10] Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988;**44:**1049–60.

[11] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993;**88:**9–25.

[12] Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res* 2007;**16:**309–330.

[13] R Development Core Team. *R*: *A Language and Environment for Statistical Computing*. Vienna, Austria. R Foundation for Statistical Computing, 2007. ISBN 3-900051-07-0.

[14] Nagelkerke N, Fidler V, Bernsen R, Borgdorff M. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Stat Med* 2000;**19:**1849–64, [Erratum, Stat Med 2001;**20**:982].

[15] Bowden RJ, Turkington DA. *Instrumental Variables*. Cambridge: Cambridge University Press, 1984.

[16] Amemiya T. The nonlinear two-stage least-squares estimator. *J Econom* 1974;**2:**105–10.

[17] Hansen LP, Singleton RJ. Generalized instrumental variable estimation of non-linear rational expectation models. *Econometrica* 1982;**50:**1269–86.

[18] Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999;**14:**29–46.

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

[20] Nitsch D, Molokhia M, Smeeth L, DeStavola BL, Whittaker JC, Leon DA. Limits to causal inference based on mendelian randomization: a comparison with randomized controlled trials. *Am J Epidemiol* 2006;**163:**397–403.

[21] Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology* 2006;**17:**260–67.

[22] Hernán MA, Robins JM. Instruments for causal inference. An epidemiologist's dream? *Epidemiology* 2006;**17:**360–72.

[23] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3 000 shared controls. *Nature* 2007;**447**:661–78.

[24] The GAIN Collaborative Research Group. New models of collaboration in genome-wide association studies: the genetic association information network. *Nat Genet* 2007; **39:**1045–51.

[25] Hardin JW, Hilbe JM. *Generalized Estimating Equations*. Boca Raton, US: Chapman and Hall/CRC, 2003.

[26] Thomas DC, Lawlor DA, Thompson JR. Re: Estimation of Bias in Nongenetic Observational Studies Using Mendelian Triangulation by Bautista *et al*. *Ann Epidemiol* 2007;**17:** 511–13.

[27] Anderson TW. *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1958.

Appendix

Formulae for the difference between the marginal and conditional parameter estimates of the three estimators

The difference between marginal and conditional parameter estimates has been investigated for the case of linear, logistic, probit and Poisson regression models.[10,25] In the case of logistic regression this difference can be expressed by a multiplicative factor,

$$\beta_{\text{marg}} \approx \beta_{\text{cond}} \cdot \frac{1}{\sqrt{1 + c^2 V}}, \text{ where } c = \frac{16\sqrt{3}}{15\pi}. \qquad (10)$$

where β_{marg} and β_{cond} are the marginal and conditional parameter estimates and V is the variance of the

covariates over which the marginal estimates are averaged. The formulae for the three estimators are derived by approximating the logistic regression as a simple regression of the log odds ratio, $\theta = \log(p/(1-p))$ on the covariates and confounders.[26] If the terms included in the linear predictor of the logistic regression are denoted by Z then the remaining variance after allowing for these terms will be given by,

$$V = \mathrm{var}(\theta|Z) = \mathrm{var}(\theta) - \frac{\mathrm{cov}(\theta, Z)^2}{\mathrm{var}(Z)} \tag{11}$$

since θ and Z can both be assumed to be normally distributed.[27] From Equation (3),

$$\theta_i = \beta_0 + \beta_1\alpha_0 + \beta_1\alpha_1 g_i + (\beta_1\alpha_2 + \beta_2)u_i + \beta_1\epsilon_i \tag{12}$$

and because u is standardized, it follows that

$$\mathrm{var}(\theta) = (\beta_1\alpha_1)^2\mathrm{var}(g) + (\beta_1\alpha_2 + \beta_2)^2 + \beta_1^2\sigma_\epsilon^2 \tag{13}$$

and we can approximate $\mathrm{var}(g)$ by $2q(1-q)$ where q is the minor allele frequency. Hence to apply Equation (10) it is necessary to derive V for each of the three estimators.

The direct estimator

The direct estimator performs a logistic regression of disease on the intermediate phenotype. In this case $Z = x_i$ where,

$$x_i = \alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \epsilon_i \tag{14}$$

so,

$$\mathrm{var}(Z) = \alpha_1^2\mathrm{var}(g) + \alpha_2^2 + \sigma_\epsilon^2. \tag{15}$$

The covariance between the log odds and the terms in the linear predictor is given by

$$\mathrm{cov}(\theta, Z) = \begin{bmatrix}\alpha_1 & \alpha_2 & 1\end{bmatrix} \cdot \begin{bmatrix}\mathrm{var}(g) & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & \sigma_\epsilon^2\end{bmatrix} \cdot \begin{bmatrix}\beta_1\alpha_1\\ \beta_1\alpha_2 + \beta_2\\ \beta_1\end{bmatrix}$$
$$= \alpha_1^2\beta_1\mathrm{var}(g) + \alpha_2(\beta_1\alpha_2 + \beta_2) + \beta_1\sigma_\epsilon^2. \tag{16}$$

Hence V_{direct} can be formed using Equations (13), (16) and (15).

The standard IV estimator

For the standard IV estimator the log odds are regressed on the fitted values from the linear regression of the phenotype on the genotype. Thus $Z \approx \alpha_0 + \alpha_1 g$ and,

$$\mathrm{var}(Z) = \alpha_1^2\mathrm{var}(g), \tag{17}$$

$$\mathrm{cov}(\theta, Z) = \alpha_1^2\beta_1\mathrm{var}(g). \tag{18}$$

Hence for the standard IV estimator V is given by,

$$V_{\mathrm{standard}} = (\beta_1\alpha_2 + \beta_2)^2 + \beta_1^2\sigma_\epsilon^2. \tag{19}$$

The adjusted IV estimator

The adjusted IV estimator makes use of the estimated residuals, r, from the regression of the phenotype on genotype to capture some of the variance explained by confounding variables not included in the standard IV estimator. Therefore the value of V is reduced compared with the standard IV estimator. For the adjusted IV estimator V is given by,

$$V = \mathrm{var}(\theta|Z) - \frac{\mathrm{cov}(\theta|Z, r)^2}{\mathrm{var}(r)}. \tag{20}$$

If the confounder u is standardized the estimated residuals and their variance are given by,

$$r_i = \alpha_2 u_i + \epsilon_i \tag{21}$$

$$\mathrm{var}(r_i) = \alpha_2^2 + \sigma_\epsilon^2 \tag{22}$$

The covariance between the log odds given the phenotype information and the estimated residuals is given by,

$$\mathrm{cov}(\theta|Z, r) = \begin{bmatrix}\beta_1\alpha_2 + \beta_2 & \beta_1\end{bmatrix} \cdot \begin{bmatrix}1 & 0\\ 0 & \sigma_\epsilon^2\end{bmatrix} \cdot \begin{bmatrix}\alpha_2\\ 1\end{bmatrix} \tag{23}$$
$$= \alpha_2(\beta_1\alpha_2 + \beta_2) + \beta_1\sigma_\epsilon^2. \tag{24}$$

Since $\mathrm{var}(\theta|Z) = V_{\mathrm{standard}}$ from the standard IV estimator above, for the adjusted IV estimator we have,

$$V_{\mathrm{adjusted}} = (\beta_1\alpha_2 + \beta_2)^2 + \beta_1^2\sigma_\epsilon^2 - \frac{(\alpha_2(\beta_1\alpha_2 + \beta_2) + \beta_1\sigma_\epsilon^2)^2}{\alpha_2^2 + \sigma_\epsilon^2}. \tag{25}$$