

# IV-estimators of the causal odds ratio for a continuous exposure in prospective and retrospective designs

JACK BOWDEN (corresponding author)

*MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge  
CB2 0SR. U.K.*

*jack.bowden@mrc-bsu.cam.ac.uk*

*FAX: 01223 330388*

STIJN VANSTEELANDT

*Department of Applied Mathematics and Computer Science, Ghent University,  
B-9000 Gent, BELGIUM.*

*Stijn.Vansteelandt@ugent.be*

## Summary

We explore the task of estimating the marginal or population averaged causal effect of an exposure, when expressed as an odds ratio, in the presence of an instrumental variable (IV). Our motivation is to provide a framework for the analysis of non-randomised epidemiological data to reflect what would have been estimated, had a randomised controlled trial in fact been possible. Specifically, we investigate two recently proposed methods - the 'Adjusted' IV estimator of Palmer et al. (2008) and the G-estimator of Vansteelandt and Goetghebeur (2003) - which are shown in a particular scenario to estimate the same theoretical quantity. General weighted estimating equations are suggested as a means to implement both methods, which can then be tailored to deal with either prospective or retrospective data. We evaluate their performance in practice in terms of bias, efficiency and sensitivity to model misspecification.

Key Words: Causal odds ratio, Adjusted IV estimator, Logistic Structural Mean Model, prospective and retrospective data.

# 1 Introduction

Ascertaining whether or not an exposure is causally related to the risk of a disease is hugely important since, if this is indeed the case, resources can then be targeted at the exposure, safe in the knowledge that the population’s health will benefit. Ethical reasons often forbid randomised trials and consequently health professionals are forced to decide on public health interventions with evidence obtained from non-randomised experimental data. However, evidence of an association gleaned from such data does not guarantee that subsequent health interventions will have an impact, because ‘association does not equal causation’.

In this paper we focus our attention on estimating the causal effect of a continuous exposure on a binary outcome. In order to motivate the problem let  $X$  - the exposure of interest - and  $U$  - an unmeasured confounder - determine the probability of disease  $Y$  through the logistic model

$$\text{logit}(\Pr(Y = 1|X = x, U = u)) = \beta_0 + \beta_1 x + \beta_2 u \quad (1.1)$$

where  $\text{logit}(p) = \log(p/1 - p)$ . Figure 1 (left) shows a causal diagram (Pearl, 1995) of the variables  $X, Y$  and  $U$  - the direction of the arrows indicate that  $U$  may affect both the exposure  $X$  and the outcome  $Y$  but not the other way round, so that  $U$  is effectively a ‘confounder’. We assume that  $U$  is the only common cause of  $X$  and  $Y$ . For now, following the interpretation of Lin et al. (1998), we think of  $U$  as a composite, normally distributed, standardised score constructed from several continuous unmeasured confounders so that  $U \sim N(0, 1)$ . Using the  $do()$  operator notation of Pearl (1995) the log odds of disease for an individual with confounder level  $U = u$  would be

$$\text{logit}(\Pr(Y = 1|do(X = x_0), U = u)) = \beta_0 + \beta_1 x_0 + \beta_2 u \quad (1.2)$$

if allocated to treatment level  $x_0$ . Since treatment allocation is independent of the confounding variable  $U$ , this effectively removes the causal arrow from  $U$  to  $X$ , as illustrated in Figure 1 (right), (Pearl, 1995). It follows that

$$\beta_1 = \log \left\{ \frac{\text{odds } \Pr(Y = 1|do(X = x_0 + 1), U)}{\text{odds } \Pr(Y = 1|do(X = x_0), U)} \right\}$$

encodes a conditional ‘causal’ log odds ratio of disease corresponding to a unit

increase in the exposure. Despite its causal effect interpretation, the magnitude of  $\exp(\beta_1)$  is of limited use for policy making. This is because it represents the change in disease odds that would be realized *in subsets of  $U$*  if the exposure were increased with a unit, which is difficult to interpret because  $U$  is unknown. In particular,  $\exp(\beta_1)$  does not reflect the marginal ‘causal’ log odds ratio (CLOR)

$$\begin{aligned} CLOR(x_0, x_0 + 1) &= \log \left\{ \frac{\text{odds } Pr(Y = 1 | do(X = x_0 + 1))}{\text{odds } Pr(Y = 1 | do(X = x_0))} \right\} \\ &= \beta_1 \{ \beta_2^2 \} \end{aligned} \tag{1.3}$$

that would have been estimated, had an ideal randomised controlled trial (i.e. with 100% compliance) in fact been possible. Here,  $\beta_1 \{ \beta_2^2 \}$  can in principle be obtained by integrating over  $\beta_2 U$ . For a normally distributed confounder,  $\beta_1$  and  $\beta_1 \{ \beta_2^2 \}$  are linked by the approximate 1-1 transform  $\beta_1 \{ \beta_2^2 \} \approx \beta_1 (c^2 \beta_2^2 + 1)^{-\frac{1}{2}}$ , where  $c = 16\sqrt{3}/15\pi$  and where the term  $\beta_2^2$  is the variance of the confounder term  $\beta_2 U$ , as in Zeger and Liang (1988). This approximation is only valid because  $x_0$  is a constant and thus evidently independent of  $U$ ; it will be referred to as the ZL approximation from now on. The above result shows that the CLOR will in general be an attenuated version of  $\beta_1$ , except when  $\beta_2 = 0$ . This is an unavoidable consequence of the fact that the odds ratio as a statistic is non-collapsible, or equivalently that (with respect to  $U$ ) the underlying marginal and conditional parameters are not equal (Greenland et al., 1999).

In an effort to obtain parameter estimates from non-randomised observational data which are free from confounding, and hence worthy of a causal interpretation, Instrumental Variable (IV) methods have been proposed. In order for a particular variable to qualify as an instrument it must be: independent of the confounder  $U$ , not independent (and hence predictive of) the exposure  $X$  and independent of the outcome  $Y$ , conditional on the exposure and confounder. These assumptions are represented by the causal diagram in Figure 2, in which  $G$  is the IV. In recent years it has become feasible and attractive to let genetic information, such as a single nucleotide polymorphism (SNP) or a set of SNP’s, to play the role of the IV. Since genes are randomly assigned from parents to their offspring at the point of conception they are highly likely to be independent of confounding factors, such as diet and other lifestyle choices that impact an individual’s health in later life. If a gene is a valid IV then one can think of the individuals with/without it as being randomised to different arms of a hypothetical clinical trial, and as

such we can use this information to help make causal inferences (Katan, 1986). The use of genetic instruments is referred to as ‘Mendelian randomisation’ and is becoming increasingly popular. At first glance their justification as instruments appears strong, however, there are many biological factors unique to genes that can prohibit their use as IV’s, due to violations of the necessary conditions. This issue is not addressed, see Didelez and Sheehan (2007); Lawlor et al. (2008) for a rigorous discussion.

The material in this paper is motivated by considering possible Instrumental Variable analyses of population data under an identical model to that considered by Palmer et al. (2008), and assessing how well they can approximate the causal effect that would have been obtained from an RCT. In Section 2 we show the precise differences between estimates for the association between exposure and outcome, and estimates for the causal effect of the exposure under their proposed Instrumental Variable method. In Section 3 we highlight its equivalence in this setting with (a particular incarnation of) the Logistic Structural Mean Model of Vansteelandt and Goetghebeur (2003). In Section 4 we explore the IV methods’ performance in estimating the CLOR with prospective data. In Section 5 we discuss their use with retrospective data from a case-control study. We conclude with a discussion.

## 2 The model

Following Palmer et al. (2008) we assume that the model for  $X$  given a measurable covariate  $G$  and the previously defined  $U$ , is

$$X = \alpha_0 + \alpha_1 G + \alpha_2 U + \epsilon \quad (2.1)$$

where  $\epsilon \sim N(0, \sigma_\epsilon^2)$  that is independent of  $G$  and  $U$ . The  $\epsilon$  term is important because it represents our belief that the exposure level is not totally deterministic, given  $G$  and  $U$ , there is always some random variation. In accordance with the causal diagram in Figure 2 we assume that  $G$  is independent of  $U$  and that  $G$  only affects  $Y$  through  $X$ . This is made explicit by re-writing model (1.1) to be

$$\text{logit}(Pr(Y = 1|X = x, U = u, G = g)) = \beta_0 + \beta_1 x + \beta_2 u \quad (2.2)$$

so that it is clear  $G \perp\!\!\!\perp Y|X, U$ . Thus, as defined in equations (2.1) and (2.2),  $G$  is an IV. When  $X$  is observed at level  $x$ , what would a logistic regression of  $Y$  on  $X$  estimate as the odds ratio increase for unit change in  $X$ ? To derive an expression for this quantity we must deal with two separate issues; firstly because  $X$  is observed at  $x$  (not fixed as in an RCT) it may not be independent of  $U$ ; secondly, since  $u$  is by definition not observable we are forced to average over it as before. To achieve these two aims, and also to motivate subsequent results, it is convenient to re-write equation (2.2) so that it is a linear function of  $x$  plus independent error. The covariance matrix

$$\Sigma_{xu} = \begin{pmatrix} \sigma_x^2 & \rho_{ux}\sigma_x \\ \rho_{ux}\sigma_x & 1 \end{pmatrix}$$

describes the conditional dependence between  $u$  and  $x$  (given  $G$ ), where  $\rho_{ux}$  equals  $\frac{\alpha_2}{\sigma_x}$ . This correlation can be removed by making the Cholesky transformation  $x = \sigma_x\omega_1$ ,  $u = \rho_{ux}\omega_1 + \sqrt{1 - \rho_{ux}^2}\omega_2$  where  $\omega_1, \omega_2$  are i.i.d  $N(0, 1)$  variables. We now re-write the right hand side of equation (1.1) so that its mean and variance are unchanged, and also so that  $X$  is independent of the residual error, as

$$\begin{aligned} \text{logit}(Pr(Y = 1|X = x, U = u, G = g)) &= \beta_0 + \beta_1x + \beta_2u \\ \text{(for an } \omega_2 \perp\!\!\!\perp x) &= \left(\beta_0 - \frac{\beta_2\alpha_2}{\sigma_x^2}\mu_x\right) + \left(\beta_1 + \frac{\beta_2\alpha_2}{\sigma_x^2}\right)x \\ &\quad + \beta_2\sqrt{1 - \frac{\alpha_2^2}{\sigma_x^2}}\omega_2 \end{aligned}$$

where  $\mu_x$  and  $\sigma_x^2$  represent the underlying mean and variance of  $X$ . It now follows that

$$\text{logit}(Pr(Y = 1|X = x, G = g)) \approx \left(\beta_0 - \frac{\beta_2\alpha_2}{\sigma_x^2}\mu_x\right) \{\sigma_{obs}^2\} + \left(\beta_1 + \frac{\beta_2\alpha_2}{\sigma_x^2}\right) \{\sigma_{obs}^2\} x$$

where  $\sigma_{obs}^2 = \beta_2^2(1 - \frac{\alpha_2^2}{\sigma_x^2})$ . As one would logically expect, it follows that when there is no confounding between  $X$  and  $Y$ , by either  $\alpha_2, \beta_2$  or both being zero, then fitting a logistic regression model for  $Pr(Y = 1|X = x, G = g)$  yields a valid estimate of the CLOR. However, when both  $\beta_2$  and  $\alpha_2$  are non-zero it will not. Note that the CLOR is in theory restricted to the same sign as  $\beta_1$  but will in

general be smaller in magnitude. The same is not true of the associational log odds ratio, since its confounding bias  $\frac{\beta_2\alpha_2}{\sigma_x^2}$  component could in theory be larger in magnitude and opposite in sign to  $\beta_1$ .

The most basic of IV analyses might take the following form. Since  $G \perp\!\!\!\perp U$  a linear regression of the IV only on the exposure should yield consistent parameter estimates for  $\alpha_0$  and  $\alpha_1$ . The fitted values of this linear regression could then themselves be used as the covariate in a logistic regression on  $Y$ , and then the coefficient of the fitted values taken as an estimate for the CLOR. As given in Palmer et al. (2008) this would lead to an estimate of the form  $\beta_1 \{\sigma_{IV}^2\}$  where  $\sigma_{IV}^2$  equals  $(\beta_1\alpha_2 + \beta_2)^2 + \beta_1^2\sigma_\epsilon^2$  since

$$\begin{aligned} \text{logit}(Pr(Y = 1|X = x, U = u, G = g)) &= \beta_0 + \beta_1x + \beta_2u \\ &= \beta_0 + \beta_1(\alpha_0 + \alpha_1g) + (\beta_1\alpha_2 + \beta_2)u + \beta_1\epsilon \\ \text{(large sample approximation)} &\approx \beta_0 + \beta_1\hat{x} + (\beta_1\alpha_2 + \beta_2)u + \beta_1\epsilon \end{aligned}$$

from which

$$\text{logit}(Pr(Y = 1|X = x, G = g)) \approx \beta_0 \{\sigma_{IV}^2\} + \beta_1 \{\sigma_{IV}^2\} \hat{x}$$

This is the standard two-stage least squares method - as applied to a continuous outcome - transferred to the binary setting. If  $Y$  was continuous and linear models were used, then this approach would yield an unbiased estimate of the causal effect of  $X$ , however in the binary setting the correction is only partial. The fitted value  $\hat{x}$  is independent of the residual error term (involving  $u$  and  $\epsilon$ ) and so confounding bias is removed. This guarantees at least that the sign of the estimate will be the same as the CLOR, but it will not in general be equal to the CLOR because of the disparity between  $\sigma_{IV}^2$  and  $\beta_2^2$ . The exception is when  $\beta_1 = 0$  and so this approach could at least be used to test the null hypothesis of no causal effect. However, a worrying facet of this estimator is that it does not necessarily coincide with the CLOR even when there is no confounding between  $X$  and  $Y$ , something that even the associational estimate achieves.

## 2.1 The adjusted IV estimate

After noting the limitations of the standard IV approach, Palmer et al. (2008) suggested an alternative estimator, with an additional term in the logistic regression

- this being the residual error from the first stage regression  $r = x - \hat{x} = \alpha_2 u + \epsilon$ . A similar idea was proposed by Nagelkerke et al. (2000) as a way to adjust for non-compliance in randomised clinical trials. To see why this is advantageous we define the random variable  $R = X - \alpha_0 - \alpha_1 G$ , employ a similar transform to remove the correlation between  $R$  and  $U$ , and note that

$$\begin{aligned}
\text{logit}(Pr(Y = 1|X = x, U = u, G = g)) &= \beta_0 + \beta_1 x + \beta_2 u \\
&= \beta_0 + \beta_1(\alpha_0 + \alpha_1 g) + \beta_1(\alpha_2 u + \epsilon) + \beta_2 u \\
(\text{large sample approximation}) &\approx \beta_0 + \beta_1 \hat{x} + \beta_1 r + \beta_2 u \\
\text{for an } \omega_2 \perp\!\!\!\perp (\hat{x}, r) &= \beta_0 + \beta_1 \hat{x} + \left( \beta_1 + \frac{\beta_2 \alpha_2}{\alpha_2^2 + \sigma_\epsilon^2} \right) r \\
&\quad + \beta_2 \sqrt{\frac{\sigma_\epsilon^2}{\alpha_2^2 + \sigma_\epsilon^2}} \omega_2
\end{aligned}$$

from which

$$\text{logit}(Pr(Y = 1|X = \hat{x}, R = r)) \approx \beta_0 \{ \sigma_{adjIV}^2 \} + \beta_1 \{ \sigma_{adjIV}^2 \} \hat{x} + \beta_r \{ \sigma_{adjIV}^2 \} r$$

where  $\sigma_{adjIV}^2 = \frac{\beta_2^2 \sigma_\epsilon^2}{\alpha_2^2 + \sigma_\epsilon^2}$  and  $\beta_r = \beta_1 + \frac{\beta_2 \alpha_2}{\alpha_2^2 + \sigma_\epsilon^2}$ .

Unlike the basic IV estimate, the adjusted IV estimate should be approximately equal to the CLOR when no confounding is present. Under the ZL approximation, it provides an estimate which lies somewhere between  $\beta_1$  and the CLOR since  $\beta_2^2 \geq \frac{\beta_2^2 \sigma_\epsilon^2}{\alpha_2^2 + \sigma_\epsilon^2} \geq 0$ . As to where in this range the estimate lies, will clearly depend on the proportion of variation carried over from the first stage regression *not* explained by  $U - \frac{\sigma_\epsilon^2}{\alpha_2^2 + \sigma_\epsilon^2}$  - which we will now refer to as  $P_\epsilon$ . To give a rough indication of the bias one might experience, Figure 3 plots the difference between the adjusted IV estimate and the CLOR under the ZL approximation as a function of  $\beta_2$  and  $P_\epsilon$ , for  $\beta_1 = 1$  and  $\beta_2$  in  $(0,1)$ . At  $\beta_2 = 1$  a unit increase in the value of a the unknown confounder increases the odds ratio of disease close to 3 times, which we feel is a fairly extreme scenario.

### 2.1.1 Implementing the adjusted IV method

Let us assume we have data  $x_i, g_i, y_i$  for  $i = 1, \dots, n$ , that is complete information on the exposure, instrumental variable and outcome for  $n$  independent subjects.

Point estimates for the CLOR can be obtained under the adjusted IV method by first regressing  $g$  on  $x$  using least squares, fitting the predicted values  $\hat{x}$  plus the residuals in a logistic regression on  $Y$  and taking the resulting coefficient of  $\hat{x}$ . Equivalently this point estimate could be obtained by solving (with respect to  $\theta \equiv (\theta_1, \dots, \theta_5)$ ) the multivariate score equation  $\sum_{i=1}^n S_i(\theta) = \underline{0}$  and taking its 4th element, where  $S_i(\theta)$  equals

$$\left( \begin{array}{c} \left( \begin{array}{c} 1 \\ g_i \end{array} \right) \\ \left( \begin{array}{c} 1 \\ \theta_1 + \theta_2 g_i \\ r_i(\theta) \end{array} \right) \end{array} \right) \begin{array}{c} W_i r_i(\theta) \\ [y_i - \text{expit} \{ \theta_3 + \theta_4 \hat{x}_i + \theta_5 r_i(\theta) \}] \end{array} \quad (2.3)$$

where  $r_i(\theta) = x_i - \theta_1 - \theta_2 g_i$ . The  $W_i$  term equals 1 for all subjects when we have standard prospective data, that is data on subjects who collectively form a representative sample from the general population of interest. When this is the case they can be ignored totally. Under the two-stage approach the variance of the second stage estimates - most importantly the regression coefficient of  $\hat{x}$  - do not take into account the uncertainty from the first stage regression, whereas the variance of the estimator  $\hat{\theta}$  does and is well approximated by the ‘sandwich’ expression

$$\frac{1}{n} E^{-1} \left( \frac{\partial S_i(\theta)}{\partial \theta} \right) \text{Var} [S_i(\theta)] E^{-1} \left( \frac{\partial S_i(\theta)}{\partial \theta} \right)^T$$

The middle term is estimated by taking the sample variance of the individual score terms at  $\theta = \hat{\theta}$ .  $E \left( \frac{\partial S_i(\theta)}{\partial \theta} \right)$  is estimated by; firstly calculating the gradient matrix  $\left( \frac{\partial S_i(\theta)}{\partial \theta} \right)$  for each subject - whose  $jl^{th}$  element is the derivative of the  $j$ 'th component of  $S_i(\theta)$  with respect to the  $l$ th element of  $\theta$  - secondly evaluating it at  $\hat{\theta}$  and finally calculating the sample average over all subjects. The fourth diagonal element of the resultant matrix is the estimated variance of the CLOR.

### 3 Estimation under a Logistic Structural Mean Model

We now discuss applying a Logistic Structural Mean Model (LSMM) (Vansteelandt and Goetghebeur, 2003) to the problem. Although it is more common to express the LSMM in counterfactual terms, for consistency we follow the notation of Didelez et al. (2008); let  $X$  denote the natural (i.e. observed) exposure level and  $\tilde{X}$  denote the exposure that the natural level is set to after an intervention. The LSMM then postulates that following relation holds in the general population

$$\log \left\{ \frac{\text{odds } Pr(Y = 1|X = x, G = g, do(\tilde{X} = x))}{\text{odds } Pr(Y = 1|X = x, G = g, do(\tilde{X} = 0))} \right\} = \psi x \quad (3.1)$$

It follows that

$$\psi = \log \left\{ \frac{\text{odds } Pr(Y = 1|X = 1, G = g, do(\tilde{X} = 1))}{\text{odds } Pr(Y = 1|X = 1, G = g, do(\tilde{X} = 0))} \right\}$$

expresses the difference in log odds that would be realized for subjects with unit exposure ( $X = 1$ ) if their exposure were set to 0. In view of this,  $\psi x$  is commonly defined as the causal effect of ‘exposure on the exposed’. This subtle difference in interpretation arises because the LSMM, unlike the previous structural equation models, does not make the untestable assumption that all subjects experience the same effect of a given exposure level, no matter their natural exposure level.

Estimation of  $\psi$  in the LSMM is based on G-estimation (Greenland et al., 2008; Vansteelandt and Goetghebeur, 2003). For convenience let  $\eta_{\tilde{X}=x}$  represent the *logit* quantity in the numerator of equation (3.1) and  $\eta_{\tilde{X}=0}$  the term in the denominator. Let  $expit(p)$  be the inverse of  $logit(p)$ . Although we do not observe  $\eta_{\tilde{X}=0}$ , under the IV assumptions  $expit(\eta_{\tilde{X}=0})$  does not depend on  $g$ . By noting that  $Pr(Y = 1|X = x, G = g, do(\tilde{X} = x)) = Pr(Y = 1|X = x, G = g)$  and specifying a working model for disease conditional on the IV and the observed exposure levels - i.e. an approximation for  $\eta_{\tilde{X}=x}$  - we can express  $\eta_{\tilde{X}=0}$  as

$$expit(\eta_{\tilde{X}=x} - \psi x) \quad (3.2)$$

and then estimate  $\psi$  as the value that makes (3.2) independent of  $g$ . Note that estimation in the LSMM is semi-parametric in the sense that it only relies on

correct specification of a working model for the probability of disease given exposure and IV. Unlike estimation under the previous structural equation models, it does not explicitly model the confounding variable  $U$ , nor the exact relationship between  $G$  and  $X$ , and additionally makes no assumptions about the distribution of  $U$  and  $X$ . Misspecification of the working model does not bias tests of the null hypothesis of no causal effect so long as it includes at least an intercept and main effect of the IV (Babanezhad et al., 2008). To maintain comparability with the adjusted IV estimate, we now evaluate what the LSMM would estimate under the structural equation model (1.1). Note that the LSMM does not hold under that model, and neither does a logistic working model for  $Pr(Y = 1|X = x, G = g)$  with main effects in  $X$  and  $G$ , so that we evaluate this approach under model misspecification.

It follows using deterministic extended causal DAGs (Robins et al., 2006) that model (1.1) implies  $\text{logit}\left(Pr(Y = 1|X = x, G = g, do(\tilde{X} = x), U = u)\right)$  equals

$$\begin{aligned}
& \beta_0 + \beta_1 x + \beta_2 u \\
\text{from (2.1)} &= \beta_0 + \beta_1 x + \frac{\beta_2}{\alpha_2}(x - \alpha_0 - \alpha_1 g - \epsilon) \\
&= \left(\beta_0 - \frac{\beta_2 \alpha_0}{\alpha_2}\right) + \left(\beta_1 + \frac{\beta_2}{\alpha_2}\right)x - \frac{\beta_2 \alpha_1}{\alpha_2}g - \frac{\beta_2}{\alpha_2}\epsilon \\
&= \beta'_0 + \beta'_1 x + \beta'_2 g + \beta'_3 \epsilon
\end{aligned} \tag{3.3}$$

and that  $\text{logit}\left(Pr(Y = 1|X = x, G = g, do(\tilde{X} = 0), U = u)\right)$  equals

$$\begin{aligned}
& \beta_0 + \beta_2 u \\
\text{from (2.1)} &= \beta_0 + \frac{\beta_2}{\alpha_2}(x - \alpha_0 - \alpha_1 g - \epsilon) \\
&= \left(\beta_0 - \frac{\beta_2 \alpha_0}{\alpha_2}\right) + \frac{\beta_2}{\alpha_2}x - \frac{\beta_2 \alpha_1}{\alpha_2}g - \frac{\beta_2}{\alpha_2}\epsilon \\
&= \beta'_0 + (\beta'_1 - \beta_1)x + \beta'_2 g + \beta'_3 \epsilon
\end{aligned} \tag{3.4}$$

To marginalise over  $\beta'_3 \epsilon$ , we need to account for the dependencies between  $x$ ,  $g$  and  $\epsilon$ , which can be represented by the  $3 \times 3$  covariance matrix

$$\begin{pmatrix} \sigma_x^2 & \rho_{gx}\sigma_x\sigma_g & \rho_{\epsilon x}\sigma_x\sigma_\epsilon \\ \rho_{gx}\sigma_x\sigma_g & \sigma_g^2 & 0 \\ \rho_{\epsilon x}\sigma_x\sigma_\epsilon & 0 & \sigma_\epsilon^2 \end{pmatrix} \tag{3.5}$$

Let  $\Sigma$  be the Cholesky decomposition of (3.5) and let  $\Sigma_{ij}$  denote the  $ij$ 'th element of  $\Sigma$ . Defining the quantities

$$a' = \frac{\beta'_3}{\Sigma_{11}} \left( \Sigma_{31} - \frac{\Sigma_{32}\Sigma_{21}}{\Sigma_{22}} \right) \quad \text{and} \quad b' = \frac{\beta'_3\Sigma_{32}}{\Sigma_{22}}$$

and making the transformation  $x, g, \epsilon \rightarrow \omega_1, \omega_2, \omega_3$ ,  $iid \sim N(0, 1)$  enables (3.3) to be written as

$$\begin{aligned} & \beta'_0 + \beta'_1 x + \beta'_2 g + \beta'_3 \epsilon \\ (\text{for } \omega_1, \omega_2, \omega_3 \perp\!\!\!\perp x, g) &= \beta'_0 + \beta'_1(\Sigma_{11}\omega_1) + \beta'_2(\Sigma_{21}\omega_1 + \Sigma_{22}\omega_2) \\ &+ \beta'_3(\Sigma_{31}\omega_1 + \Sigma_{32}\omega_2 + \Sigma_{33}\omega_3) \\ &= (\beta'_0 - a' \mu_x - b' \mu_g) + (\beta'_1 + a')x + (\beta'_2 + b')g + \beta'_3 \Sigma_{33} \omega_3 \end{aligned}$$

from which  $\text{logit} \left( Pr(Y = 1 | X = x, G = g, do(\tilde{X} = x)) \right)$  is

$$\begin{aligned} & \approx \lambda_0 \{ \sigma_{SMM}^2 \} + \lambda_1 \{ \sigma_{SMM}^2 \} x + \lambda_2 \{ \sigma_{SMM}^2 \} g \\ & = \eta_{\tilde{X}=x} \end{aligned}$$

where  $\sigma_{SMM}^2 = \beta_3'^2 \Sigma_{33}^2$ . Likewise we can equally approximate the counterfactual quantity  $\eta_{\tilde{X}=0}$  as

$$\lambda_0 \{ \sigma_{SMM}^2 \} + (\lambda_1 - \beta_1) \{ \sigma_{SMM}^2 \} x + \lambda_2 \{ \sigma_{SMM}^2 \} g$$

and therefore the difference between these two marginal quantities clearly implies that  $\psi \{ \sigma_{SMM}^2 \} x = \beta_1 \{ \sigma_{SMM}^2 \} x$ . It can be shown that

$$\begin{aligned} \sigma_{SMM}^2 &= \frac{\beta_2^2 \sigma_\epsilon^2}{\alpha_2^2} \left( 1 - \rho_{\epsilon x}^2 - \frac{\rho_{gx}^2 \rho_{\epsilon x}^2}{1 - \rho_{gx}^2} \right) \\ &= \beta_2^2 P_\epsilon \end{aligned}$$

Therefore, if the stated structural equation model holds, then LSMM approximates the same quantity as the adjusted IV estimator.

### 3.1 Implementation

In order to estimate the CLOR using this approach we could perform a logistic regression of the working model on  $Y$ , and then find, by a grid search or numerical optimisation, the  $\hat{\psi}$  that forces the independence in (3.2). This is equivalent to solving the 4-dimensional score equation  $\sum_{i=1}^n S_i(\theta) = \underline{0}$  where  $S_i(\theta)$  equals

$$\begin{pmatrix} W_i(g_i - E_W[g])(\text{expit}\{\theta_1 + (\theta_2 - \psi)x_i + \theta_3 g_i\} - E_W[\text{expit}\{\theta_1 + (\theta_2 - \psi)x + \theta_3 g\}]) \\ \begin{pmatrix} 1 \\ x_i \\ g_i \end{pmatrix} [y_i - \text{expit}\{\theta_1 + \theta_2 x_i + \theta_3 g_i\}] \end{pmatrix} \quad (3.6)$$

for  $\theta \equiv (\psi, \theta_1, \theta_2, \theta_3)$  and where the expectation terms  $E_W[\cdot]$  in the first score element are weighted means, with respect to  $W$  (i.e for the instrumental variable  $\frac{\sum_{i=1}^n W_i g_i}{\sum_{i=1}^n W_i}$ ) Again, we assume for the moment that all  $W$  terms are equal to 1. The first diagonal element of the equivalent sandwich matrix is the variance of this CLOR estimate that also properly accounts for the uncertainty in the working model. From now on we will refer to quantities derived directly from (3.6) as ‘LSMM(1)’ estimates.

### 3.2 A ‘marginalised’ LSMM estimate

Because the LSMM models the causal effect in the exposed,  $\text{exp}(\psi)$  may differ from the marginal CLOR which we identify as the target of interest in this paper. Inferring the latter requires the additional assumption that the effect of a given exposure level is the same for all subjects, no matter their natural exposure level. This assumption can be formally expressed as

$$\begin{aligned} & \log \left\{ \frac{\text{odds } Pr(Y = 1|X = x, G = g, do(\tilde{X} = x))}{\text{odds } Pr(Y = 1|X = x, G = g, do(\tilde{X} = 0))} \right\} \\ = & \log \left\{ \frac{\text{odds } Pr(Y = 1|X = x^*, G = g, do(\tilde{X} = x))}{\text{odds } Pr(Y = 1|X = x^*, G = g, do(\tilde{X} = 0))} \right\} \end{aligned}$$

and is known as the assumption of ‘no-current-treatment-interaction’ (Vansteelandt and Goetghebeur, 2005; Hernan and Robins, 2006). Note that a similar assumption is implicit in the previous structural equation models (but conditional

on  $U$ ). Under this assumption, the marginal CLOR can be estimated following the lines of Vansteelandt and Goetghebeur (2005): noting that

$$Pr(Y = 1|X = x, G = g, do(\tilde{X} = x^*)) = \text{expit} \{ \eta_{\tilde{X}=x} + \psi(x^* - x) \}$$

we can obtain the numerator of (1.3) as the sample average of  $Pr(Y = 1|X = x, G = g, do(\tilde{X} = x_0 + 1))$  over all subjects, and likewise for the denominator. We will refer to quantities derived from this extra marginalisation phase as ‘LSMM(2)’ estimates.

## 4 Estimation in practice

Although we have shown what common value the adjusted IV and LSMM approaches estimate in theory, to investigate their behaviour in practice we conduct a series of simulation studies. General population data sets of exposures, IV’s and disease states  $(x_i, g_i, y_i) \quad i = 1, \dots, 1000$  were generated from models (2.1) and (2.2) for specific parameter values and random error components  $\epsilon_i, u_i \sim N(0, 1)$ , enabling point estimates and variances to be obtained for the two methods. Following Palmer et al. (2008) the instrumental variable  $G$  was generated in a such a manner as to represent the number of copies (0,1 or 2) of a single bi-allelic SNP in Hardy-Weinberg equilibrium. The underlying allele frequency in the population was assumed to be  $p = 0.3$ , and so  $G$  was generated from a multinomial distribution with cell probabilities (0.09,0.42,0.49). The standard IV approach was fitted by firstly regressing the exposures on the IV to obtain a predicted value  $\hat{x}$ , and then performing a logistic regression of the predicted value on the disease states. The variance of this estimate was taken directly from the logistic regression output. The adjusted IV and LSMM(1) point estimates were obtained by solving score equations (2.3) and (3.6), their variances were calculated using the appropriate sandwich expression.

In order to assess the performance of the IV estimators in terms of bias and coverage we need to decide, for each simulation scenario, the CLOR of interest without appealing to the ZL approximation. In Section 1 we defined the CLOR in equation (1.3) for a unit increase in the exposure, from  $x_0$  to  $x_0 + 1$ . However this quantity will vary depending on the value of  $x_0$  chosen. To illustrate this a data set of size  $N = 5000$  was simulated under models (2.1) and (2.2) given parameter

values  $\alpha_0 = 0$ ,  $\alpha_1, \alpha_2, \beta_1, \beta_2 = 1$  and  $\beta_0 = -5$ . This induced a disease prevalence of approximately 10% in the sample. For each exposure level  $x_i$ ,  $i = 1, \dots, N$  we estimated

$$\begin{aligned} CLOR(x_i, x_i + 1) &= \log \left\{ \frac{\text{odds } Pr(Y = 1 | do(X = x_i + 1))}{\text{odds } Pr(Y = 1 | do(X = x_i))} \right\} \\ &= \log \frac{\int_{-\infty}^{\infty} \text{expit}(\beta_0 + \beta_1(x_i + 1) + \beta_2 u) \phi(u) du}{\int_{-\infty}^{\infty} \text{expit}(\beta_0 + \beta_1 x_i + \beta_2 u) \phi(u) du} \end{aligned}$$

via the Monte-Carlo analogue

$$\log \left\{ \frac{\text{odds } \frac{1}{N} \sum_{j=1}^N \text{expit}(\beta_0 + \beta_1(x_i + 1) + \beta_2 u_j)}{\text{odds } \frac{1}{N} \sum_{j=1}^N \text{expit}(\beta_0 + \beta_1 x_i + \beta_2 u_j)} \right\}$$

by making use of the simulated confounder variables  $u_1, \dots, u_N$ . The solid black line in Figure 4 shows the estimate for each  $x_i$ . This CLOR measure is equal to 1 for subjects with low exposure levels and decreases to a minimum of approximately 0.8 for subjects with higher exposures around 4. It then increases back towards 1 for those with high exposures. The ZL approximation for the CLOR -  $\beta_1 \{\beta_2^2\} \approx \beta_1 (c^2 \beta_2^2 + 1)^{-\frac{1}{2}}$  does not depend on  $X$ , and is shown by the horizontal line at  $\approx 0.86$ . A single representative measure of the causal effect would quite reasonably be  $CLOR(\bar{x}, \bar{x} + 1)$ , that is the CLOR defined for a hypothetical individual with the population mean exposure level  $\bar{x}$ . One could instead choose to estimate the CLOR averaged over all exposure levels, that is

$$E_x [CLOR(x, x + 1)] \approx \frac{1}{N} \sum_{i=1}^n \log \left\{ \frac{\text{odds } \frac{1}{N} \sum_{j=1}^N \text{expit}(\beta_0 + \beta_1(x_i + 1) + \beta_2 u_j)}{\text{odds } \frac{1}{N} \sum_{j=1}^N \text{expit}(\beta_0 + \beta_1 x_i + \beta_2 u_j)} \right\}$$

These two measures are estimated for this data and illustrated in Figure 4. To aid the interpretation of  $E_x [CLOR(x, x + 1)]$  the density of the exposures is also shown. In general they are very close, and due to its relative simplicity in calculation we choose  $CLOR(\bar{x}, \bar{x} + 1)$  as our benchmark.

Table 1 summarises the point estimates, variances and 95% CI coverages (with respect to the  $CLOR(\bar{x}, \bar{x} + 1)$ ) obtained by; the standard IV, adjusted IV, LSMM(1) and LSMM(2) approaches'. This last estimator was calculated by taking  $x^*$  to be  $\bar{x}$ , details as to how we calculated its variance can be found in the Appendix. Each

number is the average of 2000 independent simulations. Rows 1-5 show the effect of varying the parameter  $\beta_0$ , whilst holding the other parameter values fixed, rows 6-10 and 11-15 show the effect of varying  $\beta_2$  and  $\alpha_2$  respectively. The adjusted IV and LSMM(1) estimators perform significantly better than the standard IV estimator and in a highly similar fashion, as predicted. A typical correlation between the adjusted IV and LSMM(1) point estimates was 0.98. However, across all simulations the LSMM(2) estimator more closely followed the target of interest -  $CWOR(\bar{x}, \bar{x} + 1)$  - thereby exhibiting less bias.

## 4.1 Model misspecification

In order to investigate the performance of each method under model misspecification we altered the true underlying model for the disease data in the general population to include an exposure-confounder interaction term. The disease outcome  $y$  is now generated from

$$\text{logit}(\Pr(Y = 1|X = x, G = g, U = u)) = \beta_0 + \beta_1x + \beta_2u + \beta_{XU}xu$$

Since the effect of a unit increase in the exposure is now dependent upon  $U$  and thus different between subjects with different natural exposure levels, the assumption of no-current-treatment-interaction fails, which is explicitly relied on for the LSMM(2) method. Table 2 rows 1-5 show the performance of the four methods when data is generated under the above model.  $\beta_2$  was fixed to 0.5, all other previous parameter values the same and  $\beta_{XU}$  was varied between 0 and 1. In order to further remove the structure of the data from that which the theoretical results are based on we performed a second simulation, but this time modifying the confounder  $U$  to be a binary, rather than a normal variable. Each subject's confounder was simulated from independent Bernoulli trials with success probability  $p = 0.5$ , this created an exposure with a skewed continuous distribution, since the random error component  $\epsilon$  was still normal. Rows 6-10 show the results. Across both scenarios the adjusted IV and LSMM(1) estimators give indistinguishable results for their mean point estimates and prove more robust at mirroring the  $CWOR(\bar{x}, \bar{x} + 1)$  than the standard IV approach. Despite their estimates being biased the adjusted IV and LSMM(1) approaches' sandwich variances do a good job of maintaining coverage levels, across both confounder distributions. The LSMM (2) estimator performs well with a binary  $U$ , but is significantly worse

than the adjusted IV and LSMM(1) estimators with a normal  $U$ , in terms of bias and coverage.

## 5 Estimation with retrospective data

Prospective studies are sometimes not feasible for financial and practical reasons - especially if the disease outcome of interest has a low incidence rate - and the case-control design is often preferred. Let  $A = 1$  indicate the event that an individual is 'ascertained', that is recruited into a case-control study. It is well known that when the probability of disease in the general population is governed by model (2.2) then in the ascertained population

$$\text{logit}(Pr(Y = 1|X = x, G = g, U = u, A = 1)) = \tilde{\beta}_0 + \beta_1 x + \beta_2 u$$

where  $\tilde{\beta}_0 = \beta_0 + \log\left(\frac{\pi Pr(Y=0)}{(1-\pi)Pr(Y=1)}\right)$ ,  $\pi$  is the proportion of cases in the sample, and  $Pr(Y = 1) = 1 - Pr(Y = 0)$  (Prentice and Pyke, 1979; Whittemore, 1995). So, broadly speaking, estimates for the *association* between  $X$  and  $Y$  would be identical when based on observational data or on case-control data, as only the intercept parameter changes (if  $\pi = Pr(Y = 1)$  they are of course equivalent). So, can we make straightforward causal inferences using instrumental variables with case-control data? With prospective data which is a random sample from the general population of interest, we may justifiably rely on the assumption that  $G \perp\!\!\!\perp U$ . However even if this is true in the general population, the theory of d-separation tells us that conditional on being ascertained (and as such conditional on the value of  $Y$ ), a correlation is induced between  $G$  and  $U$  since  $Y$  is a common descendant of the two. This is often represented by the addition of a so called 'moral' edge between  $U$  and  $G$ , as in Figure 5. Specifically for the adjusted IV method, this would mean that a linear regression of  $X$  on  $G$  would not yield a consistent estimate for  $\alpha_1$  because the residual error term  $\alpha_2 U + \epsilon$  would be correlated with  $G$ , which in turn makes the second stage CLOR estimate invalid. For the LSMM method, this would mean that the basic idea underlying G-estimation - that (3.2) is independent of  $g$  - becomes flawed.

## 5.1 Correcting for ascertainment

Consider the distribution of  $X$  in the general population shown by the solid black line in Figure 6 (left). This was generated by simulating from population models (2.1) and (2.2) as described in Figure 4. Shown in red is the distribution of  $X$  in the disease cases. Since the sign of  $\beta_1$  means that positive values of  $X$  are associated with an increased diseased risk, the mean value of  $X|Y = 1$  is larger than in the general population. Conversely the distribution of  $X$  in the non-diseased population (shown in blue) is lower than than in the general population. However, because the disease prevalence is low this distribution is much closer to the general population than that of the cases. The grey line shows the distribution of the exposure in the *ascertained* population when the cases and controls are sampled in the ratio 1:2. Note the mean and variance are larger than in the general population. Figure 6 (right) shows the distribution of the instrumental variable  $G$  in the same groups. Given an allele frequency of 30% the proportion of hetrozygotes in the general population should be 0.42 under the HWE assumption, this is fairly close to the control group frequency but not that of the cases or the ascertained population.

The small distance between the distributions  $X$  and  $G$  in the general population and the control group lies behind the justification that, when analysing case-control data, inference for the effect of  $G$  on  $X$  should only be based on the controls (Thompson et al., 2005). This suggests that for the adjusted IV estimator one could approximately correct for ascertainment by modifying the score equation weights in (2.3) so  $W_i = 0$  if subject  $i$  is a case but 1 if they are a control. We will refer to this as the ‘crude’ weighting strategy. An alternative weighting strategy is possible if the disease prevalence is known; following the example of Whittemore (1995) we write the distribution of  $X$  in the general and ascertained populations as

$$\begin{aligned} P_{gen}(X) &= Pr(X|Y = 1)Pr(Y = 1) + Pr(X|Y = 0)Pr(Y = 0) \\ P_{asc}(X) &= Pr(X|Y = 1)\pi + Pr(X|Y = 0)(1 - \pi) \end{aligned}$$

which makes clear that, in our ascertained population we have over-sampled the case exposures relative to the general population by a factor of  $\frac{Pr(Y=1)}{\pi}$ . Likewise the control exposures have been under-sampled by a factor of  $\frac{Pr(Y=0)}{1-\pi}$ . The same

reasoning equally applies to  $G$ . This suggests that, for the adjusted IV estimator, one could more accurately correct for this selection effect by applying score equation (2.3) with modified weights  $W_i = \frac{Pr(Y=1)}{\pi}$  if subject  $i$  is a case and  $\frac{Pr(Y=0)}{1-\pi}$  otherwise. This is referred to as the ‘exact’ weighting strategy.

The LSMM is also affected by ascertainment. Let  $\eta_{\tilde{X}=x|A=1}$  represent the analogous unconditional quantity from Section 3, with a modified intercept parameter  $\tilde{\lambda}_0 = \lambda_0 + \log\left(\frac{\pi Pr(Y=0)}{(1-\pi)Pr(Y=1)}\right)$ . While it can be shown that in the case-control sample

$$\eta_{\tilde{X}=x|A=1} - \eta_{\tilde{X}=0|A=1} = \psi x$$

still holds, we cannot readily estimate estimate  $\psi$  for the following two reasons. First, whilst  $expit(\eta_{\tilde{X}=0}) \perp\!\!\!\perp G$  holds for prospective data, we do not have  $expit(\eta_{\tilde{X}=0|A=1}) \perp\!\!\!\perp G|A=1$ . Second, even if it were true that  $expit(\eta_{\tilde{X}=0|A=1}) \perp\!\!\!\perp G|A=1$ , this would require a consistent estimator of the intercept parameter  $\lambda_0$ . After some investigation, we suggest therefore to fit the same working model as before, and then to use weighted G-estimation (with an ascertainment corrected offset) to obtain  $\psi$  via the approximation

$$E_W[expit(\eta_{\tilde{X}=x|A=1} - O_A - \psi x)|A=1] \approx expit(\eta_{\tilde{X}=x} - \psi x) \perp\!\!\!\perp G \approx E_W[G|A=1]$$

where  $O_A = \log\left\{\frac{\pi Pr(Y=0)}{(1-\pi)Pr(Y=1)}\right\}$  is the offset. This approximation is implemented by solving score equation (3.6), but for the first element of the score equation substituting the  $W_i$  terms with the previously defined ‘exact’ weights, and using the same information to correctly specify the offset  $O_A$ . Crude weighting can also be used as before, but with less justification because it does not lead to an approximately valid working model for disease given exposure and IV.

How well can the crude or exact weighting methods perform with retrospective data, in the sense of producing inferences similar to those that would have been obtained from an unweighted analysis of prospective data? For the same parameter values as in Table 1 (rows 6-10), prospective and 1:2 case-control data sets of size  $n = 1500$  were simulated. Table 3 shows the average estimates for the CLOR obtained from the prospective data under the adjusted IV and LSMM (1) methods. These are compared to the estimates obtained from case-control data

(under the two weighting strategies) through their average difference. For the exact weighting method,  $Pr(Y = 1)$  was assumed fixed and known for the simulations. The variance of the case-control estimates is also shown, along with their coverage with respect to the  $CLOR(\bar{x}, \bar{x} + 1)$  - again, see Table 1 for these values.

For the case-control data analysed using exact weighting, the adjusted IV and LSMM (1) methods give near identical mean point estimates, which exhibit a negligible difference compared to those based on prospective data. Crude weighting performs fairly well but produces mean point estimates with small but non-negligible differences to their prospective counterparts. This difference increases with  $\beta_2$  and is reflected by a loss of coverage with respect to the  $CLOR(\bar{x}, \bar{x} + 1)$ . Crude weighting seems to affect the adjusted IV and LSMM(1) methods differently, with the LSMM(1) estimate performing slightly worse. The precision of both the adjusted IV and LSMM(1) the estimates is also detrimentally affected by crude weighting, since one effectively throws away a third of the data when summing particular components of score equations (2.3) and (3.6).

## 6 Discussion

Using instrumental variables to estimate the causal effect of continuous exposure on a binary outcome is a challenging task, especially when the effect is measured as an odds ratio. In this paper we have shown that under various modelling and distributional assumptions, including a multivariate normal  $X$  and  $U$ , the adjusted IV method of Palmer et al. (2008) and LSMM (1) estimate of Vansteelandt and Goetghebeur (2003) target the same quantity in theory, and in practice also perform very similarly in terms of bias, variance and their sensitivity to model misspecification. Real differences between the methods were apparent when retrospective data was analysed using ‘crude’ weighting, with the LSMM(1) performing slightly worse. In many other circumstances the two methods do behave differently when applied to prospective data, for example when the exposure is dichotomous (Babanezhad et al., 2008). There, the LSMM method is less prone to bias (and guaranteed to be unbiased at the causal null hypothesis) as a result of making fewer modeling assumptions and avoiding approximations, but can have somewhat greater imprecision. The marginalised LSMM (2) estimate was shown to potentially provide the least biased estimate of all, but to be possibly more

vulnerable to an exposure-confounder interaction.

In contrast to Palmer et al. (2008), who focus on the direct effect  $\beta_1$  of the exposure on outcome conditional on  $U$ , and to Vansteelandt and Goetghebeur (2003), who focus on the exposure effect in the exposed, we highlight the odds ratio obtained from a randomised controlled trial as the target of interest. In particular, we focused on the marginal causal log odds ratio  $CLOR(\bar{x}, \bar{x} + 1)$ , which expresses how much the odds of disease would change if each subject's exposure were increased from  $\bar{x}$  to  $\bar{x} + 1$ . Alternatively, we could have considered the following estimand

$$\log \frac{Pr(Y_{X+1} = 1)Pr(Y_X = 0)}{Pr(Y_{X+1} = 0)Pr(Y_X = 1)}$$

where  $Y_X \equiv Y$  and where

$$Pr(Y_{X+1} = 1) = \int Pr(Y_{x+1} = 1|X = x, G)f(X = x|G)f(G)dx dG$$

This estimand expresses how much the odds of disease would change if each subject's observed exposure were increased with a unit. Under the no-current-treatment-interaction assumption, it can be calculated as

$$\log \frac{\mu_1(1 - \mu_0)}{\mu_0(1 - \mu_1)}$$

with

$$\mu_1 = E \{ \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i + \psi) \}$$

and

$$\mu_0 = E \{ \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i) \}$$

where population values can be replaced with consistent estimates and population expectations with sample averages. In the economics literature  $\beta_1$  is more commonly sought and is often termed the *ceteris paribus* causal effect, that is the effect of  $X$  on  $Y$  if all other variables are held equal, Bierens and Swanson. (2000). The disparity between  $\beta_1$  and the CLOR can be large, which might mean that despite there existing a strong causal relationship within many subgroups of the population, the relationship at the aggregate level appears weak or even non-existent. Although one may therefore wish to estimate the CLOR for these specific subgroups instead, using  $\beta_1$  does not attain this goal because it expresses the effect for subgroups which are unknown (by the fact that  $U$  is unknown).

Suppose therefore that the composite confounder  $u$  can be partitioned into two sources; known and unknown. Let  $U \sim N(0, 1)$  be the unknown component as before and let  $V \sim N(0, 1)$  be the known component, reflecting the subgroups that we may be interested in. Let  $\rho_{uv}$  be the correlation between  $U$  and  $V$ . Using similar correlation removing transforms as before we rewrite the first regression equation

$$\begin{aligned} f(X = x|G = g, U = u, V = v) &= \alpha_0 + \alpha_1 g + \alpha_2 u + \gamma_1 v + \epsilon \\ &= \alpha_0 + \alpha_1 g + \gamma_1^* v + \alpha_2^* \omega_2 + \epsilon \end{aligned}$$

where  $\gamma_1^* = \gamma_1 + \alpha_2 \rho_{uv}$ ,  $\alpha_2^* = \alpha_2 \sqrt{(1 - \rho_{uv}^2)}$  and  $\omega_2$  is an independent  $N(0, 1)$  error term. The linear predictor for the logistic regression on  $Y$  becomes

$$\begin{aligned} \text{logitPr}(Y = 1|X = x, G = g, U = u, V = v) &= \beta_0 + \beta_1 x + \Delta_1 v + \beta_2 u \\ &= \beta_0 + \beta_1 x + \Delta_1^* v + \beta_2^* \omega_2 \end{aligned}$$

where  $\Delta_1^* = \Delta_1 + \beta_2 \rho_{uv}$  and  $\beta_2^* = \beta_2 \sqrt{(1 - \rho_{uv}^2)}$ . It is simple to show that when conditioning on the covariate  $V$  in the analyses as above, the adjusted IV approach will produce a log odds ratio estimate of  $\beta_1 \left\{ \frac{\beta_2^* \sigma_\epsilon^2}{\alpha_2^{*2} + \sigma_\epsilon^2} \right\}$ . This will also approximately be the quantity estimated by the Logistic structural mean model given the true model above and a working model for  $\text{logitPr}(Y = 1|X = x, G = g, V = v)$  that is linear in  $X, G, V$  with no interaction terms. Thus, any non-zero correlation between  $U$  and  $V$  will reduce the distance between the marginal log-odds ratio and the underlying conditional estimate, see Figure 7 for an illustration. However, care must be taken when including covariates in the analysis as they can invalidate the IV methods; one must be certain that  $G$  is an IV conditional on the extra covariates, which may be hard to determine, and flawed when these covariates are simultaneously influenced by the IV and outcome.

General weighted estimating equations were suggested in order to implement the adjusted IV and LSMM approaches, which allowed the analysis of prospective and retrospective data within a single framework. Unsurprisingly ‘Exact’ weighting of cases and controls using the population disease prevalence was shown to outperform crude weighting, in terms of bias and precision. However, in reality there may be uncertainty about the disease prevalence. In that case, an alternative to

crude weighting might be to postulate lower and upper bounds for the prevalence and to perform a sensitivity analysis, King and Zeng (2002). The usual confidence intervals can then be accommodated to take the uncertainty on the disease prevalence into account, Vansteelandt et al. (2006). Under the rare disease assumption, an alternative would be to use that

$$Pr(Y = 1|X = x, G = g, U = u) \approx \exp(\tilde{\beta}_0)\exp(\beta_1X + \beta_2U)$$

The fact that the intercept parameter can be isolated in this way may make it easier in particular for Structural Mean Models to be applied to retrospective data in the same manner as for prospective data (without informative weights), since G-estimation could then be employed successfully without knowledge of  $\tilde{\beta}_0$ . This is a topic for further work.

## References

- BABANEZHAD, M., VANSTEELANDT, S., and GOETGHEBEUR, E. (2008). On the performance of instrumental variable estimators of the causal odds ratio. Technical report, Ghent University.
- BIERENS, H.J., SWANSON, N.R. (2000). The econometric consequences of the ceteris paribus condition in economic theory. *Journal of Econometrics*, 95:223–253.
- DIDELEZ, V. AND SHEEHAN, N. (2007). Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16:309–330.
- DIDELEZ, V., S.MENG, AND N.SHEEHAN (2008). On the bias of iv estimators for mendelian randomisation. Technical report, University of Bristol.
- GREENLAND, S., ROBINS, J., AND PEARL, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14:29–46.
- GREENLAND, S., LANES, S., AND JARA, M., (2008). Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and G-estimation. *Clinical Trials*, 5:5–13.

- HERNAN, M.A. AND ROBINS, J.M. (2006). Instruments for causal inference - An epidemiologist's dream? *Epidemiology* 17, 360–372.
- KATAN, M. (1986). Apolipoprotein e isoforms, serum cholesterol, and cancer. *Lancet*, 327:507–508.
- KING, G. AND ZENG, L.C., (2002) Estimating risk and rate levels, ratios and differences in case-control studies *Statistics in Medicine*, 21:1409–1427.
- LAWLOR, D., HARBORD, R., STERNE, J., TIMPSON, N., AND DAVEY-SMITH, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27:1133–1163.
- LIN, D., PSATY, B., AND KRONMAL, R. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, 54:948–963.
- NAGELKERKE, N., FIDLER, V., R.BERNSEN, AND BORGDORFF, M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine*, 19:1849–64.
- PALMER, T., THOMPSON, J., TOBIN, M., SHEEHAN, N., AND BURTON, P. (2008). Adjusting for bias and unmeasured confounding in mendelian randomisation studies with binary responses. *IJE*, 37: 1161–1168
- PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82:699–710.
- PRENTICE, R. AND PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–11.
- ROBINS JM, VANDERWEELE TJ, RICHARDSON TS. (2006). Comment on causal effects in the presence of non compliance: a latent variable interpretation by Antonio Forcina METRON vol. LXIV(3)288–298.
- THOMPSON, J., MINELLI, C., ABRAMS, K., TOBIN, M., AND RILEY, R. (2005). Meta-analysis of genetic association studies using mendelian randomisation - a multivariate approach. *Statistics in Medicine*, 24:2241–2254.

- TSIATIS, A.A. (2006). *Semiparametric Theory and Missing Data*. Springer: New York.
- VANSTEELANDT, S. AND GOETGHEBEUR, E. (2003). Causal inference with generalized structural mean models. *JRSSB*, 65:817–835.
- VANSTEELANDT, S. AND GOETGHEBEUR, E. (2005). Sense and sensitivity when correcting for observed exposures in randomized clinical trials. *Statist. Med.* 24, 191–210.
- VANSTEELANDT, S., GOETGHEBEUR, E., KENWARD, M., AND MOLENBERGHS, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16:953–979.
- WHITTEMORE, A. (1995). Logistic regression of family data from case-control studies. *Biometrika*, 82:57–67.
- ZEGER, S. AND LIANG, K. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44:1049–1060.

## A Appendix

### A.1 Variance of the LSMM(2) estimator

Note that

$$CLOR(\bar{x}, \bar{x} + 1) = \log \frac{\mu_1(1 - \mu_0)}{\mu_0(1 - \mu_1)}$$

where

$$\mu_1 = E \{ \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i + \psi(\bar{x} + 1 - X_i)) \}$$

and

$$\mu_0 = E \{ \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i + \psi(\bar{x} - X_i)) \}$$

Let us denote this estimand with  $\eta$  and let the corresponding estimator be  $\hat{\eta}$ . Likewise, denote the estimators of  $\mu_1$  and  $\mu_0$  with  $\hat{\mu}_1$  and  $\hat{\mu}_0$ , respectively. Then

a Taylor series expansion shows that

$$\begin{aligned}
0 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \text{expit}(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 G_i + \hat{\psi}(\bar{x} - X_i)) - \hat{\mu}_0 \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i + \psi(\bar{x} - X_i)) - \mu_0 \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n E \left[ \frac{\partial}{\partial \theta} \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i + \psi(\bar{x} - X_i)) \right] E^{-1} \left( \frac{\partial U_i(\theta)}{\partial \theta} \right) U_i(\theta) \\
&\quad - \sqrt{n}(\hat{\mu}_0 - \mu_0)
\end{aligned}$$

from which

$$\begin{aligned}
\sqrt{n}(\hat{\mu}_0 - \mu_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i + \psi(\bar{x} - X_i)) - \mu_0 \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n E \left[ \frac{\partial}{\partial \theta} \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i + \psi(\bar{x} - X_i)) \right] E^{-1} \left( \frac{\partial U_i(\theta)}{\partial \theta} \right) U_i(\theta)
\end{aligned}$$

It follows that the influence function (Tsiatis, 2006) for  $\hat{\mu}_0$  is

$$\begin{aligned}
&\text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i + \psi(\bar{x} - X_i)) - \mu_0 \\
&+ E \left[ \frac{\partial}{\partial \theta} \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i + \psi(\bar{x} - X_i)) \right] E^{-1} \left( \frac{\partial U_i(\theta)}{\partial \theta} \right) U_i(\theta)
\end{aligned}$$

and, from the Delta method, that the influence function for  $\hat{\eta}$  is

$$\begin{aligned}
&\frac{1}{\mu_1(1 - \mu_1)} [\text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i + \psi(\bar{x} + 1 - X_i)) - \mu_0] \\
&+ E \left[ \frac{\partial}{\partial \theta} \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i + \psi(\bar{x} + 1 - X_i)) \right] E^{-1} \left( \frac{\partial U_i(\theta)}{\partial \theta} \right) U_i(\theta) \\
&- \frac{1}{\mu_0(1 - \mu_0)} [\text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i + \psi(\bar{x} - X_i)) - \mu_0] \\
&+ E \left[ \frac{\partial}{\partial \theta} \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 G_i + \psi(\bar{x} - X_i)) \right] E^{-1} \left( \frac{\partial U_i(\theta)}{\partial \theta} \right) U_i(\theta)
\end{aligned}$$

The asymptotic variance of  $\hat{\eta}$  thus equals the 1 over  $n$  times the variance of this influence function (where averages and variances can be replaced with sample analogs, and population values with consistent estimators).

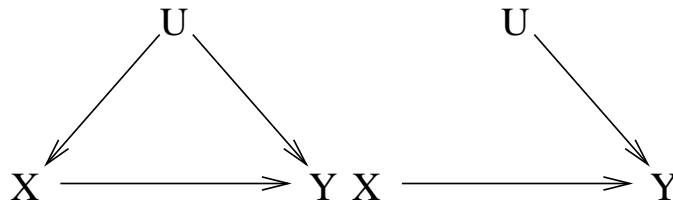


Figure 1: *Left: Causal diagram representing the relationship between observed levels of exposure  $X$  and outcome  $Y$  in the presence of an unknown confounder  $U$ . Right: Their relationship if  $X$  is fixed by design, as in an RCT.*

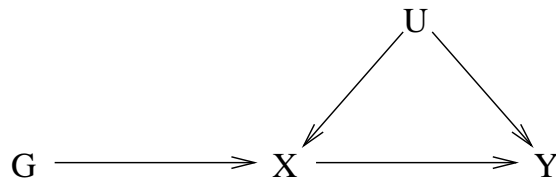


Figure 2: *The causal model commonly assumed to hold in an Instrumental Variable analysis -  $G$  being the IV.*

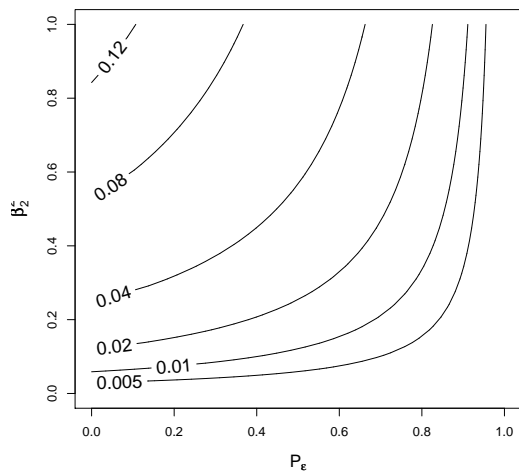


Figure 3:  $\beta_1 \{ \beta_2^2 P_\epsilon \} - \beta_1 \{ \beta_2^2 \}$  under the ZL approximation as a function of  $\beta_2^2$  and  $P_\epsilon$ .

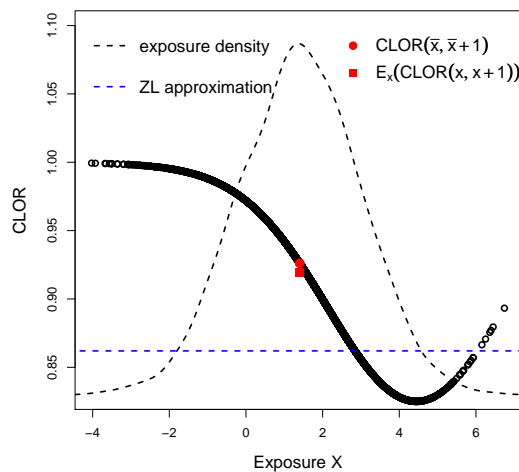


Figure 4: Exposure level  $x$  versus  $CLOR(x, x + 1)$  for a distribution of 5000 exposures generated as in Section 4. Highlighted in red are  $CLOR(\bar{x}, \bar{x} + 1)$  and  $E_x[CLOR(x, x + 1)]$  for this data.

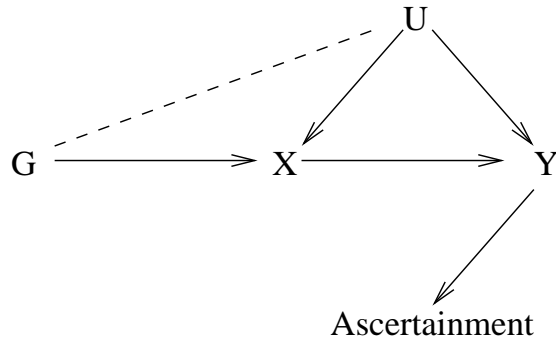


Figure 5: The previous causal graph with added moral edge between the instrumental variable  $G$  and confounder  $U$ , induced by conditioning on the value of  $Y$ .

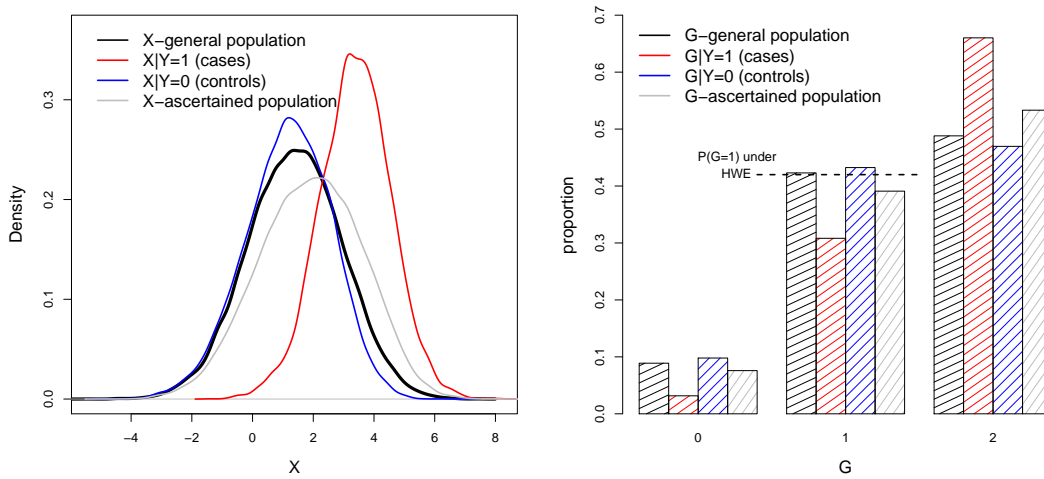


Figure 6: Left: The distribution of  $X$  in (i) the general population (ii) the disease cases (iii) the non-diseased and (iv) the ascertained sample. Right: The distribution of  $G$  in the same subgroups.

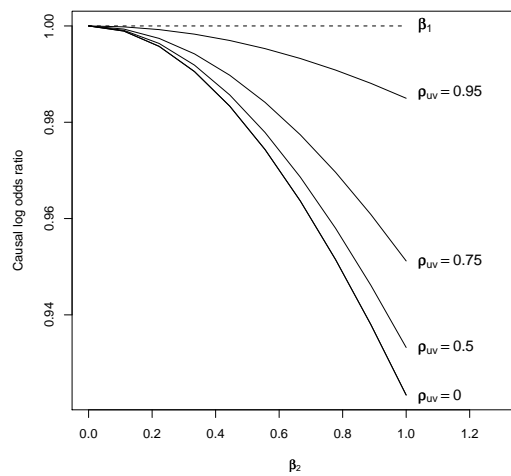


Figure 7: : *Theoretical lower and upper bounds for the adjusted IV and LSMM estimates (under the ZL approximation) when conditioning on a covariate  $v$  that is correlated with  $u$ .*

Parameter Varied	Prevalence $Pr(Y = 1)$	$CJOR(\bar{x}, \bar{x} + 1)$	Standard IV	IV methods: Adjusted IV	mean(variance)coverage	LSSM(1)	LSSM(2)
$\beta_0$			(Other parameter values - ( $\alpha_2 = 1, \beta_2 = 1$ ))				
-6	0.059	0.964	0.755 (0.066)	0.844	0.978 (0.088)	0.958	0.930 (0.087)
-5	0.111	0.932	0.687 (0.034)	0.730	0.956 (0.051)	0.954	0.922 (0.058)
-4	0.187	0.887	0.628 (0.020)	0.554	0.932 (0.033)	0.948	0.882 (0.041)
-3	0.292	0.845	0.593 (0.014)	0.410	0.926 (0.024)	0.918	0.856 (0.032)
-2	0.416	0.828	0.573 (0.011)	0.300	0.923 (0.020)	0.904	0.844 (0.027)
$\beta_2$			(Other parameter values - ( $\alpha_2 = 1, \beta_0 = -5$ ))				
0.00	0.061	1.000	0.886 (0.066)	0.914	1.016 (0.076)	0.952	0.999 (0.069)
0.25	0.071	0.997	0.844 (0.056)	0.881	1.020 (0.069)	0.942	1.003 (0.067)
0.50	0.085	0.987	0.786 (0.047)	0.823	0.997 (0.061)	0.943	0.978 (0.064)
0.75	0.094	0.967	0.744 (0.040)	0.778	0.992 (0.055)	0.949	0.968 (0.061)
1.00	0.107	0.931	0.687 (0.034)	0.724	0.960 (0.050)	0.950	0.926 (0.059)
$\alpha_2$			(Other parameter values - ( $\beta_0 = -5, \beta_2 = 1$ ))				
0.00	0.062	0.935	0.899 (0.066)	0.950	0.949 (0.069)	0.945	0.940 (0.064)
0.25	0.073	0.931	0.847 (0.056)	0.927	0.934 (0.061)	0.944	0.925 (0.062)
0.50	0.085	0.936	0.793 (0.047)	0.884	0.935 (0.057)	0.940	0.920 (0.059)
0.75	0.096	0.930	0.736 (0.040)	0.822	0.940 (0.053)	0.948	0.916 (0.058)
1.00	0.106	0.930	0.680 (0.034)	0.715	0.948 (0.050)	0.953	0.913 (0.058)

Table 1: Performance of the standard IV, adjusted IV and LSMM estimates for the  $CJOR(\bar{x}, \bar{x} + 1)$ , for varying  $\beta_0, \alpha_2$

$\beta_2$ .

Parameter	Prevalence	$CJOR(\bar{x}, \bar{x} + 1)$	Standard IV	IV methods:	mean(	(variance)	coverage				
Varied	$Pr(Y = 1)$			Adjusted IV	LSMM(1)	LSMM(2)					
$\beta_{XU}$											
(Other parameter values - ( $\beta_0 = -5, \alpha_0 = 0, \alpha_1 = 1, \beta_2 = 0.5$ ))											
Normal $u$											
0.00	0.083	0.987	0.798 (0.047)	0.856	1.012 (0.062)	0.961	1.004 (0.061)	0.961	0.994	0.064	0.952
0.25	0.123	1.107	0.720 (0.031)	0.395	1.093 (0.050)	0.958	1.088 (0.051)	0.960	1.043	0.061	0.943
0.50	0.155	1.198	0.663 (0.024)	0.061	1.122 (0.045)	0.930	1.117 (0.048)	0.938	1.026	0.062	0.877
0.75	0.185	1.203	0.615 (0.020)	0.009	1.099 (0.042)	0.915	1.094 (0.045)	0.922	0.961	0.058	0.802
1.00	0.205	1.169	0.572 (0.018)	0.006	1.031 (0.038)	0.882	1.029 (0.042)	0.897	0.876	0.050	0.708
Binary $u$											
0.00	0.102	0.995	0.894 (0.041)	0.910	1.022 (0.047)	0.953	1.017 (0.046)	0.954	1.013	(0.046)	0.947
0.25	0.149	1.137	0.922 (0.029)	0.746	1.171 (0.038)	0.956	1.167 (0.037)	0.958	1.159	(0.040)	0.952
0.50	0.201	1.259	0.921 (0.022)	0.354	1.293 (0.033)	0.946	1.291 (0.032)	0.946	1.272	(0.037)	0.945
0.75	0.256	1.307	0.886 (0.018)	0.096	1.345 (0.029)	0.947	1.343 (0.029)	0.950	1.304	(0.036)	0.946
1.00	0.300	1.263	0.848 (0.015)	0.064	1.341 (0.027)	0.926	1.340 (0.027)	0.927	1.280	(0.035)	0.935

Table 2: Performance of the standard IV, adjusted IV and LSMM estimates for the  $CJOR(\bar{x}, \bar{x} + 1)$  under model misspecification, for varying  $\beta_{XU}$  and confounder distribution.

$\beta_2$	Prospective data estimates		Retrospective data: Av.diff(variance)coverage			
	Alt IV	LSMM (1)	Adjusted IV	exact W	Adjusted IV	crude W
			LSMM(1)	LSMM(1)	Adjusted IV	LSMM(1)
(Other parameter values - ( $\beta_0 = -5, \alpha_0 = 0, \alpha_1 = 1, \beta_1 = 1$ ))						
0.00	1.010	1.010	0.001 (0.015)0.944	0.001 (0.015)0.944	0.000 (0.017)0.946	0.001 (0.017)0.944
0.25	1.009	1.009	0.003 (0.015)0.952	0.003 (0.015)0.952	0.012 (0.017)0.950	0.013 (0.018)0.950
0.50	0.998	0.998	0.003 (0.015)0.960	0.002 (0.015)0.960	0.025 (0.017)0.952	0.029 (0.018)0.948
0.75	0.978	0.978	0.004 (0.015)0.947	0.003 (0.015)0.949	0.041 (0.017)0.939	0.050 (0.018)0.930
1.00	0.944	0.943	0.000 (0.015)0.942	0.000 (0.016)0.945	0.052 (0.017)0.938	0.069 (0.018)0.924

Table 3: Average difference between the adjusted IV and LSMM(1) estimates for the CLOR( $\bar{x}, \bar{x} + 1$ ) based on case-control data versus prospective data, for the different weighting strategies and varying  $\beta_2$ . Average variance and coverage, with respect to CLOR( $\bar{x}, \bar{x} + 1$ ) is also shown.