A weighting method for simultaneous adjustment for confounding and joint exposure-outcome misclassifications

Vries, B.B.L.P. de; Smeden, M. van; Groenwold, R.H.H.

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2021, Vol. 30(2) 473–487
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280220960172
journals.sagepub.com/home/smm
SAGE

A weighting method for simultaneous adjustment for confounding and joint exposure-outcome misclassifications

Bas BL Penning de Vries[1], Maarten van Smeden[1] and Rolf HH Groenwold[1,2]

Abstract

Joint misclassification of exposure and outcome variables can lead to considerable bias in epidemiological studies of causal exposure-outcome effects. In this paper, we present a new maximum likelihood based estimator for marginal causal effects that simultaneously adjusts for confounding and several forms of joint misclassification of the exposure and outcome variables. The proposed method relies on validation data for the construction of weights that account for both sources of bias. The weighting estimator, which is an extension of the outcome misclassification weighting estimator proposed by Gravel and Platt (Weighted estimation for confounded binary outcomes subject to misclassification. *Stat Med* 2018; 37: 425–436), is applied to reinfarction data. Simulation studies were carried out to study its finite sample properties and compare it with methods that do not account for confounding or misclassification. The new estimator showed favourable large sample properties in the simulations. Further research is needed to study the sensitivity of the proposed method and that of alternatives to violations of their assumptions. The implementation of the estimator is facilitated by a new R function (ipwm) in an existing R package (mecor).

Keywords

Causal inference, confounding, inverse probability weighting, joint exposure and outcome misclassification, propensity scores, validation data

1 Introduction

In epidemiological research on causal associations between a particular exposure and a certain outcome, erroneous information on either or both of these variables poses a serious methodological obstacle in making valid inferences. In particular, joint misclassification of exposure and outcome can lead to considerable bias of standard causal effect estimators, with direction and magnitude depending on various factors, including the misclassification mechanism and the direction and magnitude of the true effect.[1–6]

Exposure and outcome misclassification is typically categorised according to two separate properties: whether or not the misclassification is differential and whether or not it is dependent relative to some covariate vector L containing patient characteristics.[1,5] Joint misclassification of exposure and outcome is said to be *nondifferential* if (1) the sensitivity and specificity of exposure classification are constant across all categories of the (true) outcome given L and (2) the sensitivity and specificity of outcome classification are constant across all categories of the (true) exposure given L; otherwise it is *differential*. Misclassification is said to be *independent* if the joint probability of any exposure and outcome classification given any true exposure and outcome categories and L can be factored into the product of the corresponding probabilities for exposure and outcome separately; otherwise, it is

[1]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
[2]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Corresponding author:
Bas B.L. Penning de Vries, Department of Clinical Epidemiology, Leiden University Medical Center, PO Box 9600, Leiden 2300, RC, The Netherlands.
Email: B.B.L.Penning_de_Vries@lumc.nl

dependent. In Dawid's notation,[7] that is, if true exposure level A and true outcome Y are (potentially mis)classified as B and Z, respectively, misclassification is nondifferential if and only if $B \perp\!\!\!\perp Y | A, L$ and $Z \perp\!\!\!\perp A | Y, L$ and independent if and only if $Z \perp\!\!\!\perp B | Y, A, L$.

Epidemiological research hampered by joint misclassification of some type is likely voluminous.[6] Examples of studies affected by exposure and outcome misclassification can be found, for example, in the literature on the causal effects of drug use, which is largely based on routinely collected data, where exposures are typically operationalised on the basis of prescription records and where outcomes are often self-reported.[8–11] In applied epidemiological research, misclassification or some of its potential consequences are often ignored.[12,13] The assertion often made in the discussion of study results that observed measures of association are biased toward the null under nondifferentiality, for example, is not generally true unless additional conditions are presupposed.[2,6]

Methods to adjust for misclassification rely on additional information that can be used to estimate or correct for bias. One potential source of information is validation data obtained through supposedly infallible measurement. Recently, Gravel and Platt proposed an inverse probability weighting (IPW) method to simultaneously address confounding and outcome misclassification by means of internal validation data.[14] Other methods likewise suppose that either the exposure or the outcome is subject to misclassification.[14–17] In what follows, we propose an extension of Gravel and Platt's method to allow for confounding adjustment and joint exposure and outcome misclassification. This flexible estimator allows for the misclassifications to be dependent, differential or both. In Section 2, inverse probability weights for confounding and joint misclassification are introduced through a hypothetical study based on the illustrative example of Gravel and Platt. Section 3 details methods for estimation of the various components of the proposed weights based on validation data. In Section 4, we describe a series of Monte Carlo simulations that were used to study properties of the proposed method in finite samples. We conclude with a summary and discussion of our findings in context of the existing literature.

2 Data distribution for illustration and development of weighting method

We first consider the data and setting described by Gravel and Platt and suppose that Table 1 represents a simple random (i.i.d.) sample from (or that its cell counts are proportional to the respective densities in) the population of interest. This illustration is based on a cohort study on the association between post-myocardial infarction statin use (A) and the one-year risk of reinfarction (Y). In what follows, we will refer to this example as the 'reinfarction example'.

Throughout we take the counterfactual framework for causal inference, formal accounts of which are given for example by Neyman et al.[18–22] The interest, we suppose, lies in estimating $g(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ for some function g, where $Y(0)$ and $Y(1)$ denote the counterfactual outcomes for hypothetical interventions setting A to 0 and 1, respectively. Common choices of g define $g(p_0, p_1) = p_1 - p_0$ (risk difference), $g(p_0, p_1) = p_1/p_0$ (risk ratio), or $g(p_0, p_1) = [p_1/(1-p_1)]/[p_0/(1-p_0)]$ (odds ratio). For our numerical example and simulation studies, we concentrate on the causal marginal odds ratio (OR) in particular, with

$$\mathrm{OR} = g(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)]) = \frac{\mathbb{E}[Y(1)]/(1-\mathbb{E}[Y(1)])}{\mathbb{E}[Y(0)]/(1-\mathbb{E}[Y(0)])} \tag{1}$$

but the results naturally extend to other effect measures.

2.1 No misclassification

Under conditional exchangeability given L (i.e. $(Y(0), Y(1)) \perp\!\!\!\perp A | L$), consistency ($Y(a) = Y$ if $A = a$) and positivity ($\Pr(A = a | L = l) > 0$ for $a = 0, 1$ and all l in the support of L), the mean counterfactuals $E[Y(0)]$ and $E[Y(1)]$ can

Table 1. Cross-classification of the reinfarction data for 33,007 individuals as given by Gravel and Platt.

	L = 0		L = 1	
	A = 0	A = 1	A = 0	A = 1
Y = 0	11602	13116	1302	5363
Y = 1	890	589	49	96

be expressed in terms of 'observables' (meaning, here, variables that would be observed in the absence of measurement error) as follows

$$\mathbb{E}[Y(0)] = \mathbb{E}[WY|A = 0] \quad \text{and} \quad \mathbb{E}[\mathrm{Y}(1)] = \mathbb{E}[\mathrm{WY}|\mathrm{A} = 1]$$

where W denotes the inverse probability of the allocated exposure level A given L (i.e. the inverse propensity score if $A = 1$ and the inverse of the complement of the propensity score if $A = 0$) multiplied by the prevalence of the allocated exposure level A (i.e. $W = \Pr(A)/\Pr(A|L)$; Supplementary Appendix I). We therefore have

$$g(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)]) = g(\mathbb{E}[WY|A = 0], \mathbb{E}[WY|A = 1]) \tag{2}$$

Replacing components of the right-hand side of equation (2) with sample analogues, we obtain the following estimator for the setting where L is binary

$$\begin{aligned} \widehat{\mathrm{OR}} &:= g(\widehat{\mathbb{E}}[\widehat{W}Y|A = 0], \widehat{\mathbb{E}}[\widehat{W}Y|A = 1]) \\ &= \frac{\widehat{\mathbb{E}}[\widehat{W}Y|A = 1]/(1 - \widehat{\mathbb{E}}[\widehat{W}Y|A = 1])}{\widehat{\mathbb{E}}[\widehat{W}Y|A = 0]/(1 - \widehat{\mathbb{E}}[\widehat{W}Y|A = 0])} \\ &= \frac{(\widehat{W}_{10}n_{110} + \widehat{W}_{11}n_{111})/(n_{110} + n_{111} + n_{010} + n_{011} - \widehat{W}_{10}n_{110} - \widehat{W}_{11}n_{111})}{(\widehat{W}_{00}n_{100} + \widehat{W}_{01}n_{101})/(n_{100} + n_{101} + n_{000} + n_{001} - \widehat{W}_{00}n_{100} - \widehat{W}_{01}n_{101})} \end{aligned} \tag{3}$$

where n_{yal} denotes the number of subjects with $Y = y$, $A = a$, $L = l$ and where $\widehat{W}_{al}$ is the product of the proportion of subjects in the sample with $A = a$ and the inverse of the proportion of subjects with $A = a$ among those with $L = l$. For the data in Table 1, we obtain $\widehat{\mathrm{OR}} \approx 0.573$. The corresponding crude odds ratio (i.e. with $\widehat{W} = 1$) is 0.509.

2.2 Joint misclassification

Suppose that rather than observing Y and A we observe Z and B, the misclassified versions of Y and A, respectively. The relation between Z and B on the one hand and Y, A and L on the other can be expressed as follows

$$\Pr(Z = z, B = b|Y = y, A = a, L = l) = (\pi_{byal})^z(1 - \pi_{byal})^{1-z}(\lambda_{yal})^b(1 - \lambda_{yal})^{1-b}$$

for $z, b \in \{0, 1\}$ and all possible realisations y, a, l of Y, A, L, and where $\pi_{byal} = \Pr(Z = 1|B = b, Y = y, A = a, L = l)$ and $\lambda_{yal} = \Pr(B = 1|Y = y, A = a, L = l)$.

To simulate (dependent differential) misclassification in the reinfarction dataset, we use the true positive and false positive rates given in Table 2. The expected cell counts for these rates are given in Table 3.

We redefine the weights in equation (2) as a function of B and L (as per Supplementary Appendix I) such that

$$W = \frac{p(B)\varepsilon_{BL}}{\sum_y \sum_a \pi_{ByaL}(\lambda_{yaL})^B(1 - \lambda_{yaL})^{1-B}(\varepsilon_{aL})^y(1 - \varepsilon_{aL})^{1-y}(\delta_L)^a(1 - \delta_L)^{1-a}} \tag{4}$$

Table 2. True and false positive rates for reinfarction example.

$\pi_{0000} = 0.050$	$\pi_{0001} = 0.020$	$\lambda_{000} = 0.010$
$\pi_{1000} = 0.060$	$\pi_{1001} = 0.108$	$\lambda_{100} = 0.181$
$\pi_{0100} = 0.930$	$\pi_{0101} = 0.806$	$\lambda_{010} = 0.880$
$\pi_{1100} = 0.938$	$\pi_{1101} = 0.692$	$\lambda_{110} = 0.910$
$\pi_{0010} = 0.030$	$\pi_{0011} = 0.109$	$\lambda_{001} = 0.100$
$\pi_{1010} = 0.060$	$\pi_{1011} = 0.050$	$\lambda_{101} = 0.265$
$\pi_{0110} = 0.906$	$\pi_{0111} = 0.765$	$\lambda_{011} = 0.930$
$\pi_{1110} = 0.950$	$\pi_{1111} = 0.861$	$\lambda_{111} = 0.823$

For $b, y, a, l \in \{0, 1\}$, $\lambda_{yal} = \Pr(B = 1|Y = y, A = a, L = l)$ and $\pi_{byal} = \Pr(Z = 1|B = b, Y = y, A = a, L = l)$.

Table 3. Expected cell counts (rounded to integers) for reinfarction example after misclassification was introduced.

	$Z=0$		$Z=1$	
	$B=0$	$B=1$	$B=0$	$B=1$
$Y=0,\ A=0,\ L=0$	10912	109	574	7
$Y=1,\ A=0,\ L=0$	51	10	678	151
$Y=0,\ A=1,\ L=0$	1527	10850	47	693
$Y=1,\ A=1,\ L=0$	5	27	48	509
$Y=0,\ A=0,\ L=1$	1148	116	23	14
$Y=1,\ A=0,\ L=1$	7	4	29	9
$Y=0,\ A=1,\ L=1$	334	4738	41	249
$Y=1,\ A=1,\ L=1$	4	11	13	68

Note: Because of rounding, the sum of all cell entries is 33,006 rather than 33,007, the size of the reinfarction dataset.

where $p(B)$ is the prevalence of level B of the potentially misclassified version of the exposure variable and where $\varepsilon_{al} = \Pr(Y = 1|A = a, L = l)$ and $\delta_l = \Pr(A = 1|L = l)$ for all possible realisations a and l of A and L, respectively. In Supplementary Appendix I, it is shown that

$$\mathbb{E}[Y(0)] = \mathbb{E}[WZ|B = 0] \quad \text{and} \quad \mathbb{E}[\mathrm{Y}(1)]{=}\mathbb{E}[\mathrm{WZ}|\mathrm{B} = 1] \tag{5}$$

which suggests the plug-in estimator

$$\begin{aligned}\widehat{\mathrm{OR}} &:= g(\widehat{\mathbb{E}}[\widehat{W}Z|B = 0], \widehat{\mathbb{E}}[\widehat{W}Z|B = 1]) \\ &= \frac{\widehat{\mathbb{E}}[\widehat{W}Z|B = 1]/(1 - \widehat{\mathbb{E}}[\widehat{W}Z|B = 1])}{\widehat{\mathbb{E}}[\widehat{W}Z|B = 0]/(1 - \widehat{\mathbb{E}}[\widehat{W}Z|B = 0])}\end{aligned} \tag{6}$$

where $\widehat{\mathbb{E}}$ denotes the sample mean operator and $\widehat{W}$ the sample analogue (i.e. consistent estimator) of W in equation (4). For other effect measures (i.e. other choices of g), the same plug-in strategy can be implemented.

In the absence of exposure misclassification, equation (4) reduces to

$$W = \left(\frac{(\delta_L)^A (1 - \delta_L)^{1-A}}{p(A)} \left[\pi_{A0AL} \frac{1 - \varepsilon_{AL}}{\varepsilon_{AL}} + \pi_{A1AL}\right]\right)^{-1} \tag{7}$$

The first term within the round brackets corrects for confounding and represents the propensity score if $A = 1$ or its complemement if $A = 0$ divided by the prevalence of exposure level A. The term within square brackets is a factor that corrects for misclassification in the outcome variable. This correction factor is similar to that proposed by Gravel and Platt.[14] The only difference is that where in equation (7) it does not depend on the fallible measurement Z of Y, Gravel and Platt define different weights for subjects with $Z = 0$. Note, however, that the choice of weights for subjects with $Z = 0$ does not affect the population quantity in equation (5) or the estimator defined by equation (6), because the weights only appear in products with Z, which equal zero if $Z = 0$.

As for the reinfarction example, the odds ratio estimate for the exposure-outcome effect based on inverse probability weighting that assumes absence of exposure or outcome misclassification is 1.120, while the corresponding misclassification naive crude odds ratio is 1.031. Estimation of the population weights W from observables using validation data is discussed in the next section. As shown below, weighting using the proposed weights that account for confounding and outcome and exposure misclassification results in an odds ratio of $\mathrm{OR} = \widehat{\mathrm{OR}} \approx 0.573$. Inference based on equation (7) rather than equation (4), i.e. using Gravel and Platt's method and ignoring misclassification in the exposure but correcting for outcome misclassification, yields an odds ratio estimate of 0.934.

2.3 Parameterisation based on positive and negative predictive values

In the foregoing discussion, the proposed weights were expressed in terms of sensitivity and specificity parameters. The sensitivity and specificity of Z with respect to Y, given (B, A, L), are π_{B1AL} and $1 - \pi_{B0AL}$, respectively.

Similarly, λ_{Y1L} and $1-\lambda_{Y0L}$ reflect the sensitivity and specificity, respectively, with respect to A, conditional on Y and L.

As discussed below, it may be more convenient to choose a parameterisation that is based on (positive and negative) predictive values. Define $\delta^*_l = \Pr(B=1|L=l)$, $\varepsilon^*_{bl} = \Pr(Z=1|B=b, L=l)$, $\lambda^*_{zbl} = \Pr(A=1|Z=z, B=b, L=l)$ and $\pi^*_{azbl} = \Pr(Y=1|A=a, Z=z, B=b, L=l)$. The weights in equation (4) can be rewritten as

$$W = \frac{\sum_y \sum_a \pi^*_{ByaL} (\lambda^*_{yaL})^B (1-\lambda^*_{yaL})^{1-B} (\varepsilon^*_{aL})^y (1-\varepsilon^*_{aL})^{1-y} (\delta^*_L)^a (1-\delta^*_L)^{1-a}}{\sum_y \sum_a (\lambda^*_{yaL})^B (1-\lambda^*_{yaL})^{1-B} (\varepsilon^*_{aL})^y (1-\varepsilon^*_{aL})^{1-y} (\delta^*_L)^a (1-\delta^*_L)^{1-a}} \times \frac{p(B)}{\varepsilon^*_{BL} (\delta^*_L)^B (1-\delta^*_L)^{1-B}} \tag{8}$$

In the absence of exposure misclassification, these weights simplify to

$$W = \frac{p(A)}{(\delta_L)^A (1-\delta_L)^{1-A}} \frac{\varepsilon_{AL}}{\varepsilon^*_{AL}}$$

3 Estimation of weights based on validation data

Estimation of the proposed weights can be done using a number of approaches and we will here consider a maximum likelihood approach that assumes the availability of internal validation data, i.e. that some study participants have their observed exposure or outcome measured by an 'infallible' or 'gold standard' (100% accurate) classifier, and that all participants have the misclassified exposure and outcome variables measured.

3.1 Validation subset inclusion mechanism

Let R_Y be the indicator variable that takes the value of 1 if the outcome is observed (i.e. measured by an infallible classifier) and 0 otherwise. Similarly, define R_A to be the indicator variable that takes the value of 1 if the exposure variable is observed and 0 otherwise. R_Y and R_A reflect which subjects have validation data available on Y and A, respectively. The subset of subjects with validation data on Y need not fully overlap with the subset with validation data on A.

The validation subsets can be approached from the missing data framework of Rubin.[23] Provided that Z, B, L are free of missing values, Rubin's missing at random (MAR) condition is met whenever the vector (R_Y, R_A) is conditionally independent of (Y, A) given (Z, B, L).

3.2 Full likelihood approach based on parameterisation in terms of sensitivities and specificities

Simultaneous estimation of the whole vector of δ, ε, λ and π parameters can be done via maximum likelihood estimation as follows. Assuming i.i.d. observations $(Z_i, B_i, Y_i, A_i, L_i)$ and ignorable missingness in the sense of Rubin[23] (MAR and distinctness), for valid likelihood-based inference it is appropriate to maximise the following log-likelihood over the parameter space of θ, the vector of δ, ε, λ and π parameters

$$\begin{aligned}\ell(\theta) = & \sum_{i:R_{Yi}=R_{Ai}=1} \log f(\theta; Z_i, B_i, Y_i, A_i, L_i) \\ & + \sum_{i:R_{Yi}=1 \wedge R_{Ai}=0} \log \sum_{A_i} f(\theta; Z_i, B_i, Y_i, A_i, L_i) \\ & + \sum_{i:R_{Yi}=0 \wedge R_{Ai}=1} \log \sum_{Y_i} f(\theta; Z_i, B_i, Y_i, A_i, L_i) \\ & + \sum_{i:R_{Yi}=R_{Ai}=0} \log \sum_{Y_i} \sum_{A_i} f(\theta; Z_i, B_i, Y_i, A_i, L_i),\end{aligned}$$

where

$$\begin{aligned} f(\theta; Z_i, B_i, Y_i, A_i, L_i) = {} & (\pi_{B_iY_iA_iL_i})^{Z_i}(1-\pi_{B_iY_iA_iL_i})^{1-Z_i}(\lambda_{Y_iA_iL_i})^{B_i}(1-\lambda_{Y_iA_iL_i})^{1-B_i} \\ & \times (\varepsilon_{A_iL_i})^{Y_i}(1-\varepsilon_{A_iL_i})^{1-Y_i}(\delta_{L_i})^{A_i}(1-\delta_{L_i})^{1-A_i} \end{aligned}$$

Evaluating this log-likelihood involves marginalising over unobserved quantities in the last three terms of $\ell(\theta)$. The log-likelihood equations may become considerably more tractable if we choose a parameterisation of the likelihood that is based on predictive values rather than sensitivities and specificities.

3.3 Full likelihood approach based on parameterisation in terms of predictive values

Inference may alternatively be based on a log-likelihood that is parameterised in terms of the vector θ^* of the δ^*, ε^*, λ^* and π^* parameters, i.e.

$$\begin{aligned} \ell^*(\theta^*) = {} & \sum_{i:R_{Yi}=R_{Ai}=1} \log h(\theta^*; Z_i, B_i, Y_i, A_i, L_i) \\ & + \sum_{i:R_{Yi}=1\wedge R_{Ai}=0} \log \sum_{A_i} h(\theta^*; Z_i, B_i, Y_i, A_i, L_i) \\ & + \sum_{i:R_{Yi}=0\wedge R_{Ai}=1} \log \sum_{Y_i} h(\theta^*; Z_i, B_i, Y_i, A_i, L_i) \\ & + \sum_{i:R_{Yi}=R_{Ai}=0} \log \sum_{Y_i}\sum_{A_i} h(\theta^*; Z_i, B_i, Y_i, A_i, L_i) \end{aligned}$$

where

$$\begin{aligned} h(\theta^*; Z_i, B_i, Y_i, A_i, L_i) = {} & (\pi^*_{A_iZ_iB_iL_i})^{Y_i}(1-\pi^*_{A_iZ_iB_iL_i})^{1-Y_i}(\lambda^*_{Z_iB_iL_i})^{A_i}(1-\lambda^*_{Z_iB_iL_i})^{1-A_i} \\ & \times (\varepsilon^*_{B_iL_i})^{Z_i}(1-\varepsilon^*_{B_iL_i})^{1-Z_i}(\delta^*_{L_i})^{B_i}(1-\delta^*_{L_i})^{1-B_i} \end{aligned}$$

If validation data is available on Y if and only if it is available on A, the complete data log-likelihood ignoring the missing data mechanism can be conveniently expressed as follows

$$\ell^*(\theta^*) = \ell_1^*(\theta^*) + \ell_2^*(\theta^*) + \ell_3^*(\theta^*) + \ell_4^*(\theta^*) \tag{9}$$

with θ^* denoting the vector of δ^*, ε^*, λ^* and π^* parameters and where

$$\begin{aligned} \ell_1^*(\theta^*) &= \sum_{i:R_{Yi}=R_{Ai}=1} Y_i \log(\pi^*_{A_iZ_iB_iL_i}) + (1-Y_i)\log(1-\pi^*_{A_iZ_iB_iL_i}) \\ \ell_2^*(\theta^*) &= \sum_{i:R_{Yi}=R_{Ai}=1} A_i \log(\lambda^*_{Z_iB_iL_i}) + (1-A_i)\log(1-\lambda^*_{Z_iB_iL_i}) \\ \ell_3^*(\theta^*) &= \sum_{i} Z_i \log(\varepsilon^*_{B_iL_i}) + (1-Z_i)\log(1-\varepsilon^*_{B_iL_i}) \\ \ell_4^*(\theta^*) &= \sum_{i} B_i \log(\delta^*_{L_i}) + (1-B_i)\log(1-\delta^*_{L_i}) \end{aligned}$$

Now, assuming distinct parameter spaces for the vectors of π^*, λ^*, ε^*, and δ^* parameters, the parameter values that maximise $\ell^*(\theta^*)$ can be found by separately maximising $\ell_1^*(\theta^*)$ and $\ell_2^*(\theta^*)$ in the validation subset with respect to the π^* and λ^* parameters, respectively, and $\ell_3^*(\theta^*)$ and $\ell_4^*(\theta^*)$ in the entire dataset with respect to ε^* and δ^*. Following Gravel and Platt[14] and Tang et al.,[24] the sum of the first and last two terms are therefore suitably labelled the internal validation and main study log-likelihood, respectively. With this parameterisation, finding the maximum likelihood estimates is readily achieved by taking advantage of standard statistical software.

3.4 Equivalence of likelihood approaches based on different parameterisations

Without restrictions imposed on

$$\theta_l := (\pi_{000l}, \pi_{100l}, \pi_{010l}, \pi_{110l}, \pi_{001l}, \pi_{101l}, \pi_{011l}, \pi_{111l}, \lambda_{00l}, \lambda_{10l}, \lambda_{01l}, \lambda_{11l}, \varepsilon_{0l}, \varepsilon_{1l}, \delta_l) \quad \text{and}$$
$$\theta_l^* := (\pi^*_{000l}, \pi^*_{100l}, \pi^*_{010l}, \pi^*_{110l}, \pi^*_{001l}, \pi^*_{101l}, \pi^*_{011l}, \pi^*_{111l}, \lambda^*_{00l}, \lambda^*_{10l}, \lambda^*_{01l}, \lambda^*_{11l}, \varepsilon^*_{0l}, \varepsilon^*_{1l}, \delta^*_l)$$

other than that $\theta_l, \theta_l^* \in (0,1)^{15}$, it can be shown that the maximum likelihood estimator based on the internal validation design is invariant to its parameterisation (sensitivities/specificities versus positive and negative predictive values). This is because there exists a function mapping every $\theta_l \in (0,1)^{15}$ to a unique $\theta_l^* \in (0,1)^{15}$ and vice versa. Maximising $\ell(\theta)$ with respect to θ is then equivalent to maximising $\ell(\sigma(\theta^*)) \, (= \ell^*(\theta^*))$ with respect to θ^* for some bijection σ such that $\theta = \sigma(\theta^*)$; that is,

$$\underset{\theta}{\operatorname{argmax}} \;\; \ell(\theta) = \sigma\Big(\underset{\theta^*}{\operatorname{argmax}} \;\; \ell(\sigma(\theta^*))\Big).$$

If more restrictions are imposed on θ or θ^*, e.g. if we assume non-saturated logistic models for the components of θ and θ^*, this equivalence no longer holds and the resulting weight estimates may differ depending on the parameterisation.

3.5 Application

For the re-infarction data example, we assume validation data are available according to a MAR mechanism characterised by

$$\Pr(R_Y \;\; = 1 | R_A = s, Z = z, B = b, Y = y, A = a, L = l) = s,$$
$$\Pr(R_A = 1 | Z = z, B = b, Y = y, A = a, L = l) = 0.25 + 0.10b$$

This mechanism assigns validation data to an individual on either both Y and A (30% of all individuals) or neither depending on their realisation of B, the misclassified version of the exposure variable A (Table S.1). Tables S.2 and S.3 (see Supplementary online Appendix) give the likelihood contributions for the parameterisation based on predictive values and the closed-form maximum likelihood expressions, respectively. Maximum likelihood estimates can also be found by fitting to the data the saturated logistic regression models of B and Z on L and (B, L), respectively, and to the validation subset the fully saturated logistic regression models of A and Y on (Z, B, L) and (A, Z, B, L), respectively. Estimated weights are then obtained by plugging in the maximum likelihood estimates into equation (8). As in the complete data setting where we assumed the weights to be known, evaluating equation (6) then yields an odds ratio of $\widehat{\text{OR}} = \text{OR} \approx 0.573$.

4 Simulations

We performed a series of Monte Carlo simulation experiments to illustrate the implementation of the proposed method, to study its finite sample properties and to compare the method to estimators that ignore the presence of confounding or joint exposure and outcome misclassification. All simulations were conducted using R-3.5.0[25] on x86_64-pc-linux-gnu platforms of the high performance computer cluster of Leiden University Medical Center.

4.1 Methods

For all 54 simulation experiments, we generated $n_{\text{sim}} = 1000$ samples of size n according to the data generating mechanisms depicted in the directed acyclic graphs of Figure 1. This multi-step data generating process included generating values on measurement error-free variables, introducing misclassification and allocating individuals validation data. We applied various estimators to each of the simulation samples to yield, for each scenario, an empirical distribution of each point estimator and corresponding precision estimators. These distributions were then summarised into various performance metrics. These metrics include the empirical bias of the estimator on the log-scale (i.e. the mean estimated log-OR minus the target log-OR across the n_{sim} samples), the empirical standard error (SE) of the estimator on the log-scale (i.e. the square root of the mean squared deviation of the estimated log-OR from the mean log-OR), the empirical mean squared error (MSE) (i.e. the sum of the squared

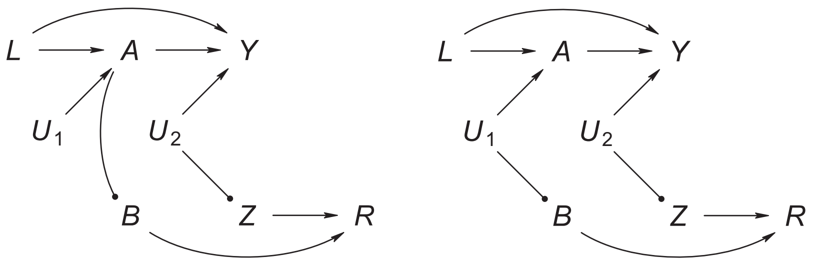

Figure 1. Data structure for scenarios with misclassification on the outcome only (left) or on both the exposure and outcome (right). Bullet arrowheads represent deterministic relationships.

SE and the squared bias), the square root of the mean estimated variance (SSE, sample standard error) and the empirical coverage probability (CP) (i.e. the fraction of simulation runs per scenario where the 95% confidence interval (95% CI) contained the target quantity).

4.1.1 Distribution of measurement error-free variables

Following Gravel and Platt,[14] we consider a setting based on that of "Scenario A" in the work of Setoguchi et al. with slight modifications to the propensity score and outcome models.[26] We consider a fully observed covariate vector $L = (L_0, \ldots, L_{10})$ whose distribution coincides with that of $h(V)$, where $V = (V_1, \ldots, V_{10})$ has the multivariate normal distribution with zero means, unit variances and correlations equal to zero except for the correlations between W_1 and V_5, V_2 and V_6, V_3 and V_8, and V_4 and V_9, which were set to 0.2, 0.9, 0.2, and 0.9, respectively. Function h was defined such that

$$h(V) = (I(V_1 > 0), V_2, I(V_3 > 0), V_4, I(V_5 > 0), I(V_6 > 0), V_7, I(V_8 > 0), I(V_9 > 0), V_{10})$$

Thus, sampling from the distribution of L is equivalent to sampling from the multivariate normal distribution with the given parameter values and dichotomising the first, third, fifth, sixth, eighth and ninth elements.

Next, let U_1 and U_2 be binary variables distributed according to the following logistic models:

$$\operatorname{logit}\Pr(U_1 = 1|L) = \eta_0 \tag{10}$$

$$\operatorname{logit}\Pr(U_2 = 1|L, U_1) = \mu_0 \tag{11}$$

The distribution of the binary exposure variable A was defined according to the model

$$\operatorname{logit}\Pr(A = 1|L, U_1, U_2) = \alpha_0 + \sum_{j=1}^{10} \alpha_j L_j + \alpha_{11} U_1 \tag{12}$$

Letting U_3 be a scalar random variable that is independent of $(A, L_1, \ldots, L_{10}, U_1, U_2)$ and uniformly distributed over the interval $[0, 1]$, we defined the counterfactual outcome $Y(a)$, under the intervention setting A to a, as

$$Y(a) = I\left(U_3 < \operatorname{expit}\left\{ \beta_0 + \gamma a + \sum_{j=1}^{10} \beta_j L_j + \beta_{11} U_2 \right\} \right) \tag{13}$$

With $Y := Y(A)$, the above implies consistency, conditional exchangeability given L and structural positivity.

4.1.2 Misclassification mechanism

For scenarios with joint misclassification, we defined $B = U_1$ and $Z = U_2$, so that the predictive values take a standard logistic form

$$\operatorname{logit}\Pr(Y = 1|A, B, L, Z) = \beta_0 + \gamma A + \sum_{j=1}^{10} \beta_j L_j + \beta_{11} Z \tag{14}$$

$$\operatorname{logit}\Pr(A=1|B,L,Z)=\alpha_0+\sum_{j=1}^{10}\alpha_jL_j+\alpha_{11}B \tag{15}$$

For scenarios without exposure misclassification, we set $\alpha_{11}=0$ and defined $B=A$ and $Z=U_2$, so that

$$\operatorname{logit}\Pr(Y=1|A,B,L,Z)=\beta_0+\gamma A+\sum_{j=1}^{10}\beta_jL_j+\beta_{11}Z \tag{16}$$

$$\operatorname{logit}\Pr(B=1|L,Z)=\alpha_0+\sum_{j=1}^{10}\alpha_jL_j \tag{17}$$

For simplicity, we removed any marginal dependence of Z on the covariates L and U_1 as well as any marginal dependence of U_1 on L (cf. equations (10) and (11)). Although models (10) through (15) take a standard logistic form, they do not imply that the corresponding sensitivities and specificities can be written in the same form. We chose the predictive values rather than the sensitivities and specificities to take a standard logistic form so as to ensure correct model specification in the estimation of the weights in the simulation experiments, in which a likelihood approach based on predictive values was adopted (cf. equation (9)).

4.1.3 Missing data mechanism

For these simulations, we stipulated L, B and Z to be observed for all subjects. We consider scenarios where the dataset can be partitioned into a subset with validation data on all misclassified variables (denoted $R=1$) and a dataset with validation data on neither ($R=0$). That is, we simulated data such that subjects have validation data on both A and Y or neither on A nor on Y. Values for the response indicator R were generated according to the following (MAR) model

$$\begin{aligned}\operatorname{logit}\Pr(R=1|Z,B,Y,A,L)&=\operatorname{logit}\Pr(R=1|Z,B,L)\\&=\xi_0+\xi_1Z+\xi_2B+\xi_3ZB\end{aligned}$$

4.1.4 Scenarios

We initially fixed most parameters of models (12) and (13) at the respective values of "Scenario A" of Setoguchi et al.[26] $\alpha_1=0.8$, $\alpha_2=-0.25$, $\alpha_3=0.6$, $\alpha_4=-0.4$, $\alpha_5=-0.8$, $\alpha_6=-0.5$, $\alpha_7=0.7$, $\alpha_8=0$, $\alpha_9=0$, $\alpha_{10}=0$, $\beta_0=-3.85$, $\beta_1=0.3$, $\beta_2=-0.36$, $\beta_3=-0.73$, $\beta_4=-0.2$, $\beta_5=0$, $\beta_6=0$, $\beta_7=0$, $\beta_8=0.71$, $\beta_9=-0.19$ and $\beta_{10}=0.26$. Parameters η_0 and α_0 were fixed at zero and ξ_1, ξ_2 and ξ_3 at 2, 1 and -1, respectively. The remaining parameters and β_0 were allowed to vary across scenarios as per Table 4.

Scenarios differ by sample size n, the presence of outcome misclassification, potentially misclassified outcome prevalence (via μ_0), the associations between the exposure and outcome on the one hand and the respective misclassified versions on the other (via α_{11} and β_{11}), outcome model intercept β_0, the conditional log-OR γ, or the size of the validation subset (via ξ_0). Based on an iterative Monte Carlo integration approach,[27] we specified γ so as to keep the target marginal log odds ratio at -0.4.

4.1.5 Estimators

We considered five estimators of the OR for the marginal exposure-outcome effect: a crude estimator (labeled Crude) that ignores both confounding and misclassication of any variable, a misclassification naive estimator (labeled PS) that addresses confounding through IPW, complete cases analysis (CCA) in which IPW is applied only to the subset of subjects with validation data, the Gravel and Platt estimator (GP) that ignores exposure misclassification, and the method proposed in this article (labeled IPWM). Both GP and IPWM are implemented using the R function mecor::ipwm,[28,29] which in the simulation settings considered uses iteratively reweighted least squares via the stats::glm function for maximum likelihood estimation. GP coincides with the approach of Gravel and Platt where it concerns point estimation, but they differ in the construction of confidence intervals.

Table 4. Simulation parameter values used in the Monte Carlo studies.

	Exposure						
Scenarios	misclassification	μ_0	α_{11}	β_0	β_{11}	Γ	ξ_0
1a,1b,1c	Absent	−2	0	−3.85	2	−0.431	−1.5
2a,2b,2c	Absent	−3	0	−3.85	2	−0.417	−1.5
3a,3b,3c	Absent	−2	0	−3.85	4	−0.624	−1.5
4a,4b,4c	Absent	−2	0	−3.85	2	−0.431	−2.5
5a,5b,5c	Present	−2	2	−3.85	2	−0.431	−1.5
6a,6b,6c	Present	−3	2	−3.85	2	−0.417	−1.5
7a,7b,7c	Present	−2	4	−3.85	2	−0.431	−1.5
8a,8b,8c	Present	−2	2	−3.85	4	−0.624	−1.5
9a,9b,9c	Present	−2	2	−3.85	2	−0.431	−2.5
10a,10b,10c	Absent	−2	0	−2	2	−0.470	−1.5
11a,11b,11c	Absent	−3	0	−2	2	−0.445	−1.5
12a,12b,12c	Absent	−2	0	−2	4	−0.641	−1.5
13a,13b,13c	Absent	−2	0	−2	2	−0.470	−2.5
14a,14b,14c	Present	−2	2	−2	2	−0.470	−1.5
15a,15b,15c	Present	−3	2	−2	2	−0.445	−1.5
16a,16b,16c	Present	−2	4	−2	2	−0.470	−1.5
17a,17b,17c	Present	−2	2	−2	4	−0.641	−1.5
18a,18b,18c	Present	−2	2	−2	2	−0.470	−2.5

Note: Scenarios indicated with ‘a’ have $n = 10{,}000$, those with ‘b’ have $n = 5000$ and those with ‘c’ have $n = 1000$.

Unlike Gravel and Platt,[14] we used a non-parametric rather than a semi-parametric bootstrap procedure for estimating standard errors and constructing confidence intervals. Semi-parametrically generating response indicators would preferably require modelling of (or making additional assumptions about) the missing data mechanism. In particular, to obtain a bootstrap dataset, we defined the record of a unit as their observed data and response indicators, imposed a uniform distribution across all records in the original dataset, and drew independently as many records from this distribution as the total number of records in the original dataset. For all methods and each original dataset, we drew 1000 bootstrap datasets for variance estimation and the construction of percentile confidence intervals.

All estimators are based on a function of the estimated outcome probability P_1 in the exposed group and the estimated outcome probability P_0 in the unexposed group. However, since P_1 and P_0 may take a value of 0 or 1, the crude odds ratio $[P_1/(1-P_1)]/[P_0/(1-P_0)]$ need not exist. In contrast to what is often (implicitly) done in simulation studies—i.e., studying the properties of the estimators after conditioning on datasets where $[P_1/(1-P_1)]/[P_0/(1-P_0)]$ is defined—we first define $P_1^* = (P_1 s + 1)/(s+2)$ and $P_0^* = (P_0 s + 1)/(s+2)$ for a large positive number s (here set to 10^6) and then regard $[P_1^*/(1-P_1^*)]/[P_0^*/(1-P_0^*)]$ as the estimator of the OR for the exposure-outcome association. This ensures the estimator is always defined and effectively shrinks the outcome probabilities towards 0.5 and the OR towards 1 (online Supplementary Appendix II).

For PS and CCA, we used a logistic regression of B and A, respectively, on covariates L_1 through L_{10} as main effects to estimate the propensity scores. Taking the crude OR for the association between B and Z (PS) or A and Y (CCA) over the data weighted by the reciprocal of the propensity scores provided an estimate of target OR. R code for the methods GP and IPWM is given in online Supplementary Appendix III.

4.2 Results

The treatment assignment mechanism detailed above resulted in average exposure rates ranging from 17% to 51%, whereas average outcome rates ranged from 3% to 22%. Across all simulation studies, the average outcome rate ranged from 6% to 18%. Across all simulation studies with exposure misclassification, exposure and joint misclassification rates ranged from 16% to 33% and from 2% to 6%, respectively. Approximately 16% to 32% of subjects were allocated validation data.

The results on the performance of the various methods in simulations studies 1–9 are provided in Table 5 (see Supplementary Table S.4 for the results on all scenarios).

Table 5. Results for simulation studies 1–9b on the performance of different causal estimators in various scenarios of confounding and misclassification in exposure and outcome.

	Crude					
Scenario	Bias	BSE	MSE	SE	SSE	CP
1b	0.394	0.004	0.169	0.119	0.118	0.122
2b	0.382	0.006	0.179	0.183	0.184	0.492
3b	0.394	0.004	0.169	0.117	0.118	0.116
4b	0.401	0.004	0.174	0.117	0.118	0.102
5b	0.401	0.003	0.169	0.090	0.088	0.007
6b	0.407	0.004	0.183	0.132	0.134	0.133
7b	0.396	0.003	0.164	0.086	0.088	0.009
8b	0.395	0.003	0.164	0.086	0.088	0.005
9b	0.398	0.003	0.166	0.089	0.088	0.005

	PS					
Scenario	Bias	BSE	MSE	SE	SSE	CP
1b	0.392	0.005	0.182	0.168	0.169	0.382
2b	0.379	0.008	0.213	0.264	0.258	0.738
3b	0.389	0.006	0.182	0.175	0.169	0.402
4b	0.389	0.006	0.182	0.176	0.168	0.392
5b	0.402	0.003	0.170	0.090	0.088	0.010
6b	0.407	0.004	0.183	0.131	0.135	0.136
7b	0.396	0.003	0.164	0.086	0.088	0.009
8b	0.395	0.003	0.164	0.086	0.088	0.004
9b	0.398	0.003	0.166	0.089	0.088	0.005

	CCA					
Scenario	Bias	BSE	MSE	SE	SSE	CP
1b	−0.078	0.015	0.226	0.469	0.491	0.899
2b	−0.117	0.019	0.375	0.601	0.900	0.887
3b	−0.020	0.010	0.091	0.301	0.300	0.919
4b	−0.093	0.020	0.407	0.631	1.158	0.899
5b	−0.145	0.009	0.103	0.286	0.286	0.903
6b	−0.109	0.011	0.131	0.345	0.362	0.930
7b	−0.213	0.007	0.101	0.237	0.250	0.865
8b	−0.209	0.006	0.079	0.187	0.186	0.775
9b	−0.175	0.012	0.184	0.392	0.411	0.902

	GP					
Scenario	Bias	BSE	MSE	SE	SSE	CP
1b	−0.036	0.011	0.130	0.359	0.428	0.958
2b	−0.097	0.016	0.265	0.505	0.861	0.938
3b	−0.019	0.007	0.055	0.233	0.240	0.939
4b	−0.045	0.016	0.253	0.501	1.087	0.944
5b	0.269	0.008	0.132	0.244	0.244	0.799
6b	0.280	0.010	0.177	0.314	0.339	0.862
7b	0.134	0.008	0.076	0.241	0.252	0.926
8b	0.259	0.004	0.087	0.140	0.144	0.570
9b	0.263	0.010	0.174	0.325	0.339	0.883

Table 5. Continued.

	IPWM					
Scenario	Bias	BSE	MSE	SE	SSE	CP
1b	−0.036	0.011	0.130	0.359	0.428	0.958
2b	−0.097	0.016	0.265	0.505	0.861	0.938
3b	−0.019	0.007	0.055	0.233	0.240	0.939
4b	−0.045	0.016	0.253	0.501	1.087	0.944
5b	−0.017	0.009	0.082	0.286	0.284	0.942
6b	−0.014	0.011	0.129	0.359	0.386	0.958
7b	0.004	0.008	0.059	0.243	0.261	0.969
8b	−0.004	0.006	0.032	0.180	0.181	0.958
9b	−0.025	0.012	0.141	0.374	0.415	0.956

PS: Propensity score method ignoring misclassification; CCA: complete case analysis; GP: Gravel and Platt estimator ignoring exposure misclassification, consistent with the methodology of Gravel and Platt for point (but not for variance) estimation[14]; IPWM: inverse probability weighting method for confounding and joint exposure and outcome misclassification; BSE: estimated standard error for the bias due to Monte Carlo error; SE: empirical standard error; SSE: sample standard error; CP: empirical coverage probability. In all scenarios, the true marginal log OR (estimand) was −0.4.

As expected, Crude, PS and CCA clearly showed bias with respect to the target log OR of −0.4. The bias associated with restricting the analysis to records with validation data is likely brought on to a large extent by collider stratification, with R acting as the collider here (cf. Figure 1). Both Crude and PS indicated a null effect, as one would anticipate in view of the marginal and L-conditional independence of B and Z implied by the simulation set-up. The empirical coverage probabilities were, although low for both estimators, similar to substantially larger for PS as compared with Crude. Paralleling this is that Crude, whose (implicit) propensity score model is inherently at least as parsimonious, yielded similar to smaller empirical and sample standard errors as compared with PS. With the average fraction of subjects with validation data being as low as 16% (in scenarios with low ξ_0) to 32%, it is not unsurprising that Crude was subject to the largest degree of variability.

The results for the IPWM approach are generally favourable for large samples and in line with its theoretical (large sample) properties. For scenarios with smaller samples (scenarios 1c, 2c and 4c, 6c and 9c in particular), however, we observed considerable bias (see Supplementary Appendix S.4). Comparing CCA with IPWM, we note a strong linear association between the methods in terms of the absolute within-method differences in estimated bias between scenarios of size 10,000 (scenarios labeled 'a') and the respective scenarios of size 1000 (scenarios labeled 'c') (Pearson correlation 0.997). Note that the results for GP and IPWM are identical for scenarios labeled 1–4 and 10–13 since the methods are equivalent in terms of point estimation in the absence of exposure misclassification. In all other scenarios, i.e. scenarios for which GP was not developed, GP performed substantially worse than IPWM. The non-zero, albeit relatively small, systematic deviations of the IPWM point estimates from the target −0.4, notably the estimated bias of −0.097 (scenario 2 b), may be attributable in part to the outcome being rare (with prevalence ranging from 3% to 8% across scenarios labeled 1–9). This is indicated by the superior performance of IPWM in scenarios where the outcome is more prevalent (cf. scenarios labeled 1–9 b versus 10–18 b, which have prevalence up to 22%). A similar observation was made by Gravel and Platt.[14]

The standard errors for GP and IPWM were noticeably higher than those of Crude and PS, which is unsurprising in view of the discrepancies in the number of estimated parameters. As expected, increasing the sample size, the true outcome rate (via β_0) or both led to a decrease in the variability of IPWM (cf. Table 4 and Supplementary Table S.4). However, despite the large discrepancies between SSE and SE for some scenarios, the empirical coverage probabilities of IPWM were close to the nominal level of 0.95, except for scenarios 1c, 2c and 4c, where we observed considerable bias.

5 Discussion

The analysis of epidemiologic data is often complicated by the presence of confounding and misclassifications in exposure and outcome variables. In this paper, we propose a new estimator for estimating a marginal odds-ratio in the presence of confouding and joint misclassification of the exposure and outcome variables. In simulation studies, this weighting estimator showed promising finite sample performance, reducing bias and mean squared error as compared with simpler methods.

The proposed IPWM estimator is an extension of the inverse probability weighting estimator recently proposed by Gravel and Platt (GP) which only addresses the misclassification in the outcome.[14] IPWM and GP are (mathematically) equivalent when the exposure is (assumed to be) measured without misclassification error.

Like the Gravel and Platt approach, IPWM relies on estimates of sensitivity and specificity or positive and negative predictive values for the misclassified variables. In this paper, we used an internal validation approach where a portion of subjects would receive error-free ('gold standard') measurements on either or both the outcome and exposure. However, we anticipate that in some settings the likelihood may not be fully identifiable from the data at hand. In these settings, it may be possible to incorporate external rather than internal information on the misclassification rates, possibly through a Bayesian approach using prior assumptions about misclassification probabilities. When validation data is external, however, it may be necessary to assume misclassification to be independent of covariates L, because external studies seldom consider the same covariates as the main study.[30] External validation approaches also require the assumption that the misclassification parameters targeted in the validation sample are transportable to the main study.

In the absence of internal and external validation data, it is possible to conduct a sensitivity analysis within the weighting framework. Formula (8) for the weights can readily be used in a sensitivity analysis in which the terms describing the distribution of true exposure and outcome variables in relation to the observed data (positive and negative predictive values) serve as sensitivity parameters of the sensitivity analysis. The models for the predictive values can take complex forms, however, thus complicating the analysis and presentation of results.

If internal validation is available, the subjects with validation data need not form a completely random subset. The proposed method, IPWM, was developed under the assumption that validation data allocation occurs in an "ignorable" fashion.[23] In practice, it may be that the researchers have limited control over the validation data allocation mechanism. For instance, it is conceivable that individuals with specific indications (e.g. with a realisation of L, B or Z) are practically ineligible to be assigned a double measurement of the exposure (A and B) and outcome (Y and Z). Further, the estimator also allows for validation subjects to receive either the double exposure or double outcome measurement. We simulated data such that subjects have validation data on both the exposure and outcome variables or on neither. Although this may greatly simplify analysis and enhance efficiency, in practice it is not necessary to assume that this condition holds. An interesting scenario is where subjects have validation data on at most one variable, i.e. on the exposure variable or the outcome variable but not both. In this case, valid estimation would require additional modelling assumptions; for example, the error-free outcome variable cannot then be regressed on the error-free exposure variable.

To accommodate settings where validation data allocation is not completely at random, we deviated from the semi-parametric bootstrap procedure for variance estimation proposed by Gravel and Platt. Instead, the non-parametric procedure we used requires less assumptions regarding the validation subset sampling procedure. The non-parametric procedure showed good performance in our simulations.

Whilst we have discussed under what conditions the proposed method consistently estimates or at least identifies the target quantity, the assumptions may be untenable in particular settings. Particularly, an infallible measurement tool for the exposure and outcome that can be performed on a subset of the data need not always exist. The robustness to deviations of infallibility is an interesting and important direction for further research. This is especially relevant where there exists considerable uncertainty about the tenability of the assumptions that is difficult to incorporate in the analysis. An obvious and flexible alternative to IPWM is to multiply impute missing values including absent measurement error-free variables before implementing IPW (MI + IPW). Although MI + IPW and IPWM may be comparable in terms of their assumptions, it is yet unclear how they behave under assumption violations such as misspecification of the outcome model.

An advantageous property of MI + IPW is that it can easily accommodate missing covariate values. Other alternatives that can accommodate missing covariates were recently developed by Shu and Yi.[31] Their proposed weighting estimators simultaneously addresses confounding, misclassification of the outcome (but not of the exposure) and measurement error on the covariates under a classical additive measurement error model. The methods can be implemented using validation data or repeated measurements and use a simple misclassification model (in which the outcome surrogate is independent of exposure or covariates given the target outcome) that is suitable for performing sensitivity analyses.

Another interesting area for further research is where the researchers do have control over who is referred for further testing by the assumed infallible measurement tool(s). An obvious choice is to adopt a completely at random strategy (simple random sampling). However, other referral (sampling) strategies exist and it is not clear what strategy leads to the most favourable estimator properties for the given setting.

In summary, we have developed an extension to an existing method, to allow for valid estimation of a marginal causal OR in the presence of confounding and a commonly ignored and misunderstood source of bias—joint exposure and outcome misclassification. The R function mecor::ipwm has been made available to facilitate implementation.[28,29]

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: RHHG was funded by the Netherlands Organization for Scientific Research (NWO-Vidi project 917.16.430). The views expressed in this article are those of the authors and not necessarily any funding body.

ORCID iDs

Bas BL Penning de Vries https://orcid.org/0000-0001-9989-7732
Maarten van Smeden https://orcid.org/0000-0002-5529-1541

Supplemental material

Supplemental material for this article is available online.

References

1. Kristensen P. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology* 1992; **3**: 210–215.
2. Brenner H, Savitz DA and Gefeller O. The effects of joint misclassification of exposure and disease on epidemiologic measures of association exposure and disease on epidemiologic measures of association. *J Clin Epidemiol* 1993; **46**: 1195–1202.
3. Vogel C, Brenner H, Pfahlberg A et al. The effects of joint misclassification of exposure and disease on the attributable risk. *Stat Med* 2005; **24**: 1881–1896.
4. Jurek AM, Greenland S and Maldonado G. Brief report: how far from non-differential does exposure or disease misclassification have to be to bias measures of association away from the null?. *Int J Epidemiol* 2008; **37**: 382–385.
5. VanderWeele TJ and Hernán MA. Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. *Am J Epidemiol* 2012; **175**: 1303–1310.
6. Brooks DR, Getz KD, Brennan AT et al. The impact of joint misclassification of exposures and outcomes on the results of epidemiologic research. *Curr Epidemiol Rep* 2018; **5**: 166–174.
7. Dawid A. Conditional independence in statistical theory. *J Royal Stat Soc, Ser B (Methodol)* 1979; 41: 1–31.
8. Marcum ZA, Sevick MA and Handler SM. Medication nonadherence: a diagnosable and treatable medical condition. *JAMA* 2013; **309**: 2105–2106.
9. Culver AL, Ockene IS, Balasubramanian R et al. Statin use and risk of diabetes mellitus in postmenopausal women in the women's health initiative. *Archives Intern Med* 2012; **172**: 144–152.
10. Leong A, Dasgupta K, Bernatsky S, et al. Systematic review and meta-analysis of validation studies on a diabetes case definition from health administrative records. *PloS One* 2013; **8**: e75256.
11. Ni J, Leong A, Dasgupta K, et al. Correcting hazard ratio estimates for outcome misclassification using multiple imputation with internal validation data. *Pharmacoepidemiol Drug Safety* 2017; **26**: 925–934.
12. Jurek AM, Maldonado G, Greenland S, et al. Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *Eur J Epidemiol* 2006; **21**: 871–876.
13. Brakenhoff TB, Mitroiu M, Keogh RH, et al. Measurement error is often neglected in medical literature: a systematic review. *J Clin Epidemiol* 2018; **98**: 89–97.
14. Gravel CA and Platt RW. Weighted estimation for confounded binary outcomes subject to misclassification. *Stat Med* 2018; **37**: 425–436.
15. Babanezhad M, Vansteelandt S and Goetghebeur E. Comparison of causal effect estimators under exposure misclassification. *J Stat Plan Inference* 2010; **140**: 1306–1319.
16. Braun D, Gorfine M, Parmigiani G, et al. Propensity scores with misclassified treatment assignment: a likelihood-based adjustment. *Biostatistics* 2017; **18**: 695–710.
17. Shu D and Yi GY. Causal inference with measurement error in outcomes: Bias analysis and estimation methods. *Stat Meth Med Res* 2019; **28**: 2049–2068.

18. Neyman J, Iwaszkiewicz K and St Kolodziejczyk. Statistical problems in agricultural experimentation. *Suppl J Royal Stat Soc* 1935; **2**: 107–180.
19. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974; **66**: 688–701.
20. Holland P. Statistics in causal inference. *J Am Stat Assoc* 1986; **81**: 945–960.
21. Holland P. Causal inference, path analysis, and recursive structural equations models. *Sociol Methodol* 1988; **18**: 449–484.
22. Pearl J. *Causality: models, reasoning and inference*. New York, NY: Cambridge University Press, 2009.
23. Rubin D. Inference and missing data. *Biometrika* 1976; **63**: 581–592.
24. Tang L, Lyles RH, Ye Y, et al. Extended matrix and inverse matrix methods utilizing internal validation data when both disease and exposure status are misclassified. *Epidemiol Meth* 2013; **2**: 49–66.
25. R Core Team. *R: A Language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2018, https://www.R-project.org/ (accessed 20 July 2020).
26. Setoguchi S, Schneeweiss S, et al. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Safe* 2008; **17**: 546–555.
27. Austin PC and Stafford J. The performance of two data-generation processes for data with specified marginal treatment odds ratios. *Communicat Stat Simulat Computat* 2008; **37**: 1039–1051.
28. Nab L. *mecor: Measurement error corrections*, 2019. R package version 0.1.0. https://github.com/LindaNab/mecor.git (accessed 30 January 2020).
29. Nab L, Groenwold RH, Welsing PM, et al. Measurement error in continuous endpoints in randomised trials: problems and solutions. 2018.
30. Lyles RH, Tang L, Superak HM, et al. Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. *Epidemiology* 2011; **22**: 589.
31. Shu D and Yi GY. Weighted causal inference methods with mismeasured covariates and misclassified outcomes. *Stat Med* 2018; **38**: 1835–1854.