SPi	Journal Code:			Article ID				Dispatch: 05.10.17	CE: Jojelyn H. Timonan
	S	I	M	7	5	2	2	No. of Pages: 12	ME:

Received: 1 August 2016 | Revised: 18 August 2017 | Accepted: 13 September 2017

DOI: 10.1002/sim.7522

RESEARCH ARTICLE

WILEY Statistics in Medicine

Weighted estimation for confounded binary outcomes subject to misclassification

Christopher A. Gravel[1,3] | Robert W. Platt[1,2]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

[2]Department of Pediatrics, McGill University, Montreal, Quebec, Canada

[3]McLaughlin Centre for Population Health Risk Assessment, University of Ottawa, Ottawa, Ontario, Canada

Correspondence
Christopher A. Gravel, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada.
Email: christopher.gravel2@mcgill.ca

In the presence of confounding, the consistency assumption required for identification of causal effects may be violated due to misclassification of the outcome variable. We introduce an inverse probability weighted approach to rebalance covariates across treatment groups while mitigating the influence of differential misclassification bias. First, using a simplified example taken from an administrative health care dataset, we introduce the approach for estimation of the marginal causal odds ratio in a simple setting with the use of internal validation information. We then extend this to the presence of additional covariates and use simulated data to investigate the finite sample properties of the proposed weighted estimators. Estimation of the weights is done using logistic regression with misclassified outcomes, and a bootstrap approach is used for variance estimation.

KEYWORDS

confounded binary data, outcome misclassification bias, propensity score, validation sampling, weighted estimation

1 | INTRODUCTION

Misclassification of a binary outcome can introduce bias in epidemiological studies even in the absence of confounding. In the presence of such misclassification, we cannot assume that the outcome we observe is the realized potential outcome for the observed exposure; in other words, such data violate the consistency assumption for the identification of causal effects.[1] Hence, in this setting, the use of propensity score (PS) weights will be insufficient to consistently estimate the causal effect of interest. Generally, sources of error in the outcome variable are unobservable in the data. For example, in a diagnostic setting, a faulty result from a laboratory test may influence a physician to diagnose a patient incorrectly. Alternatively, extensive testing may produce a temporal delay in diagnosis that can manifest as misclassification of a binary outcome if the final diagnosis is censored (a diagnostic false negative).

Methods to adjust for misclassification have been proposed through the use of additional information that can be used to estimate and mitigate the bias. One potential source of information is an internal validation sample, or a resampling procedure in which a small subset of selected observations have their observed outcomes confirmed by an assumed "infallible" (ie, 100% sensitive and specific) classifier.

A large body of literature exists discussing the impact of, and methods to offset, misclassification of a binary exposure on the estimation of the odds ratio in a 2×2 contingency table using internal validation data. Barron[2] and Marshall[3] proposed so-called matrix methods, in which observed proportions from the validation and main study data are "plugged in" to a set of equations set up in matrix form. These equations are expressions for the unconditional probabilities of the observed, possibly misclassified, outcomes as a function of either sensitivity and specificity (known as the matrix method)

or positive and negative predictive values (known as the inverse matrix method). Greenland[4] derived a delta method approximation for the asymptotic variance of these estimators and noted that the matrix method was inefficient, and introduced an inverse variance weighted estimator based on the observed proportions from both the validated and unvalidated data. Lyles[5] developed a likelihood-based approach parametrized by the positive and negative predictive values and demonstrated that the inverse matrix method produced estimates that are equivalent to the maximum likelihood estimates (MLEs) and are hence efficient. While all of these authors discussed this problem in the context of binary exposure misclassification, the results generalize to outcome misclassification because of the symmetry of the odds ratio.

In the context of logistic regression, Magder and Hughes[6] used maximum likelihood estimation with an expectation-maximization algorithm to incorporate known diagnostic error rates. Lyles et al[7] extended this to unknown error rates through the use of internal validation information in the context of differential outcome misclassification, and Edwards et al[8] used internal validation information to build a multiple imputation approach.

While the assumption that a gold standard outcome exists for a small subset of the data is not always reasonable, it is possible to conceive of a validation procedure that has high predictive value. For instance, there are many scenarios in which the error in the validation procedure can be described as so minimal that it can be ignored, such as the presence of procedure codes in claims data or codes indicating the dispensing of a prescription medication. In other words, a set of covariates might increase the positive predictive value of the validated outcome when recorded in conjunction with a standard diagnostic code such as an International Classification of Diseases code. As an example, Levine et al [9] demonstrated high positive predictive value of certain composite coding definitions for skin and soft tissue infections in a study validating an electronic health record dataset. Given this possible limitation of the internal validation methods previously discussed, an interesting future research goal might be to investigate the extent of residual bias due to imperfect validation procedures. However, this is beyond the scope of this discussion, and we assume the existence of a gold standard classifier.

Use of methods to address outcome misclassification help to restore consistency. However, it is also important to control for confounding in observational data. In this manuscript, we introduce a set of inverse probability (IP) weights that are used in conjunction with inverse PS weights to simultaneously address confounding and outcome misclassification. To motivate this discussion, we introduce an example taken from the UK Clinical Practice Research Datalink (CPRD), which is a database housing detailed medical records of a subset of general practices in the United Kingdom. Herrett et al [10] noted high discordance rates when considering the completeness of the recording of acute myocardial infarction (MI) in multiple linked data sources, namely, the CPRD, the Hospital Episode Statistics (HES) and the Myocardial Ischaemia National Audit Project (a national registry of acute coronary syndromes). Hence, we use the example of investigating post-MI statin use and the 1-year risk of a second MI (reinfarction) in these data as it possesses the bias inducing data characteristics of interest. Due to a lack of available validation information, we introduce differential misclassification at chosen error rates.

In Section 2, we introduce the problem and characterize the bias using the simplified reinfarction example. In Section 3, we present a derivation of the proposed weights and demonstrate that appropriate application of these weights to the raw data (in conjunction with PS weights) will produce a consistent estimator of the marginal causal odds ratio. In Section 4, we discuss methods for estimation of the weights, first in the simple three-way contingency table setting discussed in Section 2, followed by extension to multiple covariates using the methods outlined in Lyles et al.[7] Finally, in Section 5, we conduct a number of simulation studies in both settings using the data generation approach outlined in Setoguchi et al [11] to investigate the finite sample properties of the proposed method for estimation of the marginal causal odds ratio.

2 | CHARACTERIZATION OF BIAS DUE TO CONFOUNDING AND OUTCOME MISCLASSIFICATION

To illustrate the problem of outcome misclassification, we constructed a retrospective cohort study using the linked CPRD and HES data to explore the association between post-MI statin use and the 1-year risk of reinfarction. The study population was made up of individuals 18 years and older, with at least a year of history prior to their first recorded diagnosis of an MI in the CPRD or HES data. The study period lasted from April 1, 1998, to March 31, 2012; individuals were assessed for the occurrence of an MI using Read codes in the CPRD and ICD-10 codes in the HES data. In the 30 days after the first recorded MI (assumed to be classified accurately for demonstrative purposes), statin exposure, denoted as A, was assessed as the presence of any recorded prescription in the interval ($A = 1$, 0 otherwise). We excluded individuals who had a record of an MI in this interval as well as individuals who had a record of a statin prescription in the 365 days

prior to their first MI. After the 30-day interval, we followed individuals until the next record of an MI in the HES data (the outcome of interest), a year had passed or the end of the study interval. We chose to search for the outcome in the HES data exclusively due to the presumptive high positive predictive value of MI records in hospital-based data sources. However, it is also likely that MIs will be right censored due to fatalities occurring prior to hospital admission, creating a systemic form of outcome misclassification (false negatives). Motivated by these assumptions, we introduced misclassification error through a fixed set of values as described in Table 3. We denoted the true occurrence of a reinfarction in the follow-up period of the study as $Y = 1$, 0 otherwise, and the possibly misclassified version of Y will be denoted as Y^*, where $Y^* = 1$ if a reinfarction is recorded in the HES data during study interval, 0 otherwise.

While many covariates exist in the CPRD, for demonstrative purposes, we selected a single binary potential confounder, namely, a recording of a coronary revascularization within a year prior to the first MI. The presence of this code is indicative of a previous cardiovascular problem, which may influence the decision to prescribe statins and could heighten the risk of an MI. We denoted this covariate as L, where $L = 1$ if the code is present in this interval, 0 otherwise. In Table 1, we present the number of patients with a previous coronary revascularization in either exposure group at the time of their first MI. Note that the final cohort had $n = 33007$ subjects with 19164 exposed to statins. This study was approved by the Independent Scientific Advisory Committee for MHRA database research (protocol number 14018A2) and the Research Ethics Board of the Jewish General Hospital in Montreal, Canada. T1

To characterize the potential bias incurred from confounding and binary outcome misclassification, we treated the observed reinfarction data as the truth and artificially misclassified the outcome at chosen error rates. In Table 2, we present the cross-classification of the reinfarction data by outcome, Y, exposure, A, and potential confounder, L, where N_{yal} denotes the number of individuals with $\{Y = y, A = a, L = l\}, y, a, l = 0, 1$. Directly from Table 2, we computed the YA (crude associational) odds ratio using the correctly classified outcome, $OR_{\text{YA}} = \frac{(N_{111}+N_{110})(N_{001}+N_{000})}{(N_{011}+N_{010})(N_{101}+N_{100})} = 0.50941$. T2

We are interested in estimation of the marginal causal odds ratio, which can be written as

$$OR = \frac{P[Y^{a=1} = 1]P[Y^{a=0} = 0]}{P[Y^{a=1} = 0]P[Y^{a=0} = 1]}, \tag{1}$$

where Y^a denotes the potential outcome, or what the outcome would be under either treatment arm, $a = 0, 1$. In our example, we assumed that we lacked marginal exchangeability; however, we made the assumption that L was the only confounder and hence conditional exchangeability, $Y^a \coprod A|L$, and positivity were both satisfied.

TABLE 1 Number of patients with previous coronary revascularization at baseline exposed or unexposed to statins in the 30 days after first myocardial infarction

Characteristic	Statin	No Statin
Cohort Size	19 164 (58%)	13 843 (42%)
Previous Revascularization	5459 (16.5%)	1351 (4.1%)

TABLE 2 Cross-classification of the reinfarction data for 25 881 individuals

	$L = 1$		$L = 0$	
	$A = 1$	$A = 0$	$A = 1$	$A = 0$
$Y = 1$	$N_{111} = 96$	$N_{101} = 49$	$N_{110} = 589$	$N_{100} = 890$
$Y = 0$	$N_{011} = 5363$	$N_{001} = 1302$	$N_{010} = 13116$	$N_{000} = 11602$

TABLE 3 Chosen true and false positive rates for all y, a, l groups in the reinfarction data

TP Rate	FP Rate
$\theta_{111} = 0.9$	$\theta_{011} = 0.02$
$\theta_{101} = 0.92$	$\theta_{001} = 0.05$
$\theta_{110} = 0.85$	$\theta_{010} = 0.01$
$\theta_{100} = 0.87$	$\theta_{000} = 0.03$

Abbreviations: FP, false positive; TP, true positive.

As noted in Hernán and Robins,[1] PS weights can be used to create a pseudo-population in which marginal balance is achieved. Returning to the example, the marginal causal odds ratio was equivalent to

$$OR = OR_{\text{PS}} = \frac{\left(\frac{N_{111}}{\psi_1} + \frac{N_{110}}{\psi_0}\right)\left(\frac{N_{001}}{1-\psi_1} + \frac{N_{000}}{1-\psi_0}\right)}{\left(\frac{N_{011}}{\psi_1} + \frac{N_{010}}{\psi_0}\right)\left(\frac{N_{101}}{1-\psi_1} + \frac{N_{100}}{1-\psi_0}\right)} = 0.57329, \tag{2}$$

where $\psi_l = P(A = 1|L = l)$ is the PS and $\psi_1 = 0.80162$ and $\psi_0 = 0.52315$ were computed from Table 2.

To incorporate differential outcome misclassification as discussed in Section 1, we allowed the error rates to depend on both treatment assignment, A, and the potential confounder, L. Within each (y, a, l) subgroup, we denote the misclassification probabilities using the following notation:

$$\theta_{yal} = P(Y^* = 1|Y = y, A = a, L = l), \tag{3}$$

for $y, a, l = 0, 1$. Note that these can be easily extended to a vector of covariate information, as will be explored in Section 4, using the notation $\theta_{ya}(l_i)$ where l_i denotes the covariate vector for the ith subject. The relationship between the impact of outcome misclassification on estimation of the YA odds ratio and the magnitude of the misclassification cannot be easily described as the estimand based on the observed, possibly misclassified data is,

$$OR^*_{\text{YA}} = \frac{P(Y^* = 1|A = 1)P(Y^* = 0|A = 0)}{P(Y^* = 0|A = 1)P(Y^* = 1|A = 0)},$$

where

$$P(Y^* = 1|A = a) = \sum_{l=0,1} P(Y^* = 1|A = a, L = l)P(L = l),$$

and

$$\begin{aligned} P(Y^* = 1|A = a, L = l) = {} & P(Y^* = 1|Y = 1, A = a, L = l)P(Y = 1|A = a, L = l) \\ & + P(Y^* = 1|Y = 0, A = a, L = l)P(Y = 0|A = a, L = l). \end{aligned}$$

To simulate differential outcome misclassification in our simple example, we selected 8t values representing the rates of true positives and false positives in each (a, l) subgroup. These rates are described below in Table 3.

The possibility for right censoring due to fatalities prior to arrival at a hospital motivated the choice of larger false negative rates. We assumed that the false positive rates would be lower given the expected accuracy of inpatient hospital records. The misclassified data is presented in Table 4. Using the possibly misclassified outcome, Y^*, we computed the target odds ratios, first ignoring the covariate information, denoted as OR^*_{YA}, followed by the use of only the PS weights, denoted as OR^*_{PS}.

Computation of the target YA odds ratio from Table 4 without the use of IP weights produced

$$OR^*_{\text{YA}} = \frac{(n_{111} + n_{110})(n_{001} + n_{000})}{(n_{011} + n_{010})(n_{101} + n_{100})} = 0.46049, \tag{4}$$

and with PS weighting while continuing to ignore the impact of misclassification yielded

$$OR^*_{\text{PS}} = \frac{\left(\frac{n_{111}}{\psi_1} + \frac{n_{110}}{\psi_0}\right)\left(\frac{n_{001}}{1-\psi_1} + \frac{n_{000}}{1-\psi_0}\right)}{\left(\frac{n_{011}}{\psi_1} + \frac{n_{010}}{\psi_0}\right)\left(\frac{n_{101}}{1-\psi_1} + \frac{n_{100}}{1-\psi_0}\right)} = 0.47529. \tag{5}$$

Recall that the target marginal causal odds ratio was 0.57329; hence, both of the target odds ratios computed using the misclassified observed data in Table 4 are biased. In the next section, we present a weighted approach that can be combined with PSs to offset the bias incurred by both outcome misclassification and confounding.

TABLE 4 Observed reinfarction data with outcome misclassification

	$L = 1$		$L = 0$	
	$A = 1$	$A = 0$	$A = 1$	$A = 0$
$Y^* = 1$	$n_{111} = 193$	$n_{101} = 110$	$n_{110} = 632$	$n_{100} = 1122$
$Y^* = 0$	$n_{011} = 5266$	$n_{001} = 1241$	$n_{010} = 13073$	$n_{000} = 11370$

3 | DEVELOPMENT OF THE WEIGHTED APPROACH TO ADJUST FOR OUTCOME MISCLASSIFICATION BIAS

To design a weighted approach to mitigate misclassification bias, we will revisit the concept of IP weighting in the setting of a binary point treatment. Observed realizations of the joint distribution of the data generating process are not necessarily exchangeable, and IP weighting rectifies this by simulating a pseudo-population, which, under certain assumptions,[12,13] allows for the estimation of causal effects. However, in the presence of outcome misclassification, the observed data continues to remain insufficient for identification of these effects due to violations of the consistency assumption, or that the observed outcome is the counterfactual outcome for the given treatment.

To simulate exchangeability, the pseudo-population is created by intervening (see Pearl,[14, ch.3]) on the data generation process through an artificial assignment of each treatment status to everyone under study. However, as Y (the version of the outcome consistent for Y^A) may be measured with error, we are only able to observe information on Y^*. Specifically, the IP weighted observed data is based on $\frac{P(Y^*=y^*,A=a,L=l)}{P(A=a|L=l)}$, which may not equal $\frac{P(Y=y,A=a,L=l)}{P(A=a|L=l)}$.

To restore equality when the observed data are a realization of the version of the joint distribution based on Y^* instead of Y, we can characterize the relationship between these 2 joint distributions as

$$\begin{aligned} P(Y^*=1,A=a,L=l) &= P(Y^*=1|Y=1,A=a,L=l)P(Y=1,A=a,L=l) \\ &\quad + P(Y^*=1|Y=0,A=a,L=l)P(Y=0,A=a,L=l) \\ &= \left[\theta_{1al} + \theta_{0al}\frac{1-\pi_{al}}{\pi_{al}}\right] P(Y=1,A=a,L=l), \end{aligned} \tag{6}$$

where $\pi_{al} = P(Y=1|A=a,L=l)$ and a similar rationale can be used for observations with $Y^*=0$. Since the conditional distributions relating Y^* to A and L will be invariant to the application of the treatment intervention, we note that estimation of causal effects is possible in the pseudo-population simulated by $\frac{P(Y^*=1,A=a,L=l)}{\psi_l\left[\theta_{1al}+\theta_{0al}\frac{1-\pi_{al}}{\pi_{al}}\right]}$ and $\frac{P(Y^*=0,A=a,L=l)}{\psi_l\left[(1-\theta_{1al})\frac{\pi_{al}}{1-\pi_{al}}+1-\theta_{0al}\right]}$. We formally denote the proposed weights associated with misclassification as

$$\begin{aligned} W_{y^*=1,a,l} &= \theta_{1al} + \theta_{0al}\frac{1-\pi_{al}}{\pi_{al}} \\ W_{y^*=0,a,l} &= (1-\theta_{1al})\frac{\pi_{al}}{1-\pi_{al}} + 1 - \theta_{0al}, \end{aligned} \tag{7}$$

for $a,l=0,1$. Note that extension to a vector of covariates can be done by writing the π and θ-parameters as $\pi_a(l_i)$ and $\theta_{ya}(l_i)$ where l_i is the ith individual's covariate vector.

Returning to the reinfarction example of Section 2, we applied these adjustments to OR^*_{YA}, Equation 4, using the specified values for the θ-parameters in Table 3 and the values computed from Table 2 for the ψ and π-parameters: $\psi_1 = 0.80162$, $\psi_0 = 0.52315$, $\pi_{11} = 96/5459$, $\pi_{01} = 49/1351$, $\pi_{10} = 589/13705$, and $\pi_{00} = 890/12492$. The resulting value was

$$OR^*_{\text{W}} = \frac{\left(\frac{n_{111}}{\psi_1 W_{111}} + \frac{n_{110}}{\psi_0 W_{110}}\right)\left(\frac{n_{001}}{(1-\psi_1)W_{001}} + \frac{n_{000}}{(1-\psi_0)W_{000}}\right)}{\left(\frac{n_{011}}{\psi_1 W_{011}} + \frac{n_{010}}{\psi_0 W_{010}}\right)\left(\frac{n_{101}}{(1-\psi_1)W_{101}} + \frac{n_{100}}{(1-\psi_0)W_{100}}\right)} = 0.57355, \tag{8}$$

which demonstrates that this weighted odds ratio was equivalent to the target marginal causal odds ratio of interest, Equation 1.

4 | ESTIMATION OF THE WEIGHTS

In Sections 2 and 3, we introduced the proposed methodology using a simplified example taken from the linked CPRD and HES data. As noted in Section 1, estimation of the π and θ-parameters can be done in a number of ways and we will consider using a maximum likelihood approach that incorporates the use of internal validation data. Estimation of these parameters has been studied extensively, and closed form expressions for the MLEs can be written in our notation as [5]

$$\hat{\pi}_{al} = \frac{m_{1al}n_{0al}y_{0al} + m_{0al}n_{1al}y_{1al}}{m_{0al}m_{1al}(n_{1al} + n_{0al})}$$

$$\hat{\theta}_{1al} = \frac{m_{0al}y_{1al}n_{1al}}{m_{0al}y_{1al}n_{1al} + m_{1al}y_{0al}n_{0al}} \tag{9}$$

$$\hat{\theta}_{0al} = \frac{m_{0al}n_{1al}(m_{1al} - y_{1al})}{m_{0al}(m_{1al} - y_{1al})n_{1al} + m_{1al}(m_{0al} - y_{0al})n_{0al}},$$

where n_{y^*al} denotes the original sample size, m_{y^*al} denotes the validation sample size, and y_{y^*al} denotes the number of observations from the validation sample updated as a positive in the y^*, a, lth group. Next, the PSs, ψ_l, $l = 0, 1$, can be estimated using the respective proportions observed in the original sample since A and L are assumed to be measured accurately. Estimating the variance of $\log \widehat{OR}^*_{\mathrm{W}}$ is not straightforward, and deriving a closed-form expression is not possible. Each term is the log of a complicated expression of 6 random quantities, for example, the first term is $\log\left(\frac{n_{111}}{\hat{\psi}_1 \tilde{W}_{111}} + \frac{n_{110}}{\hat{\psi}_0 \tilde{W}_{110}}\right)$. Hence, we recommend a bootstrap estimator for the variance and will expand upon this later in the section.

We can extend this to allow for the presence of multiple covariates and describe the method presented in Lyles et al [7] in which they discussed the use of validation data to correct for the bias incurred from outcome misclassification in logistic regression. Both internal and external validation approaches were addressed; however, the authors noted that the use of external information required the additional assumption that the processes generating misclassification were similar in both the validation and original data (ie, transportability). Hence, internal validation is preferable, and we assume the ability to gather such data for a subset of the observations in the original study. It should also be noted that alternate approaches exist for estimation of the π and θ-parameters.[8,15-17]

Lyles et al[7] presented a likelihood combining information from the original data in which all observations are measured by the error prone diagnostic tool and from a subset that are measured a second time by an infallible diagnostic tool. They referred to the set of observations measured once as making up the "main study" indexed by $i = 1, \ldots, n_m$ and the observations measured twice as making up the "validation study," indexed by $j = 1, \ldots, m$, $n = n_m + m$. The main study likelihood, in our notation, is

$$L_M = \prod_{a=0,1}\prod_{i=1}^{n_m}\left[\pi_a(l_i)\theta_{1a}(l_i) + (1 - \pi_a(l_i))\theta_{0a}(l_i)\right]^{y_a^*}\left[1 - \pi_a(l_i)\theta_{1a}(l_i) - (1 - \pi_a(l_i))\theta_{0a}(l_i)\right]^{1-y_a^*}, \tag{10}$$

where l_i is the ith observation's covariate vector for $i = 1, \ldots, n_m$ and $y^*, a = 0, 1$.

The validation study likelihood is

$$L_V = \prod_{a=0,1}\prod_{j=1}^{m}\left[\theta_{1a}(l_j)\pi_a(l_j)\right]^{y_j y_j^*}\left[\theta_{0a}(l_j)(1 - \pi_a(l_j))\right]^{(1-y_j)y_j^*}\left[(1 - \theta_{1a}(l_j))\pi_a(l_j)\right]^{y_j(1-y_j^*)}$$
$$\times\left[(1 - \theta_{0a}(l_j))(1 - \pi_a(l_j))\right]^{(1-y_j)(1-y_j^*)},$$

where m denotes the chosen validation sample size and y_j is the validated outcome for the jth individual, $j = 1, \ldots, m$. Note that sensitivity and specificity, which may depend on exposure and covariates under differential error, are analogous to $\theta_{1a}(l_i)$ and $1 - \theta_{0a}(l_i)$, respectively, in our notation.

To model the θ-parameters, we propose the following set of logistic models:

$$\theta_{1A} = P(Y^* = 1|Y = 1, A) = \frac{\exp(\kappa_{10} + \kappa_{11}A)}{1 + \exp(\kappa_{10} + \kappa_{11}A)}$$
$$\theta_{0A} = P(Y^* = 1|Y = 0, A) = \frac{\exp(\kappa_{00} + \kappa_{01}A)}{1 + \exp(\kappa_{00} + \kappa_{01}A)}. \tag{11}$$

Note that, for simplicity, we remove any dependency on the additional covariates such as a previous history of coronary revascularization in the reinfarction example. However, if one wishes to model the misclassification rates as a function of K-dimensional vector of additional covariates, the models may be extended to $\frac{\exp(\kappa_{y0} + \kappa_{y1}A + \sum_{k=2}^{K+1}\kappa_{yk}L_{k-1})}{1+\exp(\kappa_{y0} + \kappa_{y1}A + \sum_{k=2}^{K+1}\kappa_{yk}L_{k-1})}$ for $y = 0, 1$.

Next, to model the π-parameters, we again choose a logistic model

$$\pi_{_A} = P(Y = 1|A, L) = \frac{\exp\left(\beta_0 + \beta_{\mathrm{treat}}A + \sum_{k=1}^{K}\beta_k L_k\right)}{1 + \exp\left(\beta_0 + \beta_{\mathrm{treat}}A + \sum_{k=1}^{K}\beta_k L_k\right)},$$

and we denote the coefficient to treatment, A, as β_{treat}. Note that $\exp(\beta_{\text{treat}})$ is a conditional odds ratio characterizing the effect of treatment, while our interest is in estimation of the marginal causal effect of treatment.

Maximization of the likelihood function $L = L_M \times L_V$ can produce the needed plug-in estimates to compute the proposed weights in Equation 7. Propensity scores must be estimated as well, and we use a logistic model applied to the whole dataset to attain the predicted values since A and L are assumed to be measured without error,

$$\psi_L = P(A = 1|L) = \frac{\exp\left(\alpha_0 + \sum_{k=1}^{K} \alpha_k L_k\right)}{1 + \exp\left(\alpha_0 + \sum_{k=1}^{K} \alpha_k L_k\right)}.$$

The final step is to apply the weights to the observed, possibly misclassified, version of the outcome, Y^*, and to compute the treatment effect of interest using the following logistic model:

$$P(Y^* = 1|A) = \frac{\exp(\gamma_0 + \gamma_{\text{treat}} A)}{1 + \exp(\gamma_0 + \gamma_{\text{treat}} A)}. \tag{12}$$

The treatment effect in (12), $\exp(\gamma_{\text{treat}})$, is equivalent to the marginal causal odds ratio, Equation 1.

Estimation of the variance of $\hat{\gamma}_{\text{treat}}$ (and $\log \widehat{OR}^*_{\text{W}}$ in the contingency table setting) is done using the bootstrap.[18] A nonparametric bootstrap may fail as it is likely that the amount of misclassification error may be small in many practical settings. For a given treatment group, as the error rates approach zero, the probability of drawing a validation sample with no observed error increases. This implies that for some iterations of the bootstrap, no variability will be present in the treatment group in question, introducing bias in the estimation of the variance. Hence, we use a nonparametric bootstrap only for the covariate vector, L, and the exposure variable, A, as they are assumed to be measured without error. For the outcome variables, a parametric bootstrap is justified, given that we are only dealing with the simple case of binary outcomes. Provided that the same model is used for the generation of bootstrapped outcomes as is used to define the π and θ-parameters, this variance estimation procedure will work.

5 | SIMULATION STUDY

In this section, we present the results of 2 sets of simulation studies designed to numerically investigate the finite sample properties of the proposed weighted estimators of the marginal causal odds ratio and to observe the behaviour of the estimators that ignore the presence of confounding and outcome misclassification. The first set investigates the maximum likelihood approach described in the beginning of Section 4 for $\log \widehat{OR}^*_{\text{W}}$ in a three-way contingency table, similar to the setting described in the example of Section 2. The second set of studies extends to the presence of multiple covariates to explore the finite sample properties of the estimator of γ_{treat} described at the end of Section 4. Recall that both of these are equivalent to the log of the marginal causal odds ratio defined in Equation 1.

For the first set of simulations, we generated data using the following approach. For a given original sample size, n, we generated the "true" subgroup sample sizes from the original data, N_{yal}, by conducting a series of draws from a binomial random generator. First, we subdivided the n observations into l subgroups with specified probability $P(L = 1)$, and for each of these groups, we generated the a, l treatment subgroups with specified probability ψ_l. The target subgroup counts, N_{yal}, are generated with probability π_{al} from each a, l subgroup. Misclassification is generated by another set of binomial draws where the correctly classified observations are generated from $CC_{yal} \sim \text{binomial}(N_{yal}, \theta_{yal})$ and the corresponding misclassified observations are $MC_{yal} = N_{yal} - CC_{yal}$. Finally, we generated the observed data as described in Table 4, $n_{y^*al} = CC_{yal} + MC_{1-y,al}, y, a, l = 0, 1$.

We generated from 9 sets of simulation parameters as noted in Table 5. T5

For all parameter settings, we generated an original sample size of $n = 10000$, and to observe the impact of differing validation sample sizes, we considered $m = 500, 1000$, and 1500. We selected a large value for n purposefully to avoid simulating y, a, l-subgroup sample sizes that are too low to observe any classification error when the target error rates are small. For each set of parameters in Table 5, we first computed the target marginal causal log odds ratio, $\log OR$, using the following expression that will be equivalent to Equation 1,

$$OR = \frac{\sum_{l=0,1} P(Y = 1|A = 1, L = l)P(L = l)\sum_{l=0,1} P(Y = 0|A = 0, L = l)P(L = l)}{\sum_{l=0,1} P(Y = 0|A = 1, L = l)P(L = l)\sum_{l=0,1} P(Y = 1|A = 0, L = l)P(L = l)}, \tag{13}$$

where the equality holds as we assume conditional exchangeability, $Y^a \coprod A|L$.[1]

TABLE 5 Target parameters used in the first set of Monte Carlo simulation studies to investigate log $\widehat{OR}^{*}_{W}$

Index	$P(L=1)$	ψ_1	ψ_0	π_{11}	π_{01}	π_{10}	π_{00}	θ_{111}	θ_{011}	θ_{101}	θ_{001}	θ_{110}	θ_{010}	θ_{100}	θ_{000}
1	0.3	0.3	0.6	0.2	0.4	0.15	0.1	0.99	0.03	0.96	0.01	0.91	0.04	0.98	0.02
2	0.3	0.3	0.6	0.2	0.4	0.15	0.1	0.95	0.1	0.9	0.08	0.99	0.01	0.99	0.01
3	0.3	0.3	0.6	0.2	0.4	0.15	0.1	0.99	0.01	0.99	0.01	0.95	0.1	0.9	0.08
4	0.3	0.3	0.6	0.4	0.25	0.333	0.15	0.99	0.03	0.96	0.01	0.91	0.04	0.98	0.02
5	0.3	0.3	0.6	0.4	0.25	0.333	0.15	0.95	0.1	0.9	0.08	0.99	0.01	0.99	0.01
6	0.3	0.3	0.6	0.4	0.25	0.333	0.15	0.99	0.01	0.99	0.01	0.95	0.1	0.9	0.08
7	0.8	0.65	0.25	0.4	0.25	0.333	0.15	0.99	0.03	0.96	0.01	0.91	0.04	0.98	0.02
8	0.8	0.65	0.25	0.4	0.25	0.333	0.15	0.95	0.1	0.9	0.08	0.99	0.01	0.99	0.01
9	0.8	0.65	0.25	0.4	0.25	0.333	0.15	0.99	0.01	0.99	0.01	0.95	0.1	0.9	0.08

TABLE 6 Results for simulation studies investigating $\widehat{OR}^{*}_{W}$ with original sample size N = 10 000

Index	log *OR*	$\text{Bias}_{\text{crude}}$	SSE	ASE	CP	Bias_{PS}	SSE	ASE	CP	Bias_{W}	SSE	BSE	CP
a) Drawing a validation sample of size M = 500													
1	−0.171	−0.160	0.048	0.050	0.084	0.157	0.049	0.052	0.575	−0.002	0.122	0.116	0.938
2	−0.171	−0.193	0.051	0.050	0.026	0.176	0.053	0.052	0.312	−0.001	0.129	0.123	0.935
3	−0.171	−0.053	0.048	0.047	0.797	0.159	0.050	0.049	0.490	0.001	0.140	0.137	0.948
4	0.911	−0.164	0.047	0.046	0.049	−0.456	0.049	0.048	0.819	0.003	0.098	0.095	0.942
5	0.911	−0.195	0.047	0.046	0.018	−0.451	0.048	0.048	0.864	−0.001	0.097	0.098	0.953
6	0.911	−0.139	0.042	0.045	0.110	−0.477	0.044	0.047	0.512	0.002	0.124	0.125	0.946
7	0.747	0.151	0.046	0.046	0.074	−0.329	0.048	0.048	0.864	−0.004	0.088	0.086	0.948
8	0.747	0.129	0.046	0.045	0.174	−0.360	0.048	0.047	0.914	0.005	0.126	0.126	0.956
9	0.747	0.019	0.047	0.045	0.918	−0.366	0.050	0.047	0.879	−0.001	0.086	0.084	0.938
b) Drawing a validation sample of size M = 1000													
1	−0.171	−0.159	0.051	0.050	0.104	0.158	0.053	0.052	0.555	−0.001	0.087	0.088	0.952
2	−0.171	−0.193	0.049	0.050	0.027	0.176	0.050	0.052	0.302	0.002	0.090	0.092	0.950
3	−0.171	−0.051	0.047	0.047	0.814	0.159	0.048	0.049	0.478	0.005	0.102	0.101	0.951
4	0.911	−0.165	0.046	0.046	0.060	−0.457	0.048	0.048	0.821	0.002	0.074	0.074	0.949
5	0.911	−0.198	0.044	0.046	0.004	−0.452	0.046	0.048	0.873	−0.003	0.075	0.076	0.955
6	0.911	−0.139	0.046	0.045	0.137	−0.477	0.048	0.047	0.506	−0.005	0.094	0.092	0.941
7	0.747	0.152	0.044	0.046	0.076	−0.329	0.047	0.048	0.874	0.002	0.069	0.069	0.948
8	0.747	0.126	0.047	0.045	0.195	−0.361	0.049	0.047	0.910	0.003	0.095	0.094	0.951
9	0.747	0.018	0.045	0.045	0.937	−0.366	0.047	0.047	0.882	0.000	0.069	0.068	0.941
c) Drawing a validation sample of size M = 1500													
1	−0.171	−0.156	0.049	0.050	0.127	0.159	0.050	0.052	0.558	0.001	0.078	0.077	0.949
2	−0.171	−0.195	0.050	0.050	0.035	0.175	0.051	0.052	0.315	0.000	0.078	0.080	0.953
3	−0.171	−0.055	0.047	0.047	0.790	0.157	0.048	0.049	0.496	−0.004	0.085	0.086	0.953
4	0.911	−0.165	0.047	0.046	0.058	−0.456	0.049	0.048	0.807	0.003	0.067	0.066	0.950
5	0.911	−0.195	0.044	0.046	0.015	−0.450	0.046	0.048	0.889	0.002	0.065	0.067	0.959
6	0.911	−0.139	0.046	0.045	0.131	−0.477	0.047	0.047	0.518	−0.002	0.081	0.079	0.942
7	0.747	0.149	0.046	0.046	0.093	−0.330	0.048	0.048	0.874	−0.002	0.061	0.062	0.942
8	0.747	0.126	0.044	0.045	0.182	−0.362	0.044	0.047	0.935	0.002	0.081	0.080	0.945
9	0.747	0.017	0.045	0.045	0.941	−0.367	0.048	0.047	0.873	−0.004	0.061	0.061	0.948

Abbreviations: ASE, average of the standard errors; BSE, bootstrap standard errors; CP, coverage percentage; SSE, sample standard errors.

For each study, we computed the average of the 1000 estimates of the marginal causal log odds ratio, Equation 1, using only the observed outcomes, Equation 4, the inverse PS weighted outcomes, Equation 5, and the outcomes weighted by both PSs and the proposed weights, Equation 8. In Table 6, we present the average bias for these estimators as $\text{Bias}_{\text{crude}} = \log \overline{\widehat{OR}}^{*}_{\text{YA}} - \log OR$, $\text{Bias}_{\text{PS}} = \log \overline{\widehat{OR}}^{*}_{\text{PS}} - \log OR$, and $\text{Bias}_{\text{W}} = \log \overline{\widehat{OR}}^{*}_{\text{W}} - \log OR$, respectively, there the bar is used to

denote the average of the estimates. We present the sample standard errors under the column "SSE" in Table 6. As noted in Section 4, we estimated the standard error of $\log \widehat{OR}^{*}_{\mathrm{W}}$ using a bootstrap approach in which we nonparametrically bootstrapped the treatment variable and additional covariate and applied a parametric bootstrap to generate the observed and validated outcomes. We present the square root of the average sample variance of the bootstrap estimates of $\log \widehat{OR}^{*}_{\mathrm{W}}$ in Table 6 under the column denoted as "BSE," which signifies the bootstrap standard error. The bootstrap procedure was conducted 200 times per generated sample that is justified as sufficiently large in Efron and Tibshriani.[18] For the crude and PS weighted approaches, we present the average of the standard errors produced by the *glm*() and *svyglm*() functions in R, respectively, under the column marked "ASE." Finally, we present coverage percentage estimates under the column "CP" as the proportion of the 1000 iterations, run for each simulation parameter set, for which a 95% asymptotically normal confidence interval covered the true value of the marginal casual log odds ratio for treatment effect.

These results affirm many of the statements discussed in previous sections. As expected, $\log \widehat{OR}^{*}_{\mathrm{YA}}$ and $\log \widehat{OR}^{*}_{\mathrm{PS}}$ are clearly biased across all simulation parameter sets. For the proposed weighted estimator, $\log \widehat{OR}^{*}_{\mathrm{W}}$, the bias appears to be approximately zero, demonstrating the successful use of the proposed weights. This holds for all validation sample sizes and across all of the simulation parameter sets. As m continues to decrease, the estimates become slightly more biased and much more variable. This is particularly true when the y^{*} validation subgroups have extremely small sample sizes (rare disease detection, for example).

To extend this to incorporate multiple covariates of different types and to create simulated data based on more realistic settings, we use an algorithm motivated from the approach used in Setoguchi et al.[11] This data generation algorithm has been used by a number of authors tailored to their individual research questions.[19-21] A hypothetical cohort study of size n was generated with a binary outcome, Y, a binary treatment, A, and 10 additional covariates, $L_1, \ldots, L_{10}$. The covariates were generated as standard normal random variables with correlations introduced using the correlation matrix specified in Table 7. T7

Four of the covariates were generated to be confounders (L_1, L_2, L_3, L_4), 3 to be predictors of treatment only (L_5, L_6, L_7), and 3 to be predictors of outcome only (L_8, L_9, L_{10}). We conducted Bernoulli trials to generate the binary treatment variable, A, characterized by the model described as "Scenario A" in Setoguchi et al[11] (reproduced here),

$$P(A = 1|L) = \frac{\exp\left(\alpha_0 + \sum_{k=1}^{7} \alpha_k L_k\right)}{1 + \exp\left(\alpha_0 + \sum_{k=1}^{7} \alpha_k L_k\right)},$$

and to generate the binary outcome variable, Y, characterized by the model,

$$P(Y = 1|A, L) = \frac{\exp(\beta_0 + \beta_{\mathrm{treat}} A + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3 + \beta_4 L_4 + \beta_5 L_8 + \beta_6 L_9 + \beta_7 L_{10})}{1 + \exp(\beta_0 + \beta_{\mathrm{treat}} A + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3 + \beta_4 L_4 + \beta_5 L_8 + \beta_6 L_9 + \beta_7 L_{10})}.$$

To misclassify the generated Y values, we conducted additional Bernoulli trials using the models outlined in the previous section to characterize the θ-parameters, Equation 11. The generated outcomes, Y^{*}, were treated as the observed, possibly

TABLE 7 Correlation matrix used to generate the additional covariates

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	L_9	L_{10}
L_1	1	0	0	0	0.2	0	0	0	0	0
L_2	0	1	0	0	0	0.9	0	0	0	0
L_3	0	0	1	0	0	0	0	0.2	0	0
L_4	0	0	0	1	0	0	0	0	0.9	0
L_5	0.2	0	0	0	1	0	0	0	0	0
L_6	0	0.9	0	0	0	1	0	0	0	0
L_7	0	0	0	0	0	0	1	0	0	0
L_8	0	0	0.2	0	0	0	0	1	0	0
L_9	0	0	0	0.9	0	0	0	0	1	0
L_{10}	0	0	0	0	0	0	0	0	0	1

misclassified, version of the outcome. We began by investigating 2 sets of misclassification parameters, $(\kappa_{10}, \kappa_{11}, \kappa_{00}, \kappa_{01})$, namely, $(2, 0.5, -1.9, -0.8)$ and $(1.5, 0.5, -1.5, -0.8)$. Since we have chosen to model the θ-parameters differentially with respect to treatment only, these values corresponded to error rates ranging from 6.3% to 13% and 9.1% to 18.2%, respectively. We assumed that only a subset of size m of the generated Y values were observable through a validation sampling procedure.

As noted in Section 4, our interest lies in estimation of $\exp(\gamma_{\text{treat}})$, which will be equivalent to the marginal causal odds ratio when estimated from the logistic regression model in Equation 12 using the proposed weighted outcomes. Setoguchi et al[11] provided the following simulation parameters (written using our notation): $\beta = (-3.85, 0.3, -0.36, -0.73, -0.2, 0.71, -0.19, 0.26)$, $\alpha = (0, 0.8, -0.25, 0.6, -0.4, -0.8, -0.5, 0.7)$, and $\beta_{\text{treat}} = -0.4$. These parameters were noted to produce a target marginal value of $\gamma_{\text{treat}} = -0.4$ (odds ratio of 0.67032) and were chosen to generate a rare outcome, which may create difficulties in sampling sufficiently from all possible (y^*, y, a) subgroups in the validation sample. Hence, we selected a sufficiently large original sample size, $n = 10000$, and drew validation sample sizes of $m = 1000$ and $m = 2000$.

To investigate the proposed method for outcomes that are not as rare, we also considered setting $\beta_0 = 0$ while keeping the rest of the covariate values as before. We introduced even more misclassification by setting $\kappa_{10} = 1$ and $\kappa_{00} = -1$, which increased the range of error rates from 13% to 26.9%. We used the iterative Monte Carlo integration approach described in Austin et al[22] to determine the value of the coefficient to treatment in the outcome model needed to keep the target marginal log odds ratio at $\gamma_{\text{treat}} = -0.4$. This value was set at $\beta_{\text{treat}} = -0.447$. Under this setting, we generated data from 4 values of (n, m), namely, $(1000, 200)$, $(1000, 100)$, $(10000, 2000)$ and $(10000, 1000)$. In Table 8, we display 8 sets of simulation parameters and note that 1000 iterations were run for each set.

The results displayed in Table 9 follow the same structure as those in Table 6. For each simulation study, we computed the average bias of the estimates of γ_{treat} computed from the crude, denoted as $\hat{\gamma}^{\text{crude}}_{\text{treat}}$, inverse PS weighted, denoted as $\hat{\gamma}^{\text{PS}}_{\text{treat}}$, and proposed modified IP weighted, denoted as $\hat{\gamma}^{\text{W}}_{\text{treat}}$, estimators. We denote the average biases as $\text{Bias}_{\text{crude}} = \bar{\hat{\gamma}}^{\text{crude}}_{\text{treat}} - \gamma_{\text{treat}}$, $\text{Bias}_{\text{PS}} = \bar{\hat{\gamma}}^{\text{PS}}_{\text{treat}} - \gamma_{\text{treat}}$ and $\text{Bias}_{\text{W}} = \bar{\hat{\gamma}}^{\text{W}}_{\text{treat}} - \gamma_{\text{treat}}$, respectively. For each estimator, we display the SSE, ASE/BSE, and CP as defined for Table 6.

TABLE 8 Target parameters used in Monte Carlo simulation studies to investigate the proposed weighted estimator

Index	β_0	β_{treat}	κ_{10}	κ_{11}	κ_{00}	κ_{01}	n	m
1	−3.85	−0.4	2	0.5	−1.9	−0.8	10 000	2000
2	−3.85	−0.4	2	0.5	−1.9	−0.8	10 000	1000
3	−3.85	−0.4	1.5	0.5	−1.5	−0.8	10 000	2000
4	−3.85	−0.4	1.5	0.5	−1.5	−0.8	10 000	1000
5	0	−0.447	1	0.5	−1	−0.8	10 000	2000
6	0	−0.447	1	0.5	−1	−0.8	10 000	1000
7	0	−0.447	1	0.5	−1	−0.8	1000	200
8	0	−0.447	1	0.5	−1	−0.8	1000	100

Parameters omitted from this table are set to be the same as in Setoguchi et al.[11]

TABLE 9 Simulation study results investigating the proposed weighted estimator

Index	$\text{Bias}_{\text{crude}}$	SSE	ASE	CP	Bias_{PS}	SSE	ASE	CP	Bias_{W}	SSE	BSE	CP
1	−0.2920	0.0649	0.0653	0.004	−0.3246	0.0783	0.0776	0.007	0.0060	0.2840	0.2921	0.962
2	−0.2924	0.0645	0.0654	0.004	−0.3256	0.0771	0.0776	0.007	0.0220	0.3848	0.3809	0.933
3	−0.3243	0.0581	0.0578	0.000	−0.3474	0.0689	0.0687	0.002	0.0106	0.3060	0.3076	0.952
4	−0.3247	0.0573	0.0578	0.000	−0.3452	0.0689	0.0687	0.001	0.0270	0.4165	0.4177	0.928
5	0.1461	0.0412	0.0402	0.046	0.0581	0.0482	0.0479	0.778	0.0055	0.0788	0.0800	0.955
6	0.1445	0.0391	0.0402	0.045	0.0560	0.0478	0.0479	0.783	0.0032	0.1071	0.1059	0.972
7	0.1471	0.1265	0.1273	0.776	0.0585	0.1500	0.1526	0.940	−0.0055	0.2539	0.2589	0.959
8	0.1451	0.1220	0.1273	0.804	0.0551	0.1456	0.1527	0.952	−0.0041	0.3545	0. 3480	0.954

Abbreviations: ASE, average of the standard errors; BSE, bootstrap standard errors; CP, coverage percentage; SSE, sample standard errors.

These results are similar to those described in the first set of studies. Bias is evident in the estimators that ignore outcome misclassification, $\hat{\gamma}^{\text{crude}}_{\text{treat}}$ and $\hat{\gamma}^{\text{PS}}_{\text{treat}}$, while the bias associated with the proposed modified IPW estimator, $\hat{\gamma}^{\text{W}}_{\text{treat}}$, appears to be approximately zero. The standard error for the proposed approach was higher than that of the PS and crude approaches, which is unsurprising given the need to estimate additional parameters. Recall that simulation parameter sets 1 through 4 in Table 8 were based on a rare outcome, and as such, we would expect inflation in the variability of the estimators due to the impact of sampling from (y^*, y, a) subgroups with overly small sample sizes. This is further noted by the improved performance of the proposed approach when the outcome is not as rare, as is seen in studies 5 through 8 of Table 8. As expected, increasing the size of the validation sample decreases the variance for all studies as is seen in Table 9. Finally, the bootstrap standard error estimates appear to approximate the standard deviation of the Monte Carlo estimates, and the coverage proportion estimates are all close to the nominal value of 0.95.

6 | CONCLUSION

The common assumption of no measurement error of the outcome is often violated in practice. We have demonstrated that modified IP weighting, using MLEs of misclassification parameters derived with internal validation data, can be implemented to offset the biases simultaneously brought on by confounding and outcome misclassification.

This is evidenced by the pseudo-population produced with the proposed weights in Equation 7. With misclassification, we are acknowledging that the observed data are a "mutilated" version of the correctly classified information based on 2 influences (in a binary setting), see Equation 6. The implication is that each contribution to the pseudo-population for outcome misclassification is scaled in 2 ways and the magnitude of the scaling depends on the estimated accuracy of the observed outcome. In other words, the presence of outcome misclassification makes it such that we are not able to trust that the observed outcome is in fact the potential outcome under the observed treatment, $a = 0, 1$, instead it may be either outcome and these weights are a function of the probabilities of both possibilities.

We conducted simulation studies to investigate the finite sample properties of the proposed weighted estimator of the marginal causal effect of treatment. We considered reasonably large sample sizes to lessen the possibilities of problematically small validation subgroup sizes. For many data applications, such as in drug safety databases, the sample sizes used in these simulation studies are realistic; however, as the error rates decrease for overly small subgroup sizes, the probability of failing to observe information on all error types in the validation data increases. In this scenario, estimated weights may become slightly biased due to the potential for the target θ-parameters at the boundary and considering alternate approaches to estimate these parameters could be considered. For example, we could use a Bayesian beta-binomial estimator in which the prior beta hyperparameters are chosen based on external estimates of the relevant diagnostic error rates. Provided that the parameter vector point estimates are unbiased, these weights will continue to restore consistency in the resulting pseudo-population.

The proposed weights were estimated using internal validation data in which the outcome was assumed to be measured correctly. This assumption may be unrealistic in practice and future research is needed to explore the extent of residual bias resulting from possible violations. The lack of a single gold standard diagnostic test with perfect sensitivity and specificity is likely and often the validation measurement tool may possess better properties, but will remain error-prone. Research has been done considering this realistic setting,[23,24] and incorporating this approach for estimation of the proposed weights is a logical extension that will enable practical implementation. Bayesian methods may be relevant.

Finally, the proposed weights provide a general structure for outcome misclassification bias adjustment that can be implemented in more complex data structures. Inverse probability weighting has been successfully used to address many other sources of bias in complex data and our weights further this objective. As the complexity increases, the difficulties implementing this approach lie in the availability and accessibility of the necessary information to produce reasonable point estimates of the π and θ-parameters. However, with the requisite data, these weights can offer an alternate approach to model a variety of data structures biased due to outcome misclassification.

ACKNOWLEDGMENTS

Dr Gravel holds a postdoctoral Mitacs Accelerate Internship with McGill University and Risk Sciences International, Ottawa, Ontario. Dr Platt is a Chercheur-National from the Fonds de recherche du Québec—Santé (FRQS) and a member of the Research Institute of the McGill University Health Centre, which receives financial support from the FRQS, and holds the Albert Boehringer I Chair in Pharmacoepidemiology.

ORCID

Christopher A. Gravel http://orcid.org/0000-0002-7780-3392

REFERENCES

1. Hernán MA, Robins JM. *Causal Inference*. Boca Raton: Chapman & Hall/CRC. forthcoming.
2. Barron BA. The effects of misclassification on the estimation of relative risk. *Biometrics*. 1977;33:414-418.
3. Marshall RJ. Validation study methods for estimating proportions and odds ratios with misclassified data. *J Clin Epidemiol*. 1990;43:941-947.
4. Greenland S. Variance estimation for epidemiologic effect estimates under misclassification. *Stat Med*. 1988;7:745-757.
5. Lyles RH. A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. *Biometrics*. 2002;58:1034-1037.
6. Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. *Am J Epidemiol*. 1997;146(2):195-203.
7. Lyles RH, Tang L, Superak HM, et al. Validation data-based adjustments for outcome misclassification in logistic regression. *Epidemiology*. 2011;22:589-98.
8. Edwards JK, Cole SR, Troester MA, Richardson DB. Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *Am J Epidemiol*. 2013;177(9):904-12.
9. Levine PJ, Elman MR, Kullar R, et al. Use of electronic health record data to identify skin and soft tissue infections in primary care settings: a validation study. *BMC Infect Dis*. 2013;13:171.
10. Herrett E, Shah AD, Boggon R, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ*. 2013;20:346:f2350.
11. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf*. 2008;17(6):546-55.
12. Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist Sci*. 1990;5(4):465-472. Trans. Dorota M. Dabrowska and Terence P. Speed.
13. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688-701.
14. Pearl J. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York, NY: Cambridge University Press; 2000.
15. Prescott GJ, Garthwaite PH. A simple Bayesian analysis of misclassified binary data with a validation substudy. *Biometrics*. 2002;58(2):454-8.
16. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Boca Raton FL: Chapman & Hall; 2003.
17. McInturff P, Johnson WO, Cowling D, Gardner IA. Modelling risk when binary outcomes are subject to error. *Stat Med*. 2004;23(7):1095-109.
18. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York NY: Chapman & Hall; 1993.
19. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;10;29(3):337-46.
20. Austin PC. Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based G-computation. *Multivariate Behav Res*. 2012;47(1):115-135.
21. Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Stat Methods Med Res*. 2015. pii: 0962280215584401.
22. Austin PC, Stafford J. The performance of two data-generation processes for data with specified marginal treatment odds ratios. *Commun Stat-Simul C*. 2008;37:6.
23. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol*. 1995;141(3):263-72.
24. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*. 2001;57(1):158-67.

How to cite this article: Gravel CA, Platt RW. Weighted estimation for confounded binary outcomes subject to misclassification. *Statistisc in Medicine*. 2017;1–12. https://doi.org/10.1002/sim.7522