Title

Simultaneous adjustment of uncontrolled confounding, selection bias and misclassification in multiple-bias modelling.

Journal

International Journal of Epidemiology, 52(4)

Authors

Brendel, Paul
Torres, Aracelis
Arah, Onyebuchi

Publication Date

2023-08-02

DOI

10.1093/ije/dyad001

Peer reviewed

International Epidemiological Association

International Journal of Epidemiology, 2023, 1220–1230
https://doi.org/10.1093/ije/dyad001
Advance Access Publication Date: 30 January 2023
Original article

OXFORD

Methods

Simultaneous adjustment of uncontrolled confounding, selection bias and misclassification in multiple-bias modelling

Paul Brendel,[1,2]* Aracelis Torres,[3] and Onyebuchi A Arah [1,4,5]

[1]Department of Epidemiology, Fielding School of Public Health, UCLA, Los Angeles, CA, USA, [2]Valo Health, Boston, MA, USA, [3]Verana Health, San Francisco, CA, USA, [4]Department of Statistics, College of Letters and Science, UCLA, Los Angeles, CA, USA and [5]Department of Public Health, Section for Epidemiology, Aarhus University, Aarhus, Denmark

*Corresponding author. Valo Health, 399 Boylston Street, Suite 505, Boston, MA 02116, USA.
E-mail: pbrendel@valohealth.com

Received 15 December 2021; Editorial decision 15 December 2022; Accepted 23 January 2023

Abstract

Background: Adjusting for multiple biases usually involves adjusting for one bias at a time, with careful attention to the order in which these biases are adjusted. A novel, alternative approach to multiple-bias adjustment involves the simultaneous adjustment of all biases via imputation and/or regression weighting. The imputed value or weight corresponds to the probability of the missing data and serves to 'reconstruct' the unbiased data that would be observed based on the provided assumptions of the degree of bias.
Methods: We motivate and describe the steps necessary to implement this method. We also demonstrate the validity of this method through a simulation study with an exposure-outcome relationship that is biased by uncontrolled confounding, exposure misclassification, and selection bias.
Results: The study revealed that a non-biased effect estimate can be obtained when correct bias parameters are applied. It also found that incorrect specification of every bias parameter by +/-25% still produced an effect estimate with less bias than the observed, biased effect.
Conclusions: Simultaneous multi-bias analysis is a useful way of investigating and understanding how multiple sources of bias may affect naive effect estimates. This new method can be used to enhance the validity and transparency of real-world evidence obtained from observational, longitudinal studies.

Key words: Multi-bias modelling, confounding, information bias, selection bias, parameters, imputed, regression weight, mis-specification, simulation

Key Messages

- Multi-bias modelling has usually involved the sequential adjustment of uncontrolled confounding, information bias and selection bias, using knowledge of the sequence in which the biases took place.
- A new approach to bias adjustment allows for multiple biases to be adjusted simultaneously by combining individual-level data with bias parameters to obtain imputed values or a regression weight.
- In the case of bias parameter mis-specification, this method can still produce reasonable bias-reduced estimates, as demonstrated by a simulation study.
- Bias parameters can be varied to assess what biases and bias strengths would be necessary to observe a null effect or other reported effect.

Introduction

Bias analyses are used to quantify threats to validity in research[1] and are essential tools in studies involving real-world data from the electronic health records (EHR), claims databases, and registries.[2] Methods to account for bias are constantly being improved[3] and typically have involved adjusting for single biases, including simple sensitivity analysis, Monte Carlo risk analysis, Bayesian uncertainty assessment and external adjustment formulas, among others.[4–9] These methods can be varied by different bias parameters, and considerations include whether bias parameters are fixed or probabilistic and whether the bias parameters are applied to the data as in a missing data approach or to the observed estimate as in external adjustment.

Outside of recent innovations in quantifying the bounds of the composite contribution of multiple biases on causal effect estimates,[10] the development of methods for the adjustment of multiple biases has been comparatively stagnant.[11,12] An approach in which biases are adjusted sequentially has been described, but to implement this method, biases should be adjusted in the proper order (the reverse of the order in which the biases occurred in the data generation process).[1,13,14] As described by Greenland: 'One can imagine each correction moving a step from the biased data back to the unbiased structure, as if hypothetically "unwrapping the truth from the data package"'.[1,13,14] Since the order of adjustments can influence the results of a sequential bias analysis, such analyses can be difficult if the true sequence of biases is unknown or hard to ascertain. In addition, these adjustments can be time-consuming and prone to error if many biases are to be evaluated.

To overcome these problems, we introduce a new method to adjust for multiple biases of binary variables simultaneously. We specifically address uncontrolled confounding, selection bias, and exposure misclassification, although the method also applies to outcome misclassification and other, more complex forms of these biases. This method generalizes the concept of combining inverse probability of selection weighting (IPSW) with predictive value weighting, as introduced by Johnson *et al.*[15–17] It relies on predicting the probability of the missing data (uncontrolled confounders, misclassified exposure or selection bias) using the available data and externally obtained information or assumptions on the effect of these data on that which is missing (i.e. bias parameters). These predicted probabilities are then incorporated as simulated values or weights in the outcome regression. We outline the steps for performing simultaneous multi-bias adjustment on any combination of uncontrolled confounding, exposure misclassification, and selection bias. We also verify the validity of this method and explore the sensitivity of effect estimates to misspecified bias parameters, through a simulation study.

Methods

Notation and assumptions

The following binary variables are defined: X = exposure, Y = outcome, $\mathbf{C}$ = vector of known confounders, U = unknown or uncontrolled confounder, X^* = misclassified exposure, S = selection. These variables are used to represent potential multi-bias scenarios, depicted in directed acyclic graphs (DAGs) (Figure 1). Although this paper specifically focuses on the simplified bias scenarios in Figure 1, this method can be generalized to bias adjustment with alternative or additional causal paths to these DAGs.

Bias can be evaluated in DAGs using the backdoor criterion and other rules.[15,18–21] The direct effect of X on Y in these causal models is distorted by backdoor paths stemming from uncontrolled confounding ($X \leftarrow U \rightarrow Y$), information bias in the form of exposure misclassification ($X^* \leftarrow (X) \rightarrow Y$) and selection (i.e. collider stratification) bias ($X \rightarrow |S| \leftarrow Y$). Values of variables U and X are unknown and observations with $S = 0$ are missing. An overlined variable refers to a variable whose value is assigned by the investigator to a particular data replicate to perform regression weighting. In the multi-bias analysis, $\overline{X}$ and $\overline{U}$

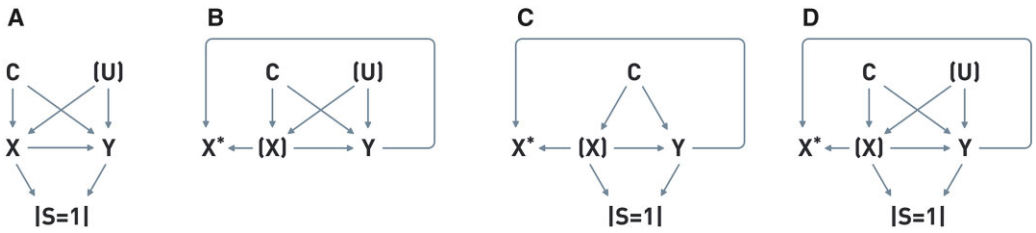

Figure 1 Graphs of four potential multi-bias scenarios. • X = exposure, Y = outcome, $\boldsymbol{C}$ = vector of known confounders, U = unknown or uncontrolled confounder, X^* = misclassified exposure, S = selection. • DAG A: uncontrolled confounding and selection bias. • DAG B: uncontrolled confounding and exposure misclassification. • DAG C: exposure misclassification and selection bias. • DAG D: uncontrolled confounding, selection bias and exposure misclassification

represent assigned values of exposure and uncontrolled confounder, and X_{IMP} and U_{IMP} represent imputed values of exposure and uncontrolled confounder from a single Bernoulli trial.

Performing simultaneous multi-bias analysis relies on understanding the equivalence of causal systems represented as either DAGs or structural equation models (SEMs). Since the exposure, confounder, and selection indicator are each binary in this study, they were modelled using a system of logistic regression models. The exponentiated parameters of these models correspond to covariate-outcome odds ratios, which are conditional on the other covariates in the model. For ease of exposition, we ignore product terms in the logistic regression models, although the methods described here can accommodate such interaction terms in real-world applications. Further, whereas this study does not specifically address non-binary exposures and confounders, the overall framework of this bias analysis method remains applicable to non-binary variables. The multi-bias analysis method in this paper relies on the causal assumptions of positivity, consistency, the stable unit treatment value assumption, and exchangeability given the measured and unmeasured (de)confounders.

Simultaneous multi-bias analysis

Simultaneous multi-bias analysis combines subject-level data with assumptions about the plausible causal structure to recreate an unbiased dataset. There are eight steps to performing this adjustment (Table 1). Steps 1–5 lead to the calculation of the predicted probabilities of the unknown or incomplete variables for each subject. After these probabilities are established, the investigator can choose to incorporate these probabilities into the data through imputed values, a regression weight or a combination of both. Table 1 outlines these steps for bias adjustment in the presence of uncontrolled confounding, exposure misclassification, and selection bias. Supplementary Table S1 (available as Supplementary data at *IJE* online) similarly shows these steps under the different combinations of two biases. We describe these steps below.

First, the investigator can graph the assumed causal relationships between each variable in the observed (biased) and desired (non-biased) data as two DAGs. Statements for the observed and desired joint probabilities can be expressed for both causal models. Using these two probabilities, a bias-adjusting conditional probability is found by the following equation:

$$\textit{bias–adjusting probability} = \textit{desired joint probability} \,/\, \textit{observed joint probability}$$

This probability derivation is similar to inverse probability of treatment weighting, where the weight may have a numerator with the probability of the treatment conditional on previous treatments (that which is desired) and the denominator would have the probability of the treatment conditional on previous treatments and the bias-inducing covariates (that which is observed).[22] Adjustments for selection bias will include a bias-adjusting probability whose denominator is equal to the conditional probability of selection, as seen in single-bias IPSW.[15] The expression is rewritten such that the unknown values (true exposure, confounder, selection indicator) are conditional on the known values; see Table 2 for a full example of the necessary probability manipulations.

The numerator and denominator of the bias-adjusting conditional probability are then expressed as structural equation models, specifically logistic regression models. In the case of both uncontrolled confounding and exposure misclassification, two separate regression models or a single multinomial logistic regression can be used to represent the conditional probability of the confounder and exposure. In the latter case, the model's dependent variable corresponds to each potential combination of the X, U values.

Regression coefficients for the logistic regression models for exposure, confounder, and study selection (i.e. bias parameters) are then externally obtained. Plausible parameter values can derive from a variety of sources: previous

Table 1 Steps to performing simultaneous multi-bias analysis (three biases)

Step	Uncontrolled confounding, Exposure misclassification, and selection bias
1. Determine the observed probability and the desired joint probability	Observed: $P(x^*, \boldsymbol{c}, y \mid S=1)$ Desired: $P(\boldsymbol{x}, y, \boldsymbol{c}, \boldsymbol{u})$
2. Divide the desired probability by the observed probability and rewrite in terms of the unknown values	$\frac{P(x,\ u \mid x^*,\ y\boldsymbol{c})}{P(S=1 \mid x^*,\ y,\ \boldsymbol{c})}$
3. Rewrite this bias-adjusting probability as corresponding statistical models (i.e. bias models)	$\log\left(\frac{P(X=1,\ U=0)}{P(X=0,\ U=0)}\right) = \beta_{1,0} + \beta_{1,1}X^* + \beta_{1,2}\ Y + \boldsymbol{\beta}_{1,3}\ \mathbf{C}$ $\log\left(\frac{P(X=0,\ U=1)}{P(X=0,\ U=0)}\right) = \beta_{2,0} + \beta_{2,1}X^* + \beta_{2,2}\ Y + \boldsymbol{\beta}_{2,3}\ \mathbf{C}$ $\log\left(\frac{P(X=1,\ U=1)}{P(X=0,\ U=0)}\right) = \beta_{3,0} + \beta_{3,1}X^* + \beta_{3,2}\ Y + \boldsymbol{\beta}_{3,3}\ \mathbf{C}$ $logit(P(S=1)) = \delta_0 + \delta_1 X^* + \delta_2 Y + \boldsymbol{\delta}_3 \mathbf{C}$
4. Externally obtain the regression coefficients for the bias models (i.e. the bias parameters)	
5. Using the bias parameters and individual-level data, solve for each probability for each patient	
Option A: Regression weighting	
6a. Replicate the data and assign values for exposure and/or uncontrolled confounder.	In first replicate: $\overline{X}=1,\ \overline{U}=1$ In second replicate: $\overline{X}=1,\ \overline{U}=0$ In third replicate: $\overline{X}=0,\ \overline{U}=1$ In fourth replicate: $\overline{X}=0,\ \overline{U}=0$
7a. Create individual weights based on the predicted exposure/outcome conditional probabilities (corresponding to the assigned $\overline{X}$ and/or $\overline{U}$) divided by the conditional probability of selection	When $\overline{X}=1,\ \overline{U}=1$: Weight $= \frac{P(X=1,\ U=1 \mid x^*, y, \boldsymbol{c})}{P(S=1 \mid x^*, y, \boldsymbol{c})}$ When $\overline{X}=1,\ \overline{U}=0$: Weight $= \frac{P(X=1,\ U=0 \mid x^*, y, \boldsymbol{c})}{P(S=1 \mid x^*, y, \boldsymbol{c})}$ When $\overline{X}=0,\ \overline{U}=1$: Weight $= \frac{P(X=0,\ U=1 \mid x^*, y, \boldsymbol{c})}{P(S=1 \mid x^*, y, \boldsymbol{c})}$ When $\overline{X}=0,\ \overline{U}=0$: Weight $= \frac{P(X=0,\ U=0 \mid x^*, y, \boldsymbol{c})}{P(S=1 \mid x^*, y, \boldsymbol{c})}$
8a. Perform weighted outcome regression	$logit(P(Y=1)) = \omega_0 + \omega_1\overline{X} + \omega_2\overline{U} + \boldsymbol{\omega}_3\mathbf{C}$
Option B: Exposure/outcome imputation combined with selection weighting	
6b. Using the predicted probabilities for X and U, impute values for each by sampling from the binomial distribution	$U_{IMP} =$ Random.binomial(n, 1, P($U=1$)) $X_{IMP} =$ Random.binomial(n, 1, P($X=1$))
7b. Determine the selection weight for each patient	$\frac{1}{P(S=1 \mid x^*, y, \boldsymbol{c})}$
8b. Perform outcome regression using the imputed values and selection weight	$logit(P(Y=1)) = \omega_0 + \omega_1 X_{IMP} + \omega_2 U_{IMP} + \boldsymbol{\omega}_3\mathbf{C}$

[a]Variable key: X = exposure, Y = outcome, $\mathbf{C}$ = vector of known confounders, S = selection.
[b]Variable modifiers: X^* = misclassified exposure, $\overline{X}$ = assigned value for the exposure in a data replicate, X_{IMP} = imputed value for the exposure.

studies, a validation sub-study, an expert opinion, machine learning model benchmarking or simply the investigator's best estimate. As with all bias analyses, it is important to clearly communicate the derivation of each parameter and to consider representing each parameter as a distribution of values to iteratively sample over instead of just a point estimate.[23] Once these parameters are obtained, they are combined with the individual-level data to determine the predicted probabilities of exposure, confounder, and study selection for each subject.

The investigator may choose an imputation approach or a weighting approach to incorporate the predicted probabilities of X and/or U. With the imputation approach, the new value (X_{IMP} and/or U_{IMP}) is simulated via a single Bernoulli trial for each patient, using the predicted probabilities. In the weighting approach, subjects are first assigned each potential

Table 2 Deriving the bias-adjusting probability in the triple-bias scenario[a]

$\frac{P(x,\ y,\ c,u)}{P(x^*,\ y,\ c\|S=1)}$	Dividing the observed joint probability by the desired joint probability
$\frac{\sum_{x^*} P(x,\ y,\ c,u,\ x^*)}{P(x^*,\ y,\ c\|S=1)}$	Law of total probability
$\frac{\sum_{x^*} P(x,\ u\ \|x^*,\ y,\ c)P(x^*,\ y,c)}{P(x^*,\ y,\ c\|S=1)}$	Multiplication rule
$\frac{\sum_{x^*} P(x,\ u\ \|x^*,\ y,\ c)P(x^*,\ y,c)}{\frac{P(S=1\|x^*,\ y,c)P(x^*,\ y,c)}{P(S=1)}}$	Bayes rule
$\frac{P(x,\ u\|x^*,\ yc)}{P(S=1\|x^*,\ y,\ c)}$	Rearrange terms, (optionally) remove constant (P($S=1$))

[a]See Figure 1D for corresponding causal graph.

value of exposure and/or unobserved confounder as $\overline{X}$ and/or $\overline{U}$. Thus, the data are replicated twice (if X or U is missing) or four times (if X and U are missing), with each replicate containing a different value of the assigned variable(s). The weight is then assigned according to the value of $\overline{X}$ and/or $\overline{U}$; if $\overline{X}$ and/or $\overline{U}=1$, then the weight equals the predicted probability; if $\overline{X}$ and/or $\overline{U}=0$, then the weight equals 1—the predicted probability. The selection probability is incorporated as the denominator of this weight.

The final logistic regression outcome model therefore includes: (i) the imputed values, X_{IMP} and/or U_{IMP}, with a weight equalling the inverse of the predicted probability of selection; or (ii) the assigned values, $\overline{X}$ and/or $\overline{U}$, with a weight equalling the predicted probability of exposure and/or confounder divided by the predicted probability of selection. In this final model, the exponential of the exposure coefficient (X or $\overline{X}$) represents the bias-adjusted odds ratio of the direct effect of X on Y.

Imputation and weighting serve to reconstruct the data that would have been observed in the absence of bias, given the assumptions implicit in the bias parameters. In the case of selection weighting, observations with a lower probability of selection are given a greater weight in the analysis and vice versa, which serves to restore the initial variable distributions seen in the source population.[15] In the case of weighting with predicted probabilities of exposure and unobserved confounder, subjects are assigned each potential value of exposure and/or unobserved confounder, and the values that are most probable are given the most weight. Using $\overline{X}$ and $\overline{U}$ along with the weights serves to recreate a dataset in which the correctly classified exposure and unobserved confounder are both included.[5]

Simulation study

Aims

A simulation study was performed to demonstrate the proof-of-concept of the simultaneous multi-bias analysis method. The study will assess whether the method remains robust under (i) data with varying bias strengths and (ii) mis-specification of bias parameters.

Data-generating mechanism

Monte Carlo simulation was used to generate two datasets of binary variables whose causal relations were based off the DAG in Figure 1D—the triple bias scenario (Table 3). Each dataset of n_{obs} = 100 000 patients represents real-world data used for observational analysis for causal inference. One dataset has stronger individual biasing paths (Simulation A) and one has weaker individual biasing paths (Simulation B). The strengths of these biasing paths are guided by three parameters: ψ_1, (i) the conditional effect of the unknown confounder on exposure and outcome and (ii) the conditional effect of the exposure and outcome on selection; ψ_2, the log odds of misclassified exposure when true exposure and outcome are both absent; and ψ_3, the conditional effect of true exposure on misclassified exposure.

In Simulation A, $\psi_1=\log(2)$, $\psi_2=-1$ and $\psi_3=\log(5)$. These values were intended to create strong confounding by U and strong effects of X and Y on selection. The misclassified exposure could be simulated with a Bernoulli conditional probability as high as 0.70 (P($X^*=1|X=1,\ Y=1$)) or as low as 0.27 (P($X^*=1|X=0,\ Y=0$). Intercepts were selected to keep the conditional probabilities of true exposure, outcome, and selection bound within (0.12, 0.29), (0.08, 0.33), and (0.50, 0.80), respectively. In Simulation B, $\psi_1=\log(1.25)$, $\psi_2=-1.5$ and $\psi_3=\log(15)$. These values were intended to create weak confounding by U and weak effects of X and Y on selection. The misclassified exposure could be simulated with a Bernoulli conditional probability as high as 0.81 (P($X^*=1|X=1,\ Y=1$)) or as low as 0.18 (P($X^*=1|X=0,\ Y=0$). Intercepts were selected to keep the probabilities of true exposure, outcome, and selection bound within (0.12, 0.20), (0.08, 0.24) and (0.50, 0.61), respectively.

The datasets of binomial random variables were simulated and subsequently analysed using R version 3.2.2. To significantly increase the speed of bootstrapping, multiple CPU cores were used via parallel processing. The input seed was '1234'; see Supplementary Code for Simulation (available as Supplementary data at *IJE* online) to inspect the R code used for analysis.

Estimands

The estimand OR_{YX} represents the odds ratio for X = 1 versus X = 0, which would correspond to a causal effect estimate in an observational study.

Table 3 Data generating mechanism of binary variables for two simulated datasets

Variable	Description	Probability of variable = 1
C	Known confounder	0.5
U	Unknown confounder	0.5
X	Unknown, true exposure	$expit(-2 + log(1.5)C + \psi_1 U)$
Y	Outcome	$expit(-2.5 + log(2)X + log(1.5)C + \psi_1 U)$
S	Selection	$expit(\psi_1 X + \psi_1 Y)$
X*	Misclassified exposure	$expit(\psi_2 + \psi_3 X + log(1.25)Y)$

Methods

The following logistic regression model (Equation 1) fit among selected patients ($S = 1$) represents the 'real-world' scenario in which an investigator would model the outcome without consideration to (i) the values of true exposure and confounder U and (ii) the impact of exposure and outcome on patient selection:

$$logit(P(Y = 1|X^*, C)) = \alpha_Y + \alpha_{YX^*}X^* + \alpha_{YC}C \quad (1)$$

Here the biased $OR_{YX} = \exp(\alpha_{YX^*})$ does not equal the unbiased $OR_{YX} \approx 2$, which is known to be based on the simulation of Y in both datasets. The analysis assessed the ability to obtain an unbiased estimate of OR_{YX} using simulated, biased data and simultaneous multi-bias analysis.

The analysis began by identifying the observed (biased) joint probability, $(X^* = x^*, C = c, Y = y \mid S = 1)$, and the desired (bias-free) joint probability, $P(X = x, C = c, U = u, Y = y)$. After dividing the desired probability by the observed probability, the bias-adjusting conditional probability was $P(x, u \mid x^*, c, y)/P(S = 1 \mid x^*, c, y)$ (see Table 2 for detailed steps). The numerator was simplified by the multiplication rule: $P(x, u \mid x^*, c, y) = P(u \mid x, x^*, c, y)P(x \mid x^*, c, y)$ and $P(u \mid x, x^*, c, y)$ was rewritten as $P(u \mid x, y)$ due to the conditional independencies. The three probabilities for U, X, and S were rewritten as logistic regression models (Equations 2–4):

$$logit(P(X = 1)) = \delta_S + \delta_{SX^*}X^* + \delta_{SY}Y + \delta_{SC}C \quad (2)$$

$$logit(P(U = 1)) = \alpha_U + \alpha_{UX}X + \alpha_{UY}Y \quad (3)$$

$$logit(P(S = 1)) = \beta_S + \beta_{SX^*}X^* + \beta_{SY}Y + \beta_{SC}C \quad (4)$$

Having U, X, and S in the data allowed for the fitting of these models to obtain the correct bias parameters which, although impossible in real-world practice, is necessary for proper evaluation of the bias-adjustment method. To get the 11 parameters, models 2–4 were fit using data from Simulations A and B. As a reminder, real-world strategies for obtaining bias parameters do not involve calculating these parameters from within the sample data but require obtaining the parameters from external sources.

The imputation approach was used to incorporate the predicted probabilities of X and U. X_{IMP} was simulated via Bernoulli trial using the probabilities obtained from combining the bias parameters with the individual data for X^*, C, and Y, as in Equation 3. U_{IMP} was simulated via Bernoulli trial using the probabilities obtained from combining the bias parameters with the individual data for X_{IMP} and Y, as in Equation 2. The probability of selection was estimated using the bias parameters and individual data for X^*, C, and Y, as in Equation 4. Last, the logistic regression outcome model was fit, weighted by the inverse of the predicted probability of selection (Equation 5):

$$logit(P(Y = 1)) = \phi_Y + \phi_{YX}X_{IMP} + \phi_{YC}C + \phi_{YU}U_{IMP} \quad (5)$$

To account for the uncertainty of Monte Carlo procedures and to obtain the sampling distribution and confidence interval for OR_{YX}, the analysis ran on $n_{sim} = 1000$ bootstrap samples. These confidence intervals represent uncertainty due to random error, but since all the bias parameters are treated as known, fixed values, these intervals due not quantify any uncertainty due to systematic error. The median, 2.5th percentile and 97.5th percentiles from the distribution of n_{sim} OR_{YX} estimates were used for the point estimate and 95% simulation interval. Each patient in the bootstrap sample had a value of $S = 1$. Since selection is causally determined by the exposure and outcome, this approach incorporates selection bias into the samples.

In real-world applications, an investigator will not know if the obtained bias parameters are correct, so understanding the sensitivity and resilience of OR_{YX} to misspecification of the bias parameters is essential. The above analyses were therefore repeated using different, incorrect bias parameters to assess changes in OR_{YX} in response to changes in the bias parameters. The percent misspecification of each parameter is defined on a logarithmic scale; the exponentiated parameter (odds ratio) is multiplied by the percent mis-specifiation.

Performance measures

To evaluate how well bias was corrected, $\exp(\phi_{YX}) = OR_{YX}$ should approximately equal 2, corresponding to the bias-free OR_{YX} seen in the derivation of Y (Table 3). Comparisons of the estimated parameter to the correct

value were assessed based on the *Bias* $(2 - OR_{YX})$ and *RMSE* $(\sqrt{(2 - OR_{YX})^2 + SD(OR_{YX})^2})$.

Results

Both datasets were sampled with replacement over $S = 1$ and fit to Equation 1 to obtain OR_{YX} estimates biased from uncontrolled confounding, exposure misclassification and selection bias: Simulation A biased $OR_{YX} = 1.46$ (95% CI: 1.41, 1.50), Simulation B biased $OR_{YX} = 1.54$ (95% CI: 1.48, 1.60).

Results from the simultaneous multi-bias analysis are provided in Table 4. As expected, when correct bias parameters were used, bias – adjusted $\text{OR}_{YX} \approx 2$. Modifying single bias parameters by +/-25%, while leaving the others at the correct value, seemed to have a minimal effect on OR_{YX} in both simulations, with $|Bias|$ usually less than 0.1. However, mis-specification of $e^{\delta_{XY}}$ was particularly impactful in creating bias in OR_{YX}, with $|Bias|$ between 0.3 and 0.5. Larger bias was observed when multiple-bias parameters were distorted compared with single parameter mis-specification.

The sensitivity of OR_{YX} to changes in the bias parameters when all 11 bias parameters were mis-specified by a common factor was assessed (Figure 2). The degree of mis-specification and the amount of bias were positively related. In both simulations, it was found that the odds ratio resulting from multi-bias adjustment, in which each bias parameter was mis-specified by +/-25%, still produced an odds ratio estimate that better approximated the true effect when compared with the odds ratio with no bias adjustment.

Discussion

This paper introduced a novel, simultaneous approach to multiple-bias adjustment and a tutorial on how to perform this method on any combination of epidemiological biases. A simulation study using data with an exposure-outcome relationship biased by uncontrolled confounding, selection bias, and exposure misclassification confirmed that an estimate with near-zero bias was obtained when the correct bias parameters were applied. The robustness of effect estimates to distortions in the bias parameters was assessed. Single parameter mis-specification of +/-25% generally led to $|Bias| < 0.10$. In both simulations, it was observed that +/-25% mis-specification of all bias parameters produced a bias-adjusted effect estimate with less bias than the observed effect estimate with no bias adjustment. Thus, one can be confident that a biased effect estimate adjusted via simultaneous multi-bias analysis with near-accurate bias parameters is more valid than the estimate without bias adjustment, assuming biases were correctly identified in the DAG.

An obstacle to performing simultaneous multiple-bias adjustment involves deriving the large number of bias parameters. Obtaining an accurate parameter estimate can be a difficult task, particularly when the parameter has several variable dependencies. Fortunately, there are many potential strategies an investigator can pursue. Searching for parameter estimates from the literature or expert opinion may be applied.[14] One should be mindful, however, that expert opinion is subject to unique bias and influence, even if sampled in aggregate. A more efficient strategy would be to use internal or external validation data to inform the bias parameters.[23] For example, a subset of the data may have information for the uncontrolled confounder or a better-classified exposure. In this case, models for U and X may be fitted to the data subset to obtain bias parameters. A similar approach may be used to obtain the selection bias parameters if information is present for subjects who were invited but chose not to participate. Last, more advanced simulation strategies may be used to avoid having to reason about the bias parameters backwards (e.g. from X and C to U).[24] All the observed relationships can be combined with hypothesized effects to simulate a new dataset. These new data could then be fit to models to obtain the bias parameters. Some of these different strategies for obtaining bias parameters were applied in adjusting for exposure misclassification in a study of the effect of Parkinson's disease (PD) on cancer.[25]

It is important to consider that every study inherently makes assumptions regarding the biases impacting on the effect of interest.[13] Studies without bias adjustment inherently assume that if all models are correctly specified, all estimates are valid and that the only source of uncertainty in these estimates is random error. Such assumptions are generally incorrect. Any attempt to improve on these implausible assumptions is worth the effort of the investigator. Multiple effect estimates that derive from different DAGs or different bias assumptions can and should be presented.[23] This transparency allows the reader to understand the resilience of the effect estimate to various bias scenarios and can also help editors identify key areas of improvement.[26] It is possible and advisable to incorporate uncertainty into each bias parameter. Bias parameters may be represented by probability distributions instead of fixed values (i.e. probabilistic bias analysis).[13,27] This semi-Bayesian approach allows for the simulation interval of the effect estimate to incorporate uncertainty due to both random error and systematic error.[14]

It is important to consider other applications of bias analysis besides attempting to produce the best bias-adjusted effect estimate with the best bias parameters. These alternative methods can be used prior to data collection, to optimize resource allocation by identifying where

Table 4 Results of parameter mis-specification in multi-bias analysis of simulated data

Simulation A

Mis-specified bias parameter(s)	Correct value(s)	Mis-specified value(s)	% Mis-specification	Bias-adjusted OR_{YX}[a]	Bias	RMSE[b]
None	–	–	0	2.03 (1.96, 2.11)	−0.0326	0.0491
e^{α_0}	0.80	1.00	+25	2.04 (1.97, 2.12)	−0.0433	0.0568
e^{α_0}	0.80	0.60	−25	2.03 (1.95, 2.10)	−0.0257	0.0455
$e^{\alpha_{UX}}$	1.84	2.30	+25	1.97 (1.90, 2.04)	0.0306	0.0473
$e^{\alpha_{UX}}$	1.84	1.38	−25	2.13 (2.05, 2.21)	−0.1272	0.1329
$e^{\alpha_{UY}}$	2.01	2.51	+25	1.98 (1.90, 2.05)	0.0247	0.0446
$e^{\alpha_{UY}}$	2.01	1.51	−25	2.11 (2.05, 2.19)	−0.1135	0.119
e^{δ_0}	0.08	0.11	+25	2.02 (1.95, 2.09)	−0.0178	0.0392
e^{δ_0}	0.08	0.06	−25	2.05 (1.98, 2.13)	−0.0543	0.0659
$e^{\delta_{XX^*}}$	4.93	6.16	+25	2.02 (1.95, 2.09)	−0.0185	0.0409
$e^{\delta_{XX^*}}$	4.93	3.69	−25	2.04 (1.96, 2.12)	−0.0365	0.0531
$e^{\delta_{XY}}$	2.07	2.58	+25	2.47 (2.38, 2.56)	−0.4680	0.4701
$e^{\delta_{XY}}$	2.07	1.54	−25	1.58 (1.52, 1.64)	0.4215	0.4226
$e^{\delta_{XC}}$	1.44	1.80	+25	2.02 (1.96, 2.10)	−0.0242	0.0443
$e^{\delta_{XC}}$	1.44	1.08	−25	2.04 (1.97, 2.12)	−0.0411	0.0553
e^{β_0}	1.02	1.28	+25	2.03 (1.96, 2.10)	−0.0302	0.0478
e^{β_0}	1.02	0.77	−25	2.03 (1.96, 2.11)	−0.0330	0.0499
$e^{\beta_{SX^*}}$	1.19	1.48	+25	2.06 (1.99, 2.13	−0.0608	0.0717
$e^{\beta_{SX^*}}$	1.19	0.89	−25	2.00 (1.92, 2.07)	0.0044	0.0372
$e^{\beta_{SY}}$	2.15	2.69	+25	2.04 (1.97, 2.12)	−0.0404	0.0554
$e^{\beta_{SY}}$	2.15	1.62	−25	2.03 (1.95, 2.10)	−0.0228	0.0434
$e^{\beta_{SC}}$	1.04	1.30	+25	2.03 (1.97, 2.10)	−0.0326	0.0483
$e^{\beta_{SC}}$	1.04	0.78	−25	2.03 (1.96, 2.11)	−0.0333	0.0503
e^{α_0}, $e^{\alpha_{UX}}$, $e^{\alpha_{UY}}$	0.80, 1.84, 2.01	1.00, 2.30, 2.51	+25	1.92 (1.85, 1.99)	0.0781	0.0856
e^{α_0}, $e^{\alpha_{UX}}$, $e^{\alpha_{UY}}$	0.80, 1.84, 2.01	0.60, 1.38, 1.51	−25	2.18 (2.10, 2.25)	−0.1762	0.1806
e^{δ_0}, $e^{\delta_{XX^*}}$, $e^{\delta_{XY}}$, $e^{\delta_{XC}}$	0.08, 4.93, 2.07, 1.44	0.11, 6.16, 2.58, 1.80	+25	2.40 (2.32, 2.48)	−0.3958	0.3978
e^{δ_0}, $e^{\delta_{XX^*}}$, $e^{\delta_{XY}}$, $e^{\delta_{XC}}$	0.08, 4.93, 2.07, 1.44	0.06, 3.69, 1.08, 0.77	−25	1.57 (1.50, 1.65)	0.4250	0.4266
e^{β_0}, $e^{\beta_{SX^*}}$, $e^{\beta_{SY}}$, $e^{\beta_{SC}}$	1.02, 1.19, 2.15, 1.04	1.28, 1.48, 2.69, 1.30	+25	2.06 (2.00, 2.14)	−0.0630	0.0734
e^{β_0}, $e^{\beta_{SX^*}}$, $e^{\beta_{SY}}$, $e^{\beta_{SC}}$	1.02, 1.19, 2.15, 1.04	0.77, 0.89, 1.62, 0.78	−25	2.00 (1.93, 2.07)	0.0017	0.0362

Simulation B

Mis-specified parameter(s)	Correct value(s)	Mis-specified value(s)	% Mis-specification	Bias-adjusted OR_{YX}[a]	Bias	RMSE[b]
none	–	–	n/a	2.01 (1.93, 2.10)	−0.0149	0.0466
e^{α_0}	0.94	1.18	+25	2.01 (1.93, 2.10)	−0.0108	0.0456
e^{α_0}	0.94	0.71	−25	2.01 (1.92, 2.10)	−0.0106	0.0483
$e^{\alpha_{UX}}$	1.20	1.51	+25	1.99 (1.90, 2.08)	0.0147	0.0475
$e^{\alpha_{UX}}$	1.20	0.90	−25	2.04 (1.96, 2.14)	−0.0447	0.0651
$e^{\alpha_{UY}}$	1.26	1.58	+25	1.99 (1.91, 2.08)	0.0082	0.0445
$e^{\alpha_{UY}}$	1.26	0.95	−25	2.04 (1.94, 2.12)	−0.0377	0.0589
e^{δ_0}	0.04	0.05	+25	1.98 (1.91, 2.07)	0.0182	0.0448
e^{δ_0}	0.04	0.03	−25	2.05 (1.97, 2.15)	−0.0467	0.0666
$e^{\delta_{XX^*}}$	14.53	18.16	+25	1.98 (1.90, 2.07)	0.0151	0.0460
$e^{\delta_{XX^*}}$	14.53	10.90	−25	2.03 (1.94, 2.12)	−0.0305	0.0563
$e^{\delta_{XY}}$	1.78	2.23	+25	2.35 (2.25, 2.46)	−0.3546	0.3583
$e^{\delta_{XY}}$	1.78	1.34	−25	1.63 (1.56, 1.71)	0.3704	0.3724
$e^{\delta_{XC}}$	1.48	1.84	+25	2.00 (1.91, 2.07)	0.0044	0.0412
$e^{\delta_{XC}}$	1.48	1.11	−25	2.03 (1.94, 2.12)	−0.0306	0.0555
e^{β_0}	0.98	1.23	+25	2.01 (1.93, 2.10)	−0.0093	0.0447
e^{β_0}	0.98	0.74	−25	2.01 (1.93, 2.10)	−0.0106	0.0454

(Continued)

Table 4 Continued

Simulation B						
Mis-specified parameter(s)	Correct value(s)	Mis-specified value(s)	% Mis-specification	Bias-adjusted OR_{YX}[a]	Bias	RMSE[b]
$e^{\beta_{SX^*}}$	1.10	1.37	+25	2.03 (1.93, 2.12)	−0.0271	0.0539
$e^{\beta_{SX^*}}$	1.10	0.82	−25	1.99 (1.91, 2.07)	0.0139	0.0461
$e^{\beta_{SY}}$	1.29	1.62	+25	2.02 (1.93, 2.11)	−0.0181	0.0491
$e^{\beta_{SY}}$	1.29	0.97	−25	2.00 (1.92, 2.10)	−0.0023	0.0463
$e^{\beta_{SC}}$	1.02	1.28	+25	2.01 (1.92, 2.10)	−0.0094	0.0473
$e^{\beta_{SC}}$	1.02	0.77	−25	2.01 (1.92, 2.10)	−0.0143	0.0464
e^{α_0}, $e^{\alpha_{UX}}$, $e^{\alpha_{UY}}$	0.94, 1.20, 1.26	1.18, 1.51, 1.58	+25	1.95 (1.86, 2.03)	0.0521	0.0675
e^{α_0}, $e^{\alpha_{UX}}$, $e^{\alpha_{UY}}$	0.94, 1.20, 1.26	0.71, 0.90, 0.95	−25	2.03 (1.94, 2.12)	−0.0295	0.0551
e^{δ_0}, $e^{\delta_{XX^*}}$, $e^{\delta_{XY}}$, $e^{\delta_{XC}}$	0.04, 14.53, 1.78, 1.48	0.05, 18.16, 2.23, 1.84	+25	2.24 (2.16, 2.33)	−0.2442	0.2484
e^{δ_0}, $e^{\delta_{XX^*}}$, $e^{\delta_{XY}}$, $e^{\delta_{XC}}$	0.04, 14.53, 1.78, 1.48	0.03, 10.90, 1.34, 1.11	−25	1.64 (1.55, 1.72)	0.3637	0.3665
e^{β_0}, $e^{\beta_{SX^*}}$, $e^{\beta_{SY}}$, $e^{\beta_{SC}}$	0.98, 1.10, 1.29, 1.02	1.23, 1.37, 1.62, 1.28	+25	2.04 (1.96, 2.13)	−0.0437	0.0633
e^{β_0}, $e^{\beta_{SX^*}}$, $e^{\beta_{SY}}$, $e^{\beta_{SC}}$	0.98, 1.10, 1.29, 1.02	0.74, 0.82, 0.97, 0.77	−25	2.01 (1.92, 2.09)	−0.0064	0.0441

[a]Odds ratio of the effect of exposure (X) on outcome (Y).
[b]Root mean square error.

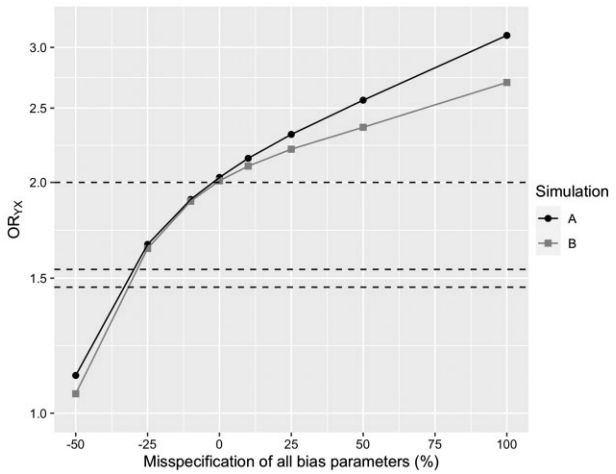

Figure 2 Multi-bias analysis results in Simulations A and B under misspecification of all bias parameters. • OR_{YX} = odds ratio of the effect of exposure (X) on outcome (Y). • Horizontal lines represent the non-biased OR_{YX} (2.00), the biased OR_{YX} in Simulation A (1.46) and the biased OR_{YX} in Simulation B (1.54)

additional data would be most impactful in minimizing bias.[28,29] One can also perform the exercise of evaluating bias strengths that would lead to a null effect estimate or an effect estimate whose confidence interval includes the null, as in the E-value.[30] By understanding the bias strengths that would 'explain away' the observed effect, one gets a better idea of how likely the effect is to be non-null. Existing tools for these analyses have mostly focused on the case of uncontrolled confounding; corresponding tools for multiple biases have seen less development.

The multi-bias analysis presented here is limited to binary exposures and outcomes, but can easily incorporate other variable types by adjusting the SEM form of the bias models. Future work should illustrate the application of this method with other model families and measures of effect (e.g. risk differences). Guidance on using this method to adjust for biases in more complex scenarios should be explicitly laid out; the examples here only evaluated the case of a single uncontrolled confounder and a single selection bias mechanism. For example, compared with the DAGs in Figure 1A or 1D, a different type of selection bias could result in S caused by U instead of Y, as could be the case in time-varying exposure or confounder settings. Last, considering the computational complexity of these methods, additional work is needed to make simultaneous bias-adjustment more accessible to researchers across disciplines. Analytical tools should be created which can assist with performing the bias adjustment.

Considering the toolkit of potential methods for bias analysis, simultaneous multi-bias analysis can serve in an important niche. In adjusting for the aggregate impact of multiple biases, this method is one of few that account for the joint impact of these biases. If one is willing to devote the effort to identify and specify the many necessary parameters and to make the assumptions underlying these parameters, simultaneous multi-bias analysis should be the preferred method. If one wants a simpler method that does a reasonable job of making this adjustment, an alternative method, such as bounding the composite bias, may be advised.[10]

Conclusion

Simultaneous multi-bias analysis is a useful tool to help understand how multiple biases could affect an observed effect estimate in observational studies. This new method

expands the field of quantitative bias analysis to help researchers provide high-quality insights into important public health questions.

Ethics approval

Approval was not needed as no human data were used.

Data availability

The data underlying this article was generated through statistical simuation. The code to generate the data can be seen in the online supplementary material.

Supplementary data

Supplementary data are available at *IJE* online.

Author contributions

P.B. and O.A. developed the method and the simulation described in this manuscript. P.B. wrote the first draft, and A.T. and O.A. critically and substantively reviewed and edited it.

Funding

P.B. was supported by the Dissertation Year Fellowship from the University of California Los Angeles (UCLA). P.B. wishes to thank his doctoral dissertation committee for providing constructive feedback that helped improve this article.

Acknowledgements

Thank you to Rick Shaffer for assistance on graphics.

Conflict of interest

P.B. started this work during his graduate training at UCLA, was formerly employed by Verana Health, and is currently employed by Valo Health. A.T. is an employee of Verana Health.

References

1. Lash T, Fox MP, Flink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. New York, NY: Springer Science and Business Media, 2009.
2. Miksad RA, Abernethy AP. Harnessing the power of Real-World Evidence (RWE): a checklist to ensure regulatory-grade data quality. *Clin Pharmacol Ther* 2018;**103**:202–05.
3. Zura R, Irwin DE, Mack CD, Aldridge ML, Mackowiak JI. Real-world evidence: a primer. *J Orthop Trauma* 2021;**35**(**Suppl 1**):S1–S5.
4. Arah OA, Chiba Y, Greenland S. Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. *Ann Epidemiol* 2008;**18**:637–46.
5. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *Int J Epidemiol* 2005;**34**:1370–76.
6. Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol* 1996;**25**:1107–16.
7. Greenland S. Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment. *Risk Anal* 2001;**21**: 579–83.
8. Thompson CA, Arah OA. Selection bias modelling using observed data augmented with imputed record-level probabilities. *Ann Epidemiol* 2014;**24**:747–53.
9. Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* 2011;**22**:42–52.
10. Smith L, Mathur M, VanderWeele T. Multiple-bias sensitivity analysis using bounds. *Epidemiology* 2021;**32**:625–34.
11. Greenland S, Robins JM. Confounding and misclassification. *Am J Epidemiol* 1985;**122**:495–506.
12. Lash TL, Silliman RA. A sensitivity analysis to separate bias due to confounding from bias due to predicting misclassification by a variable that does both. *Epidemiology* 2000;**11**:544–49.
13. Greenland S. Multiple-bias modelling for analysis of observational data (with discussion). *J R Stat Soc Ser A (Stat Soc)* 2005; **168**:267–306.
14. Rothman K, Greenland S, Lash TL. *Modern Epidemiology*. 3rd edn. London: Lippincott Williams & Wilkins, 2008.
15. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;**15**:615–25.
16. Johnson CY, Howards PP, Strickland MJ, Waller DK, Flanders WD; National Birth Defects Prevention Study. Multiple bias analysis using logistic regression: an example from the National Birth Defects Prevention Study. *Ann Epidemiol* 2018;**28**:510–14.
17. Lyles RH, Lin J. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Stat Med* 2010;**29**:2297–309.
18. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 2003;**14**: 300–06.
19. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;**10**:37–48.
20. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol* 2002;**155**:176–84.
21. Pearl J. Causal diagrams for empirical research. *Biometrika* 1995;**82**:669–88.
22. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**:550–60.
23. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014;**43**:1969–85.
24. Arah OA. Bias analysis for uncontrolled confounding in the health sciences. *Annu Rev Public Health* 2017;**38**:23–38.
25. Brendel PC. *Applications of Multi-Bias Analysis in Studies of the Associations Between Parkinson's Disease and Cancer*. https://escholarship.org/uc/item/4c50w909 (9 September 2022, date last accessed).

26. Fox MP, Lash TL. On the need for quantitative bias analysis in the peer-review process. *Am J Epidemiol* 2017;**185**:865–68.
27. Hunnicutt JN, Ulbricht CM, Chrysanthopoulou SA, Lapane KL. Probabilistic bias analysis in pharmacoepidemiology and comparative effectiveness research: a systematic review. *Pharmacoepidemiol Drug Saf* 2016; **25**:1343–53.
28. Lash TL, Ahern TP. Bias analysis to guide new data collection. *Int J Biostat* 2012;**8**:1–26.
29. Fox MP, Lash TL. Quantitative bias analysis for study and grant planning. *Ann Epidemiol* 2020;**43**:32–36.
30. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the e-value. *Ann Intern Med* 2017;**167**:268–74.