The Impact of Nondifferential Exposure Misclassification on the Performance of Propensity Scores for Continuous and Binary Outcomes

A Simulation Study

Mollie E. Wood, MPH, PhD,† Stavroula Chrysanthopoulou, PhD,‡ Hedvig M.E. Nordeng, PhD,*†§ and Kate L. Lapane, PhD‡*

Purpose: To investigate the ability of the propensity score (PS) to reduce confounding bias in the presence of nondifferential misclassification of treatment, using simulations.

Methods: Using an example from the pregnancy medication safety literature, we carried out simulations to quantify the effect of nondifferential misclassification of treatment under varying scenarios of sensitivity and specificity, exposure prevalence (10%, 50%), outcome type (continuous and binary), true outcome (null and increased risk), confounding direction, and different PS applications (matching, stratification, weighting, regression), and obtained measures of bias and 95% confidence interval coverage.

Results: All methods were subject to substantial bias toward the null due to nondifferential exposure misclassification (range: 0%–47% for 50% exposure prevalence and 0%–80% for 10% exposure prevalence), particularly if specificity was low (<97%). PS stratification produced the least biased effect estimates. We observed that the impact of sensitivity and specificity on the bias and coverage for each adjustment method is strongly related to prevalence of exposure: as exposure prevalence decreases and/or outcomes are continuous rather than categorical, the effect of misclassification is magnified, producing larger biases and loss of coverage of 95% confidence intervals. PS matching resulted in unpredictably biased effect estimates.

Conclusions: The results of this study underline the importance of assessing exposure misclassification in observational studies in the context of PS methods. Although PS methods reduce confounding bias, bias owing to nondifferential misclassification is of potentially greater concern.

Key Words: propensity score, nondifferential exposure misclassification, simulation

(*Med Care* 2018;56: e46–e53)

Propensity score (PS) methods are used to estimate causal effects in observational studies when there are systematic imbalances in confounders across treatment groups under study, under assumptions of consistency, exchangeability, positivity, and correct model specification.[1,2] This technique has gained in popularity in the medical literature,[3] and much of the recent methodologic literature has focused on what covariates should be included in the PS,[4] as well as the performance of different PS methods under different outcome types.[5,6] However, the role of treatment misclassification as a threat to the ability of the PS to reduce bias in the estimation of treatment effects has, to the best of our knowledge, not been explored, although one previous study suggested that this potential for bias should be explored and quantified.[7] Misclassification of exposure is known to cause bias, which may be toward or away from the null, depending on the type of misclassification,[8–10] and so appreciating the potential impact of misclassification of the PS is vital to understanding its operating characteristics.

To ground this methodologic exploration in the real world, we will use an example from the pregnancy medication safety literature. Research on the safety and efficacy of medication use during pregnancy poses particular exposure misclassification problems. Birth cohort studies, such as the Norwegian Mother and Child Cohort Study,[11] collect medication use data directly from mothers via prospective self-report. Several studies on the accuracy of prospective maternal recall of medication use during pregnancy suggest that while specificity is often high (values of 0.99–1.00),

From the *PharmacoEpidemiology & Drug Safety Research Group, Department of Pharmacy, School of Pharmacy; †PharmaTox Strategic Research Initiative, Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo, Norway; ‡Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA; and §Department of Child Health, Norwegian Institute of Public Health, Oslo, Norway.

Selected portions of this work were presented as a poster at the 2016 annual meeting of the Society for Epidemiologic Research and as an oral presentation at the annual meeting of the International Society for Pharmacoepidemiology.

M.E.W. and S.C. are supported by internal institutional funds. H.M.E.N is funded by the ERC StG DrugsInPregnancy (639377). K.L.L. receives salary support from investigator initiated grants from NCI (R21CA198172), NCATS (TL1TR001454), NIA (R21AG046839), NINR (R56NR015498), NIOSH (R21OH010769), NIGMS (R25GM113686), and Merck.

Reprints: Mollie E. Wood, MPH, PhD, PB 1068 Blindern, Oslo 0316, Norway. E-mail: mollie.wood@farmasi.uio.no.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.lww-medicalcare.com.

ISSN: 0025-7079/18/5608-e046

sensitivity may be low (0.17–0.41), particularly for medications taken intermittently, such as analgesics.[12,13] Comparing prescription redemptions in administrative databases[14] to self-report data often shows substantial disagreement between these information sources,[12,13] and because pregnancy is a major predictor of medication discontinuation,[15] women may incorrectly be classified as exposed when they have reduced or discontinued medication use. Many medications that women take during pregnancy may be acquired over-the-counter (OTC), or from other sources, and will not be captured in databases relying on prescription fills. In all of these scenarios, misclassification of a binary exposure is likely to be nondifferential with respect to outcome, and so has an expectation of bias toward the null over many studies, although individual studies may be biased toward or away from the null.[9,10,16]

To our knowledge, no study has examined the impact of nondifferential exposure misclassification on the performance of the most common PS methods used in the pharmacoepidemiology literature. Using a simulation study constructed with realistic parameters from the pregnancy medication safety literature, we compared the validity of estimates of the exposure effect derived from the application of PS methods under varying degrees of nondifferential exposure misclassification. Secondarily, we have compared the bias and coverage resulting from misclassified PSs across a variety of common applications of the PS. Our aim was to determine the extent to which the ability of PS methods to reduce bias was affected by nondifferential exposure misclassification, and additionally, whether some applications of PS methods perform better or worse under certain misclassification scenarios.

METHODS

No ethics review was required because this is a simulation study.

For this simulation study, we consider the case of nonsteroidal anti-inflammatory drugs (NSAID) use during pregnancy and a continuous outcome, birth weight. NSAIDs are analgesic medications available through both prescription and OTC avenues. Prior research on the safety of NSAID use during pregnancy has produced inconsistent results, with some studies suggesting an increased risk of malformations[17,18] or low birth weight,[19] while others find no effect.[20] Despite recommendations that women discontinue NSAID use during the first and third trimester in pregnancy, as many as 19% of women use NSAIDs during pregnancy[21,22] with wide variation in prevalence (7%–19%). This variation is due to whether drug utilization studies considered only prescription, or prescription plus OTC drug use as well as differences in prescribing practices between countries. Further, among persons with certain pain indications such as migraine or arthritis, prevalence of NSAID use is even higher. For studies using drug registries or administrative records, women who acquired NSAIDs OTC will not be counted as exposed, meaning that studies will consider some women unexposed when they were truly exposed (ie, decreased sensitivity). Conversely, registry and administrative data reflect only medications prescribed or dispensed, not medications actually consumed, which means that some women classified as exposed are truly unexposed (ie, decreased specificity).

Data Generation

The data were generated to closely follow realistic scenarios for NSAID exposure, pregnancy outcome, and confounders. Details on the parameters used for the simulation are outlined in Table 1. We simulated datasets that included exposure (A), confounders (X_1 through X_5), and an outcome Y; the proposed causal model is shown in Figure 1. The 5 confounders were simulated with properties similar to those found in the birthing population: X_1 (analogous to indication for NSAID use, e.g. severe pain, with prevalence 0.50),[23] X_2 (analogous to folate supplementation, with prevalence 0.60),[24] X_3 (analogous to smoking during pregnancy, with prevalence 0.15),[25] X_4 (analogous to concomitant opioid use, with prevalence 0.05),[26] X_5 (analogous to maternal age, with a mean of 30 and a SD of 5). Exposure A, with a prevalence of 50% or 10%, was simulated conditional on these confounders. The nodes A^* and U_A represent the misclassified exposure and all sources of error leading to misclassification of the exposure, respectively. Exposure was simulated as in Equation 1.

Equation 1. Model used to simulate probability of treatment, A, conditional on confounders.

$$p(A\,|\,X_1,\ \dots,\ X_5)=$$
$$[\exp(\alpha_0+X_1\alpha_1+X_2\alpha_2+X_3\alpha_3+X_4\alpha_4+X_5\alpha_5)]\,/$$
$$[1+\exp(\alpha_0+X_1\alpha_1+X_2\alpha_2+X_3\alpha_3+X_4\alpha_4+X_5\alpha_5)].$$

TABLE 1. Parameters of the Simulation Study

Variables	Prevalence or Mean (SD)	Effect Size (Treatment Model)	Parameter (Treatment Model)	Effect Size (Outcome Model)	Parameter (Outcome Model)
A	10%, 50%	—	α_0	−200 [0.7, 0]	β_A
X_1	50%	0.8 [−0.8]	α_1	−20 [0.2, −0.2]	β_1
X_2	60%	−0.05 [0.05]	α_2	−38 [0.1, −0.1]	β_2
X_3	15%	0.4 [−0.4]	α_3	−300 [0.7, −0.7]	β_3
X_4	5%	1.2 [−1.2]	α_4	−50 [0.15, −0.15]	β_4
X_5	30 (5)	0.01 [−0.01]	α_5	5 [0.05, −0.05]	β_5
Y_1	5%, 3500 (500)	—	—	—	β_0

Confounders X_1 to X_5 were simulated to have parameters analogous to indication for medication use (X_1), folic acid supplementation (X_2), smoking during pregnancy (X_3), concomitant opioid use (X_4), and maternal age (X_5). A has a prevalence similar to nonsteroidal anti-inflammatory drugs use during pregnancy, and its simulated effect on Y_1 indicates a 200 g decrease in birth weight. Y was simulated to have the mean and variation of birth weight (−200 for continuous outcomes) and the expected prevalence of low birth weight (5% for categorical outcomes) among term infants in the normal birthing population. Effect sizes for simulation parameters are shown with alternate scenarios in square parentheses: for example. the treatment effect size for the continuous outcome model was −200, with 0.7 and 0 considered as alternate scenarios. Effect sizes for categorical variables are given in log odds, for example, an odds ratio of 2.0 is equal to log odds of 0.7.

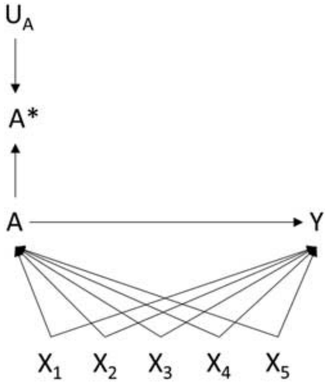

FIGURE 1. Causal diagram showing measurement error, U_A, leading to nondifferential misclassification of the exposure, A, into A^*. X_1 through X_5 are confounders of the A-Y association. In this simulation study, A is nonsteroidal anti-inflammatory drugs use in pregnancy, Y is birth weight, X_1 is indication for nonsteroidal anti-inflammatory drugs use, X_2 is folate supplementation, X_3 is smoking during pregnancy, X_4 is concomitant opioid use, and X_5 is maternal age.

We considered 2 possible outcome specifications: a continuous variable with a mean of 3500 g and a SD of 500 g, analogous to the mean and SD of birth weight (g) among term births (Equation 2), and a binary outcome with a prevalence of 5%, analogous to low birthweight (<2500 g) among live births (Equation 3).[27] The outcome variables were generated conditional on the exposure A and the confounders X_1 through X_5, from the models described in Equations 2 and 3.

Equation 2. Model used to simulate continuous outcome, Y, conditional on treatment A and confounders.

$$E(Y\,|\,A, X_1, \ldots, X_5) = \beta_0 + A\beta_A + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + X_5\beta_5.$$

Equation 3. Model used to simulate binary outcome, Y, conditional on treatment A and confounders.

$$p(Y\,|\,A, X_1, \ldots, X_5) = \left[\exp\left(\beta_0 + A\beta_A + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + X_5\beta_5\right)\right] / \left[1 + \exp\left(\beta_0 + A\beta_A + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + X_5\beta_5\right)\right].$$

For the continuous outcome, the true mean difference in birth weight was set to 200 g, a difference that would be of clinical concern.[28] Similarly, the true effect size for the binary outcome was set to an odds ratio of 2.0 (log odds of 0.7). We also considered continuous and binary outcome scenarios in which the true effect of treatment was zero. We simulated joint confounding by the confounders X_1 through X_5 to produce effect estimates that (1) overestimated the true effect size by about 15% or (2) underestimated the true effect size by about 15%. Overall, the data generation process was planned in order to show realistic, clinically meaningful true effect sizes, which were moderately biased due to confounding.

To assess the impact of varying degrees of exposure misclassification, we created misclassified exposure variables A^* from our original (correctly classified) exposure variable A. Values for sensitivity and specificity included: 1.0, 0.99, 0.97, 0.95, 0.90, 0.80, and 0.70; these values represent measured (if somewhat optimistic) values for sensitivity and specificity found in the pregnancy medication literature.[12,13] We investigated all 49 possible combinations, assuming nondifferential misclassification; because results from the 70% scenarios were uniformly poor, we have included these data only in the Supplementary Material, Supplemental Digital Content 1 (http://links.lww.com/MLR/B463).

Analytic Approaches

We first fit a PS model using logistic regression with the correctly classified exposure, A, as the dependent variable and the confounders X_1 through X_5 as independent variables; we derived the predicted probability of prenatal exposure to NSAIDs from this model. We then fit PS models for all combinations of sensitivity and specificity, resulting in 49 predicted probabilities of exposure. We used the PSs to calculate inverse probability of treatment weights, in which exposed units received weights of [1/PS] and unexposed received weights of [1/(1−PS)]. We also calculated standardized morbidity/mortality rate weights (SMRW), where exposed individuals received weights of 1 and unexposed received weights of [PS/(1−PS)]. After deriving the PS and weights, we fit outcome models in 6 ways: (1) fitting an unadjusted model; (2) adjusting for the PS as a covariate in the multivariable model; (3) matching on PS, using a nearest neighbor 1:1 matching algorithm and a caliper equal to 0.2 of the SD of the logit of the PS; (4) calculating 5 strata of the PS based on the distribution of the PS in the exposed, stratifying the outcome model to estimate the effect of exposure on outcome within each stratum, and calculating a pooled effect estimate across strata; (5) fitting an IPT-weighted model; and (6) fitting an SMR-weighted model. Matching and stratification were performed using the R package *matchit*.[29] Weighted models were fit with robust SEs using the R package *sandwich*.[30] We performed steps 1–6 on the perfectly classified exposure, A, and then repeated them using each combination of misclassified exposure, A^*, resulting in 294 estimates (6 methods×49 misclassification scenarios), and then repeated this process for scenarios with varying exposure prevalence (50%, 10%), outcome type (continuous, categorical), confounding structure (overestimate vs. underestimate of the true effect) and effect size (mean difference of −200 for continuous outcomes, log odds of 0.7 for categorical outcomes, effect size of 0 for both continuous and categorical outcomes). Results are reported as the mean difference between exposed and unexposed individuals (β), mean percent bias (difference between truth and observed, divided by truth; only reported for scenarios where the true effect was not zero), and coverage of 95% confidence intervals (CI) (percent of estimates in which the 95% CI contained the true value). For each scenario, we simulated 1000 cohorts with sample size N = 10,000. All simulations were carried out using RStudio.

RESULTS

The results of the simulation are presented by outcome type (continuous or categorical), direction of confounding (underestimate vs. overestimate), exposure prevalence (50% vs. 10%), and true effect size (increased risk vs. no effect).

Performance of PS methods varied according to the scenarios considered, with some overall trends emerging. Results for continuous and categorical outcome models, in which exposure prevalence was 10%, are presented in Figures 2 and 3, and selected results are shown in Table 2. Additional, expanded results from these and other analyses can be found in the online Supplementary Material, Supplemental Digital Content 1 (http://links.lww.com/MLR/B463). Continuous outcome models were more biased and had lower coverage than categorical outcome models across scenarios where exposure was misclassified; PS stratification generally outperformed other methods, and this advantage increased as specificity (and to a lesser extent, sensitivity) worsened: for example, at 97% sensitivity and specificity, PS stratified estimates were biased by about 15% versus biases equal to or in excess of 20% for other methods, and for 97% sensitivity and 90% specificity, stratified estimates underestimated the true effect by almost 40%, while other estimates underestimated by 45% or more (Table 2). PS adjusted and SMR-weighted models produced similar estimates regardless of the degree of misclassification, with IPT-weighted models performing marginally less well. Scenarios where exposure prevalence was 50% were less biased due to misclassification (Supplementary Figs. S2, S3, Supplemental Digital Content 1, http://links.lww.com/MLR/B463), and as misclassification increased, PS stratified models exhibited less bias and better coverage than weighted or adjusted results. Models with 50% exposure prevalence were also more susceptible to losses in sensitivity, compared with models with 10% exposure prevalence.

In Figures 2 and 3, (A) and (B) show scenarios where confounding resulted in an overestimate of the true effect. PS stratified models generally had a slight advantage over other methods for scenarios where exposure prevalence was 10%; however, increasing exposure prevalence to 50% resulted in PS-matched estimates with improved coverage.

PS-matched estimates initially appeared to outperform other methods as specificity decreased, particularly in cases where confounding was negative. However, this observation is limited to better coverage, as percent bias for matched estimates tended to be comparable to inverse probability of treatment weights, SMRW, and PS adjusted results. Examination of the sample sizes included in the PS matched samples suggests that coverage is improved due to smaller sample size and correspondingly wider CIs (Supplementary Fig. 7, Supplemental Digital Content 1, http://links.lww.com/MLR/B463); this was a particular problem for rarer (10%) exposures but was also present for the more common exposure. In addition, when confounding was negative and exposure prevalence was 50%, PS matching methods produced estimates that were biased in the opposite direction as the confounded estimates; for example, control for confounding using matching resulted in an overestimate of the true effect, even when no misclassification was present (Supplementary Figs. S3, S4, C, and D, Supplemental Digital Content 1, http://links.lww.com/MLR/B463) and when the true effect of exposure was null (Supplementary Figs. S5 and S6, C, and D, Supplemental Digital Content 1, http://links.lww.com/MLR/B463). Overall, PS matching estimates were less predictably biased than other methods, and showed substantial sensitivity to misclassification, exposure prevalence, and confounding structure.

We observed one other phenomenon that deserves attention. Examining the graph of coverage shown in Figure 2B shows a “crossing of the curves,” in which PS methods are close to 95% and then decline, whereas unadjusted estimates begin close to 40% higher than the true effect, decrease until at specificity values of 99% and 97% their bias is close to zero, and then decline. This pattern, although obviously limited to a specific scenario in which confounding results in an overestimation, illustrates the balance between 2 sources

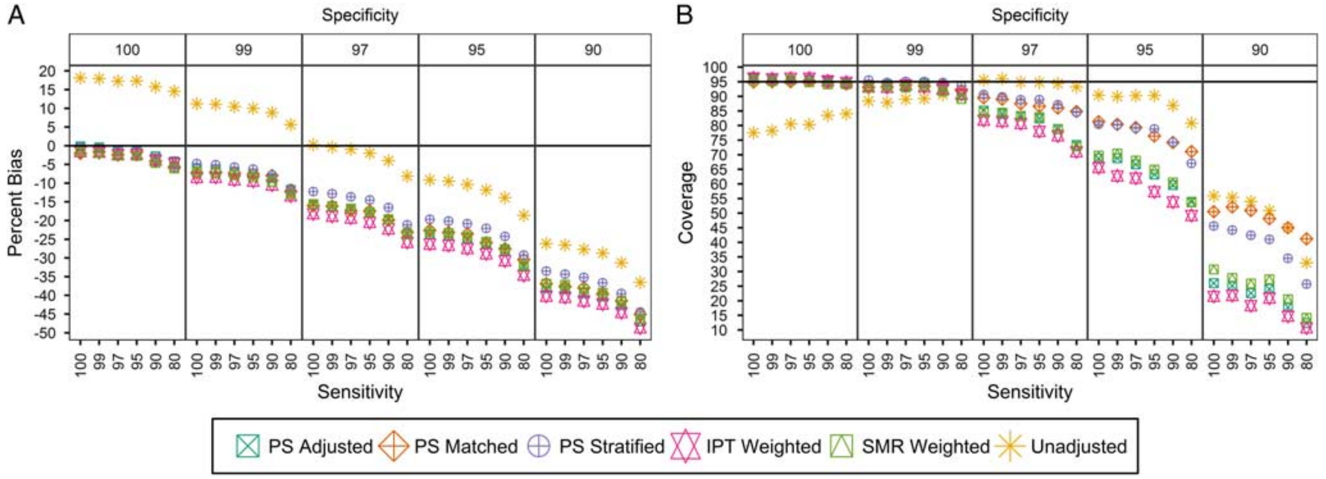

FIGURE 2. Results from continuous outcome models showing (A) percent bias and (B) coverage, for positive confounding scenarios. Results are for propensity score matching, regression adjustment, stratification, weighted, and unadjusted models, under varying values of sensitivity and specificity, with a true mean difference of −200 and exposure prevalence set to 10%. Percent bias is calculated as [(observed−truth)/truth]×100%. Coverage is defined as the percent of simulations in which the confidence interval of the effect estimate contained the true effect. IPT indicates inverse probability of treatment; PS, propensity score; SMR, standardized morbidity/mortality rate.

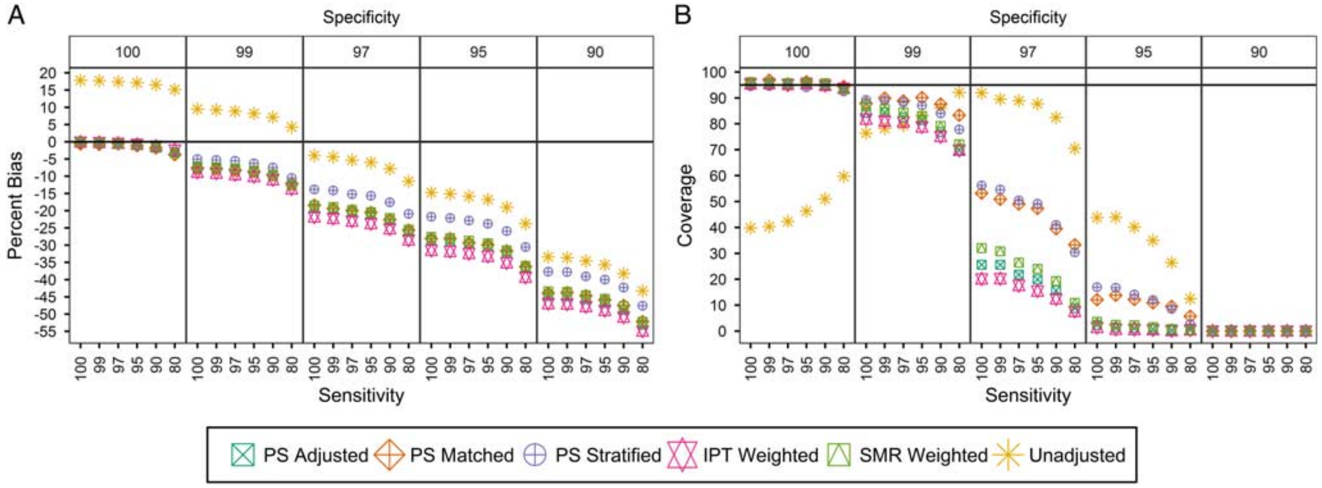

FIGURE 3. Results from categorical outcome showing (A) percent bias and (B) coverage, for positive confounding scenarios. Results are for propensity score matching, regression adjustment, stratification, weighted, and unadjusted models, under varying values of sensitivity and specificity, with the true log odds of *Y* set to 0.7 and exposure prevalence set to 10%. Percent bias is calculated as [(observed–truth)/truth]×100%. Coverage is defined as the percent of simulations in which the confidence interval of the effect estimate contained the true effect. IPT indicates inverse probability of treatment; PS, propensity score; SMR, standardized morbidity/mortality rate.

of systematic error: bias due to misclassification, and bias due to confounding.

DISCUSSION

We compared the performance of 5 common implementations of PS methods, including adjustment, 2 methods of weighting, matching, and stratification, for varying exposure prevalence, and for both continuous and categorical outcome models. We found that all methods were vulnerable to bias due to misclassification of exposure, and that losses of specificity had a greater impact on effect estimates than losses of sensitivity. PS matching more often produced estimates with worse coverage and greater bias, although in the presence of even moderate misclassification, all methods showed substantial loss of coverage and increase in bias.

The effects of misclassification were more extreme for an exposure with 10% prevalence compared with 50% prevalence. A recent simulation study on applying PS methods to rare exposures found that an exposure prevalence of 10% produced estimates with low bias and acceptable variability in comparison to very rare exposure[31]; however, this study did not consider systematic error due to misclassification. In reality, studies with exposure prevalence of 10% may be more biased than expected if exposure data are misclassified.[32] In addition, we found that estimates from categorical outcome models less biased than estimates from continuous outcome models where exposure prevalence and misclassification were similar.

It is unsurprising that PS stratification is less vulnerable to misclassification than PS matching. In PS matching, an exposed individual is matched to unexposed individuals, conditional on the PS. In the case of moderate to low specificity and low sensitivity, few truly exposed individuals are included in the outcome analysis, which will clearly result in bias toward the null. Prior research on the performance of matching estimators compared with other PS methods is conflicting, with one study showing that matching on the PS is preferable to stratification for purposes of reducing bias due to confounding[33]; however, these studies assumed perfect classification of exposure. Other recent work suggests that PS matching can substantially increase bias compared with other estimators.[34] Our results are more in line with the latter study, and suggest that other sources of bias, such as misclassification, should be considered when selecting an adjustment method.

It is less clear why PS stratification should outperform PS adjustment or weighting. PS stratification methods estimate the treatment effect within each stratum, and so fitting the parametric model within each stratum relies only on local, rather than global assumptions; this increases the robustness of the estimate,[29] and could explain why PS stratification methods appear less vulnerable to bias due to misclassification of the exposure.

A possible, and alluring, conclusion to be drawn from these results, given that the unadjusted estimates are often less biased, with better coverage than the adjusted estimates, particularly for poor sensitivity and specificity, is that researchers are better off not adjusting for confounders. Although this is true in the case of this simulation, when the magnitude of bias due to confounding is known and fixed, this conclusion should not be generalized to observational research, where the magnitude of this bias is unknown. Rather, this finding illustrates a long-understood phenomenon when working with real data: that effect estimates are subject to multiple sources of bias.

PS stratification, adjustment, and inverse probability of treatment weighting estimate the average treatment effect in

TABLE 2. Selected Results for Continuous and Categorical Outcome Models With 10% Exposure Prevalence

	PS Adjusted			PS Matched*			PS Stratified[†]			IPT Weighted[‡]			SMR Weighted[§]		
Sens/Spec	**β[‖]**	**Bias (%)[¶]**	**Cov.[#]**	**β**	**Bias (%)**	**Cov.**	**β**	**Bias (%)**	**Cov.**	**β**	**Bias (%)**	**Cov.**	**β**	**Bias (%)**	**Cov.**
Continuous outcome models															
1.00/1.00	−199.9	−0.1	95.7	−198.7	−0.7	95.2	−200.5	0.2	94.5	−199.8	−0.1	95.4	−199.8	−0.1	96.1
1.00/0.99	−185.1	−7.4	84.2	−184.5	−7.7	87.9	−189.9	−5.0	89.2	−182.3	−8.8	81.8	−185.9	−7.1	86.8
1.00/0.97	−161.4	−19.3	25.6	−163.2	−18.4	53.2	−172.3	−13.8	56.2	−156.3	−21.8	20.1	−163.0	−18.5	32.0
0.99/1.00	−199.6	−0.2	95.4	−198.5	−0.8	96.8	−200.1	0.1	94.6	−199.5	−0.2	95.5	−199.5	−0.3	96.0
0.97/1.00	−198.9	−0.5	95.2	−198.9	−0.5	94.8	−199.5	−0.3	94.8	−199.1	−0.5	95.2	−198.8	−0.6	95.7
0.95/1.00	−198.2	−0.9	95.5	−197.7	−1.2	96.2	−198.5	−0.7	94.1	−198.5	−0.8	95.5	−198.0	−1.0	96.1
0.90/1.00	−196.8	−1.6	95.0	−196.4	−1.8	94.9	−197.0	−1.5	94.6	−197.4	−1.3	94.8	−196.5	−1.8	95.7
0.99/0.99	−184.5	−7.7	84.4	−184.4	−7.8	89.9	−189.4	−5.3	89.0	−181.8	−9.1	81.0	−185.2	−7.4	85.7
0.99/0.97	−160.6	−19.7	25.6	−161.2	−19.4	50.9	−171.7	−14.2	54.6	−155.5	−22.2	20.2	−162.2	−18.9	30.8
0.97/0.99	−183.8	−8.1	82.4	−183.6	−8.2	88.8	−189.0	−5.5	88.3	−180.9	−9.5	80.7	−184.4	−7.8	84.4
0.97/0.97	−159.0	−20.5	21.8	−160.0	−20.0	49.0	−169.6	−15.2	50.4	−153.9	−23.0	17.6	−160.5	−19.7	26.4
0.95/0.99	−182.5	−8.7	81.5	−182.7	−8.6	90.1	−187.5	−6.3	87.1	−179.9	−10.1	78.8	−183.2	−8.4	83.0
0.95/0.97	−157.6	−21.2	19.9	−159.0	−20.5	47.3	−168.7	−15.7	49.1	−152.5	−23.7	15.4	−159.2	−20.4	24.0
0.90/0.99	−180.4	−9.8	76.8	−180.8	−9.6	87.6	−185.0	−7.5	84.0	−178.0	−11.0	75.1	−180.8	−9.6	79.2
0.90/0.97	−154.4	−22.8	15.7	−154.8	−22.6	39.5	−164.8	−17.6	40.9	−149.5	−25.3	12.4	−155.8	−22.1	19.3
Categorical outcome models															
1.00/1.00	0.70	−0.2	96.1	0.69	−2.0	94.9	0.70	−0.4	96.1	0.69	−1.6	96.3	0.69	−1.7	95.5
1.00/0.99	0.66	−6.2	93.5	0.65	−7.3	93.1	0.67	−4.8	95.5	0.64	−8.5	93.2	0.65	−7.0	92.9
1.00/0.97	0.59	−15.6	85.1	0.59	−16.0	89.5	0.61	−12.3	90.6	0.57	−18.2	81.8	0.59	−15.6	84.4
0.99/1.00	0.70	−0.4	95.7	0.69	−1.6	95.0	0.69	−0.7	96.0	0.69	−1.8	96.0	0.69	−1.9	95.2
0.97/1.00	0.69	−1.1	95.7	0.68	−2.4	95.0	0.69	−1.6	95.7	0.68	−2.4	96.2	0.68	−2.7	95.5
0.95/1.00	0.69	−1.1	95.7	0.69	−1.7	95.3	0.69	−1.6	95.9	0.68	−2.3	96.2	0.68	−2.7	94.8
0.90/1.00	0.68	−2.9	95.4	0.67	−4.2	94.7	0.68	−3.5	95.6	0.67	−3.8	95.3	0.67	−4.6	94.1
0.99/0.99	0.66	−6.4	94.1	0.65	−7.1	93.2	0.66	−5.1	94.7	0.64	−8.5	93.1	0.65	−7.2	93.5
0.99/0.97	0.59	−16.2	84.0	0.59	−16.1	89.0	0.61	−12.9	89.8	0.57	−18.9	81.4	0.59	−16.2	84.4
0.97/0.99	0.65	−7.0	93.4	0.65	−7.3	93.9	0.66	−5.8	95.1	0.64	−9.1	93.9	0.64	−7.9	93.0
0.97/0.97	0.58	−16.8	83.3	0.58	−17.3	87.5	0.60	−13.7	88.8	0.56	−19.3	80.7	0.58	−16.8	83.2
0.95/0.99	0.65	−7.5	94.5	0.65	−7.6	93.5	0.66	−6.3	95.0	0.63	−9.5	93.2	0.64	−8.4	93.6
0.95/0.97	0.58	−17.8	82.4	0.58	−17.5	86.6	0.60	−14.5	88.8	0.56	−20.5	78.0	0.58	−17.8	82.9
0.90/0.99	0.64	−8.7	93.1	0.64	−8.0	93.9	0.65	−7.8	94.7	0.63	−10.5	92.6	0.63	−9.6	91.7
0.90/0.97	0.56	−19.7	78.8	0.56	−19.9	86.0	0.58	−16.5	87.1	0.54	−22.2	76.5	0.56	−19.8	78.7
1.00/1.00	0.70	−0.2	96.1	0.69	−2.0	94.9	0.70	−0.4	96.1	0.69	−1.6	96.3	0.69	−1.7	95.5

*1:1 matching on caliper equal to 20% of the SD of the PS.

[†] Five strata based on distribution of the propensity score within the exposed group.

[‡] IPT weighted analyses where exposed individuals received weights of [1/PS] and unexposed received weights of [1 /(1−PS)].

[§] SMR weighted analyses where exposed individuals received weights of 1 and unexposed received weights of [PS/(1−PS)].

[‖] Observed effect of exposure, *A*, on outcome *Y*, expressed as mean difference between exposed and unexposed (for continuous outcome models) and log odds (log odds of 0.7 is equal to an odds ratio of 2.0) for categorical outcome models.

[¶] Percent bias, calculated as: $[(\beta_{\mathrm{TRUE}}-\beta_{\mathrm{OBSERVED}})/\beta_{\mathrm{TRUE}}]\times 100$.

[#] Percentage of simulations in which the true effect (−200 for continuous outcome models and 0.7 for categorical outcome models) is included in the 95% confidence interval.

Cov indicates coverage; IPT, inverse probability of treatment; PS, propensity score; Sens, sensitivity; SMR, standardized mortality/morbidity rate; Spec, specificity.

the population (ATE), or the effect we would expect to see if all exposed individuals were unexposed. Matching and standardized morbidity/mortality rate weighting, by contrast, estimate the average treatment effect in the treated (ATT). One recent study examining the performance of different PS methods for a rare exposure found that ATT estimators were more reliable than ATE estimators,[31] which is not consistent with our findings, although this may be explained by the fact that we simulated our outcome model in a full cohort (ATE). Further research should seek to clarify whether ATE or ATT measures are more susceptible to exposure misclassification.

This study has several limitations that should be kept in mind when considering the results. As with all simulation studies, this study is likely an oversimplification of reality. We simulated independent, rather than correlated confounders, and examined scenarios where only the exposure, not the confounders, was misclassified. One prior study on the impact of misclassification of confounders included in the PS found that even small levels of covariate measurement error reduced the ability of the PS to control confounding; further, higher correlation among covariates led to increased bias.[35] This suggests that we might expect to see more extreme levels of bias, if we had used a more complex confounding structure. We elected to limit our study to misclassification of the exposure, rather than exposure and confounders, but future studies should examine the impact of joint misclassification of exposure, as well as differential exposure misclassification with respect to outcome, on effect estimation within the PS context.

This is not the first study to have observed and described a problem of exposure misclassification in the pregnancy medication literature,[12,13] and others have noted that misclassification of exposure in epidemiologic studies is an endemic and serious problem in the field.[16] Indeed, the idea that misclassification of exposure will result in bias of effect estimates, likely toward the null, is not new in observational research,[8,9,36] and various methods, including regression calibration and multiple imputation as well as probabilistic bias analysis, have emerged to address this problem,[37–41] although the application of these techniques has not yet been tested in PS methods. Further, studies of medication safety during pregnancy have recognized,[42] and in some cases taken steps to correct for,[43] exposure misclassification. Our study adds to the current literature by underlining the importance of considering multiple sources of systematic bias, not just confounding, when using PS methods, particularly PS matching.

It is important to note that values of sensitivity and specificity that we refer to as "low" are in fact common in studies of medication use during pregnancy,[12,13] and that for some medications such as OTC analgesics, both maternal recall and capture in automated electronic databases may be far worse than the "worst case" 80% scenario we used in this study. Nondifferential misclassification tends to result in bias toward the null,[16] so this kind of systematic error will generally not result in false-positive studies. However, because studies of safety and efficacy of drugs used in pregnancy are almost exclusively performed using observational studies, the fact that we are certainly failing to detect meaningful risks should be of major concern.

REFERENCES

1. Rosenbaum BYPR, Rubin DB, Rosenbaum PR, et al. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41–55.
2. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using score on the propensity subclassification. *Am Econ Rev*. 1984;79:516–524.
3. Borah BJ, Moriarty JP, Crown WH, et al. Applications of propensity score methods in observational comparative effectiveness and safety research: where have we come and where should we go? *J Comp Eff Res*. 2014;3:63–78.
4. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163:1149–1156.
5. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med*. 2013;32: 2837–2849.
6. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*. 2007;26:734–753.
7. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. *Stat Methods Med Res*. 2009;18:67–80.
8. Barron BA. The effects of misclassification on the estimation of relative risk. *Biometrics*. 1977;33:414–418.
9. Copeland KT, Checkoway H, McMichael AJ, et al. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol*. 1977; 105:488–495.
10. Höfler M. The effect of misclassification on the estimation of association: a review. *Int J Methods Psychiatr Res*. 2005;14:92–101.
11. Magnus P, Irgens LM, Haug K, et al. Cohort profile: the Norwegian Mother and Child Cohort Study (MoBa). *Int J Epidemiol*. 2006;35: 1146–1150.
12. van Gelder MMHJ, van Rooij IALM, de Walle HEK, et al. Maternal recall of prescription medication use during pregnancy using a paper-based questionnaire. *Drug Saf*. 2013;36:43–54.
13. Skurtveit S, Selmer R, Tverdal A, et al. Drug exposure: inclusion of dispensed drugs before pregnancy may lead to underestimation of risk associations. *J Clin Epidemiol*. 2013;66:964–972.
14. Palmsten K, Huybrechts KF, Mogun H, et al. Harnessing the Medicaid Analytic eXtract (MAX) to evaluate medications in pregnancy: design considerations. *PLoS One*. 2013;8:e67405.
15. Lupattelli A, Spigset O, Twigg MJ, et al. Medication use in pregnancy: a cross-sectional, multinational web-based study. *BMJ Open*. 2014;4: e004365.
16. Jurek AM, Greenland S, Maldonado G, et al. Proper interpretation of non-differential misclassification effects: expectations vs observations. *Int J Epidemiol*. 2005;34:680–687.
17. Ofori B, Oraichi D, Blais L, et al. Risk of congenital anomalies in pregnant users of non-steroidal anti-inflammatory drugs: a nested case-control study. *Birth Defects Res B Dev Reprod Toxicol*. 2006;77: 268–279.
18. Ericson A, Källén BA. Nonsteroidal anti-inflammatory drugs in early pregnancy. *Reprod Toxicol*. 2001;15:371–375.
19. Nezvalová-Henriksen K, Spigset O, Nordeng H. Effects of ibuprofen, diclofenac, naproxen, and piroxicam on the course of pregnancy and pregnancy outcome: a prospective cohort study. *Br J Obstet Gynaecol*. 2013;8:14–18.
20. Nielsen GL, Sørensen HT, Larsen H, et al. Risk of adverse birth outcome and miscarriage in pregnant users of non-steroidal anti-inflammatory drugs: population based observational study and case-control study. *BMJ*. 2001;322:266–270.
21. Glover DD, Amonkar M, Rybeck BF, et al. Prescription, over-the-counter, and herbal medicine use in a rural, obstetric population. *Am J Obstet Gynecol*. 2003;188:1039–1045.
22. Cleves MA, Savell VH, Raj S, et al. Maternal use of acetaminophen and nonsteroidal anti-inflammatory drugs (NSAIDs), and muscular ventricular septal defects. *Birth Defects Res A Clin Mol Teratol*. 2004;70: 107–113.
23. Aegidius K, Zwart J-A, Hagen K, et al. The effect of pregnancy and parity on headache prevalence: the Head-HUNT study. *Headache*. 2009; 49:851–859.
24. Suren P, Roth C, Bresnahan M, et al. Association between maternal use of folic acid supplements and risk of autism spectrum disorders in children. *J Am Med Assoc*. 2013;309:570–577.
25. Shipton D, Tappin DM, Vadiveloo T, et al. Reliability of self reported smoking status by pregnant women for estimating smoking prevalence: a retrospective, cross sectional study. *BMJ*. 2009;339:b4347.
26. Bateman BT, Hernandez-Diaz S, Rathmell JP, et al. Patterns of opioid utilization in pregnancy in a large cohort of commercial insurance beneficiaries in the United States. *Anesthesiology*. 2014;120: 1216–1224.
27. Skjaerven R, Gjessing HK, Bakketeig LS. Birthweight by gestational age in Norway. *Acta Obstet Gynecol Scand*. 2000;79:440–449.
28. Bacci S, Bartolucci F, Chiavarini M, et al. Differences in birthweight outcomes: a longitudinal study based on siblings. *Int J Environ Res Public Health*. 2014;11:6472–6484.
29. Ho DE, Imai K, King G, et al. MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Softw*. 2011;42:1–28.
30. Zeileis A. Object-oriented computation of sandwich estimators. *J Stat Softw*. 2006;16:1–16.
31. Hajage D, Tubach F, Steg PG, et al. On the use of propensity scores in case of rare exposure. *BMC Med Res Methodol*. 2016;16:38.
32. Kelsey JL, Whittemore A, Evans AS, et al. *Methods in Observational Epidemiology*, 2nd ed. New York, New York: Oxford University Press; 1996.
33. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Stat Med*. 2006;25:2084–2106.
34. King G, Nielsen R. Why propensity score should not be used for matching. 2016. Available at: http://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-formatching. Accessed January 5, 2017.
35. Rodríguez de Gil P, Bellara AP, Lanehart RE, et al. How do propensity score methods measure up in the presence of measurement error? A Monte Carlo study. *Multivariate Behav Res*. 2015;3171:1–13.
36. Bross I. Misclassification in 2 X 2 tables. *Biometrics*. 1954;10:478–486.
37. MacLehose RF, Olshan AF, Herring AH, et al. Bayesian methods for correcting misclassification. *Epidemiology*. 2009;20:27–35.

38. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology*. 2003;14:451–458.
39. Lash TL, Fox MP, MacLehose RF, et al. Good practices for quantitative bias analysis. *Int J Epidemiol*. 2014;43:1–17.
40. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *Int J Epidemiol*. 2005;34:1370–1376.
41. Corbin M, Haslett S, Pearce N, et al. A comparison of sensitivity-specificity imputation, direct imputation and fully Bayesian analysis to adjust for exposure misclassification when validation data are unavailable. *Int J Epidemiol*. 2017;46:1–10.
42. Ehrenstein V, Sørensen HT, Bakketeig LS, et al. Medical databases in studies of drug teratogenicity: methodological issues. *Clin Epidemiol*. 2010;2:37–43.
43. Ahrens K, Lash TL, Louik C, et al. Correcting for exposure misclassification using survival analysis with a time-varying exposure. *Ann Epidemiol*. 2012;22:799–806.