Published in final edited form as:
Am J Epidemiol. 2007 May 15; 165(10): 1110–1118.

Performance of propensity score calibration – a simulation study

Til Stürmer[1,2], **Sebastian Schneeweiss**[1,3], **Kenneth J Rothman**[1,4,5], **Jerry Avorn**[1], and **Robert J. Glynn**[1,2,6]

1 *Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA.*

2 *Division of Preventive Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA.*

3 *Department of Epidemiology, Harvard School of Public Health, Boston, MA.*

4 *Department of Epidemiology, Boston University School of Public Health, Boston, MA.*

5 *Research Triangle Institute, Research Triangle Park, NC.*

6 *Department of Biostatistics, Harvard School of Public Health, Boston, MA.*

Abstract

Confounding can be a major source of bias in non-experimental research. The authors recently introduced propensity score calibration (PSC), which combines propensity scores (PS) and regression calibration to address confounding by variables unobserved in the main study by using variables observed in a validation study. Here, the authors assess the performance of PSC using simulations in settings with and without violation of the key assumption of PSC: that the error-prone PS estimated in the main study is a surrogate for the gold-standard PS (i.e. contains no additional information on the outcome). The assumption can be assessed if data on the outcome are available in the validation study. If data are simulated allowing for surrogacy to be violated, results largely depend on the extent of violation. If surrogacy holds, PSC leads to bias reduction between 74 and 106 percent (>100 percent representing an overcorrection). If surrogacy is violated, PSC can lead to an increase in bias. Surrogacy is violated when the direction of confounding of the exposure-disease association caused by the unobserved variable(s) differs from that of the confounding due to observed variables. When surrogacy holds, PSC is a useful approach to adjust for unmeasured confounding using validation data.

Keywords

bias (epidemiology); cohort studies; confounding factors (epidemiology); epidemiologic methods; propensity score calibration; research design

---

Confounding can be a major source of bias in non-experimental research. Studies often lack measures of important potential confounders, such as smoking and body mass index in pharmacoepidemiologic studies that use claims data, or laboratory or blood pressure measurements in questionnaire-based studies. Various methods have been proposed to assess the sensitivity of observed associations to the possible effect of unobserved confounders (1-12), but only one of these can address the joint confounding due to multiple unobserved confounders (12). We recently introduced propensity score calibration (PSC, 13), combining

Author for correspondence: Til Stürmer, MD, MPH Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120 Phone: +1 617 278 0627 Fax: +1 617 232 8602 Email: til.sturmer@post.harvard.edu

propensity scores (14) and regression calibration developed to correct for measurement error (15,16). Our goal was to address the joint confounding by variables unobserved in the main cohort study by using variables observed in a cross-sectional validation study. We previously demonstrated that this method worked well in one specific pharmacoepidemiologic example, assessing the effect of nonsteroidal anti-inflammatory drugs on short-term all-cause mortality (13), without requiring outcome information in the validation study.

As we have noted (13), PSC, like regression calibration, is dependent on the assumption that the error-prone variable is a surrogate for the gold-standard variable, i.e. that the error-prone propensity score (PS) is independent of disease given the gold-standard PS and exposure (17,18). Thus, under surrogacy, the error-prone PS serves as a proxy for the gold-standard PS with measurement error that is independent of the outcome. Surrogacy is plausible in many settings, especially when the gold-standard and error-prone variables are observed at baseline in a cohort study in which the disease outcome occurs later in time (17). For example, it is plausible that a single day's blood pressure contributes no information on incidence of cardiovascular disease beyond that given by true long-term blood pressure (17). By contrast, self-reported values of total cholesterol, which might be considered as surrogates for (unavailable) serum cholesterol values, have been observed to be stronger predictors of cardiovascular outcomes compared with measured serum cholesterol in the Women's Health Study (unpublished data by RJG). Thus, self-reported cholesterol is not a surrogate for measured cholesterol in this setting and regression calibration to correct for measurement error in the self-reported cholesterol values, based on measured serum cholesterol in a validation study, would be invalid because surrogacy is violated.

We here present the results of a simulation study assessing the performance of PSC under a wide range of parameter constellations and in settings with and without violation of surrogacy, and we discuss the meaning of surrogacy using practical examples.

METHODS

Propensity score calibration

Assume a main cohort study with a dichotomous exposure of interest, A, a dichotomous outcome of interest, Y, and information on two confounders, X_1 and X_2. An additional third confounder, C, is observed in a separate validation study only, together with exposure A and confounders X_1 and X_2.

To control for confounding in the main study, we first estimate the propensity score given the two observed confounders, X_1 and X_2, by fitting a logistic regression model with exposure as the dependent variable. Since this propensity score is estimated without information on the third confounder C, we call this the error-prone (EP) propensity score:

$$PS_{EP} = \Pr(A = 1 \mid X_1, X_2) = \left(1 + \exp\left\{-\left(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2\right)\right\}\right)^{-1}. \quad (1)$$

This estimated propensity score is then used as a continuous summary confounding variable to control for confounding in the main study. We use a logistic model to approximate the cumulative incidence of disease:

$$\Pr\left[Y = 1 \mid A, PS_{EP}\right] = \left(1 + \exp\left\{-\left(\beta_0 + \beta_1 A + \beta_2 PS_{EP}\right)\right\}\right)^{-1}. \quad (2)$$

In the validation sample, we then estimate both PS_{EP}, as a function of X_1 and X_2 (equation 1), and the gold-standard (GS) propensity score PS_{GS}, as a function of X_1,X_2, and C. Following the general notation introduced for the PS, we define

$$PS_{GS} = \Pr(A = 1 \mid X_1, X_2, C) = \left(1 + \exp\left\{-\left(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 C\right)\right\}\right)^{-1} \quad (3)$$

as the gold-standard propensity score. The linear measurement error model which is the basis for regression calibration (15,16) then is:

$$E\left[PS_{GS} \mid A, PS_{EP}\right] = \delta_0 + \delta_1 A + \delta_2 PS_{EP}. \quad (4)$$

Assuming that the outcome is a function of the exposure and PS_{GS}

$$Pr\left[Y = 1 \mid A, PS_{GS}\right] = \left(1 + \exp\left\{-\left(\eta_0 + \eta_1 A + \eta_2 PS_{GS}\right)\right\}\right)^{-1}, \quad (5)$$

the regression calibration adjusted estimator for the effect of A under the assumption of no additional unmeasured confounding given PS_{GS} then is (13):

$$\hat{\eta}_1 = \hat{\beta}_1 - \hat{\delta}_1 \hat{\beta}_2 / \hat{\delta}_2 \quad (6)$$

based on the estimates from equations 2 ($ß_1$ $ß_2$) and 4 (δ_1, δ_2). We used the logistic link in equations 2 and 5 despite the non-collapsibility of the odds ratio under exchangeability of exposed and unexposed given PS_{GS}, because regression calibration has not yet been evaluated for relative risk or Poisson outcome models. Corrected estimates for the variances account for the additional uncertainty caused by the estimation of δ in the validation study.

Regression calibration can also be implemented as a single imputation of the gold-standard variable based on the parameters of the measurement error model in the validation study (equation 4) and the values of X_1,X_2, and A observed in the main study (17). Since in PSC the gold-standard variable is a propensity score, single imputation of the PS_{GS} makes it possible to implement matching on or stratification by this imputed PS, rather than controlling for the PS as a single continuous covariate in the outcome model (19). Analyses matched on the PS might be advantageous, since they are not based on comparisons outside a common range of PS for exposed and unexposed (19). Stratification by the PS can also be restricted to this common range.

For these reasons we implemented PSC in all simulations by first imputing missing values of PS_{GS} based on equation 4 (i.e. as a function of exposure and PS_{EP}, but not disease outcome) rather than by using equation 6. We then matched a single unexposed observation to every exposed observation on this imputed value of PS_{GS} using greedy matching (20). Greedy matching starts using a very narrow caliper of the PS (to the 5^{th} decimal place) to find an unexposed match for every exposed observation and if unsuccessful widens the caliper in 1 decimal steps up to the first decimal place (20). Greedy matching is a frequently-used algorithm in this setting (21) because it achieves close matching with a high proportion of exposed observations for whom an unexposed match can be found. The proportion of exposed observations that could be matched to unexposed ones on the imputed PS_{GS} is an inverse function of the ability of the PS to predict exposure and was above 85 percent in most scenarios (range: 72 to 99 percent). High values are necessary for a causal contrast ("what would have happened to the exposed had they been unexposed"). The values in our simulations are well within the range observed in published applications of propensity score analyses (21).

The exposure-outcome association in matched pairs was then estimated using conditional logistic regression to increase efficiency. To obtain 95 percent confidence limits for this estimate, we took 1,000 bootstraps sampling on matched pairs with replacement that were again analyzed with conditional logistic regression. We used the empirical distribution of these estimates (2.5^{th} and 97.5^{th} percentiles) to assign a lower and an upper bound of the confidence interval to the PSC estimate and assessed whether the true value of the log odds ratio of the exposure-outcome association was covered by this non-parametric confidence interval.

Simulation study

Let the exposure of interest, A, and the outcome of interest, Y, both be dichotomous variables. The confounders X_1, X_2, and C are independent standard normal variables with a mean of 0 and unit variance. The probability that an observation is exposed (A=1) given confounders X_1, X_2, and C corresponds to PS_{GS} (equation 3).

The probability that an observation has the outcome (Y=1) given the exposure A and PS_{GS} is given by equation 5. Using this model, the association between individual confounders X_1, X_2, C and disease is defined by their association with exposure and the association of PS_{GS} with disease. In particular, the association between the confounder C and disease cannot be varied independently of the confounders X_1 and X_2.

In these simulations with disease as a function of A and PS_{GS} (equation 5), surrogacy (13,17, 18) of PS_{EP} is present by design, because PS_{EP} is based on a subset of covariates contained in PS_{GS} and disease is a (log) linear function of PS_{GS} (and A) only.

To allow surrogacy to be violated, we conducted a second set of simulations where the expected value of the dichotomous disease outcome Y given the exposure A as well as the confounders X_1, X_2, and C is defined as

$$\Pr\left[Y = 1 \mid A, X_1, X_2, C\right] = \left(1 + \exp\left\{-\left(\theta_0 + \theta_1 A + \theta_2 X_1 + \theta_3 X_2 + \theta_4 C\right)\right\}\right)^{-1}. \quad (7)$$

Because the disease is now no longer simulated as a function of PS_{GS}, there are fewer constraints with respect to the association between PS_{EP} and the disease outcome. Thus, in contrast to the first set of simulations, surrogacy might be violated in the second set.

Using these expected values, we simulated 1,000 datasets for each parameter constellation. Although simulated for the whole dataset, the confounder C is deleted from the main study and only observed in a random validation sample, whereas X_1 and X_2 are observed in the main study and the validation sample.

Since our validation study contains outcome information, we are able to assess surrogacy. Following the logistic form of equation 7, we fitted a logistic regression model with the disease outcome Y as a function of A, PS_{GS}, and PS_{EP}

$$\Pr\left[Y = 1 \mid A, PS_{GS}, PS_{EP}\right] = \left(1 + \exp\left\{-\left(\nu_0 + \nu_1 A + \nu_2 PS_{GS} + \nu_3 PS_{EP}\right)\right\}\right)^{-1}. \quad (8)$$

In the absence of a specific test for surrogacy, we used two measures: first, we performed a likelihood ratio test (LRT) for the predictive value of PS_{EP} independent of PS_{GS} and A, i.e. comparing the full model (equation 8) to a model without PS_{EP}. Second, we assessed the percent of the variance in Y explained by PS_{GS} and PS_{EP} which is due to PS_{GS}. This ratio of pseudo R-squares was calculated as the ratio of the likelihood-ratio comparing logistic regression model $\text{logit}(Y) = \nu'_0 + \nu'_1 A + \nu'_2 PS_{GS}$ with the nested logistic regression model $\text{logit}(Y) = \nu''_0 + \nu''_1 A$ and of the likelihood-ratio comparing the full model (equation 8) with the nested logistic regression model $\text{logit}(Y) = \nu''_0 + \nu''_1 A$ times 100 (22). Values close to the maximum possible value of 100 percent suggest that surrogacy holds.

Parameters

The parameters used in the basic scenario as well as the range of parameters covered in these simulations is shown in table 1. In the basic scenario we assume a prevalence of the exposure of 20 percent ($P_{A=}0.2$), a cumulative incidence of disease of 1 percent ($I_{Y=}0.01$), no association between the exposure and disease ($OR_{AY=}1$), a main study size of 10,000 (N_{MAIN}=10,000), and a 10 percent validation sample ($\%_{VAL=}10$).

In all scenarios of both the first and second set of simulations, both X_1 and X_2 are inversely associated with exposure ($\gamma_1 = -0.405$ and $\gamma_2 = -0.405$, corresponding to an OR_{XA} of 0.67).

In the first set of simulations (equation 5), the associations between the confounders and disease are defined by the association between PS_{GS} and disease. The value for the basic scenario ($\eta_2 = -9$) reflects a reasonable relation between the PS, which is bounded between 0 and 1 and often has low variability, and risk of disease. In the second set of simulations (equation 7), both X_1 and X_2 are risk factors for disease (π_1=0.405 and π_2=0.405, corresponding to an OR_{XY} of 1.5). Therefore, X_1 and X_2 lead to confounding towards lower values of the exposure-disease association (see downward arrow in table 3).

RESULTS

In table 2 we present the results of the first set of simulations, i.e. when simulating the disease as a function of the exposure and PS_{GS} (equation 5) and thus surrogacy holds by design. The following parameters are varied around the value of the basic scenario: the cumulative incidence of disease I_Y, the odds ratio of the exposure-of-interest disease association OR_{AY}, the odds ratio of the unobserved confounder-exposure association OR_{CA}, the log odds ratio of the PS_{GS}-disease association η_2, the size of the main study N_{MAIN}, and the percentage of persons in the random validation sample $\%_{VAL}$. These parameters are varied while keeping all other parameters at the value of the basic scenario typed in italics and presented in table 1. In particular, the true OR_{AY} is 1.0 in all scenarios, except for the two rows with OR_{AY}=2 and OR_{AY}=0.5. For easy comparison of the magnitude and the direction of confounding, we present the median crude OR_{AY} and the median OR_{AY} adjusting for PS_{EP} based on the observed covariates X_1 and X_2 only (equation 2) in the main study.

In all scenarios assessed, median estimates of OR_{AY} from PSC are close to the true value and the percent bias reduction (where applicable) is between 71 and 110 percent, except when N_{MAIN}=1,000. A bias reduction of 100 represents complete control of bias (no residual confounding) and values exceeding 100 represent overcorrection. In some scenarios (marked with ‡), percent bias reduction is either undefined (since the expected value of the estimator controlling for X_1 and X_2 only is unbiased) or not meaningful, since there is little residual bias (and thus the denominator of the percent bias reduction is close to 0). The coverage of the 95 percent confidence interval ranges from 86.0 to 95.9 percent (except when N_{MAIN}=1,000) and is near nominal in many scenarios. Coverage is reduced with increasing incidence of disease (I_Y), odds ratios of the exposure-disease association (OR_{AY}) away from the null, and decreasing size of the validation study ($\%_{VAL}$).

In table 3 we show the results when the disease occurrence is simulated as a function of exposure and the three individual covariates (equation 7) to allow surrogacy to be violated. We present the median and interquartile range of the odds ratios of the exposure disease association OR_{AY}, the median percent bias reduction for selected parameters using PSC, and the two diagnostics for a violation of the surrogacy assumption. Instead of the log odds ratio of the PS_{GS}-disease association η_2 varied in the first set of simulations, OR_{CY} is varied for OR_{CA} of 0.5 and 2.0, respectively.

Since surrogacy can be violated when disease occurrence is simulated as a function of the individual covariates (equation 7), we include the results of the two proposed diagnostics for the surrogacy assumption, i.e. the LRT for PS_{EP} and the percent variance of Y due to PS_{GS} and PS_{EP} explained by PS_{GS}.

When the surrogacy assumption holds, i.e. when the median p-value of the LRT is higher than 0.3 and percent variance explained by PS_{GS} is more than 73.8 percent in the scenarios assessed, the median OR_{AY} is very close to the true value and percent bias reduction ranges from 32 to

106 percent, accordingly (except when N_{MAIN}=1,000). The coverage of the 95 percent confidence interval is close to nominal in these scenarios, with coverage decreasing with increasing incidence of disease (I_Y) and decreasing size of the validation sample ($\%_{VAL}$).

PSC is biased, however, when $OR_{CA=}2$ and $OR_{CY}=2$, when $OR_{CA=}0.5$ and $OR_{CY}=0.5$, when $OR_{CA=}0.5$ and $OR_{CY}=1$, when $OR_{CA=}2$ and $OR_{CY}=1$ (some scenarios are presented twice to allow easy assessment of variation of one of the parameters). These scenarios can be characterized by the additional confounding due to the unobserved covariate C not acting in the same direction as the confounding by the observed covariates **X** (see arrows). They all show indications for violation of the surrogacy assumption: the median p-value of LRT is 0.2 or less and the percent variance explained by PS_{GS} is low (less than 45.5 percent in the scenarios assessed).

The only apparent exception seems to be the scenario with $OR_{CA}=1$, where the LRT (0.05) and the percent variance explained by PS_{GS} (43.6 percent) indicate a violation of the surrogacy assumption but PSC is nevertheless unbiased. Since there is by definition no residual confounding when $OR_{CA}=1$, the analysis controlling for X_1 and X_2 leads to an unbiased estimate. When $OR_{CA}=1$, PSC is unbiased despite indications for violation of surrogacy since C is not associated with exposure and therefore not a confounder. When $OR_{CA}=1$, surrogacy is violated since the inclusion of C adds unnecessary variability to PS_{GS} compared with PS_{EP} (23). Therefore, surrogacy is a sufficient but not always necessary condition for PSC to be valid.

DISCUSSION

We evaluated the performance of propensity score calibration using simulations over a wide range of parameter-values. These results should be interpreted in light of the specific parameter-values we selected for our settings. These values resulted in strong, but not unrealistic, unmeasured confounding in the main study (e.g. such as might be plausible for the association between self-selected hormone therapy and myocardial infarction in postmenopausal women). PSC was always valid in the first set of simulations (table 2), i.e. when simulating the disease as a (log)linear function of PS_{GS} according to the target model of PSC and thus surrogacy holds by design. The second set of simulations, however, indicates that the approach may increase rather than decrease bias if surrogacy is violated (table 3). Generally speaking, surrogacy is violated when the direction of confounding of the exposure-disease association caused by the unobserved variable(s) differs from that of the confounding due to observed variables. One can use different diagnostics to assess violations of surrogacy if the validation study contains sufficient information on the outcome.

Despite the intuition that adding an unmeasured confounder to the PS would always introduce differential measurement error and thus violate surrogacy, surrogacy holds when the direction of confounding of the observed and unobserved variables(s) is the same, as evinced by our simulations. In such settings, adding the confounder to the PS increases the strength of association between the PS and the disease outcome. All of the association between PS_{EP} and the outcome might therefore be captured in PS_{GS}, which results in surrogacy. If the direction of the confounding by the unmeasured confounder(s) is different from the direction of the measured confounder(s), including the unmeasured confounder in the PS reduces the strength of association between the PS and disease outcome. Therefore, PS_{EP} is more strongly associated with disease risk than PS_{GS}, thus violating surrogacy.

Simulations cannot allow a quantitative assessment of how frequently surrogacy holds or is violated in epidemiologic studies. Many informal and other formal sensitivity analyses of residual confounding also depend on the assumption of uni-directionality of confounding (e.g. 24,25,26). This assumption is plausible in many but not all epidemiologic settings, especially

since PSC addresses the joint confounding of a set of observed and unobserved covariates rather than a single covariate. In such a setting, the surrogacy assumption might be plausible if an underlying and well understood framework for confounding is consistent with surrogacy. Practical examples for such a framework include variables used in claims data (e.g. chronic obstructive pulmonary disease, being admitted to a nursing home) as a proxy for the unmeasured covariate of interest (e.g. smoking and frailty, respectively). In such settings, a more refined PS, based on alternative measures with less error compared with those measured in the main study (e.g. smoking and activities of daily living or cognitive function, respectively), might be hypothesized to contain all the relevant information on propensity of exposure captured in an error-prone PS. Thus surrogacy might be a plausible assumption in such settings.

The direction of confounding introduced by any single unobserved covariate may be unpredictable and thus clearly lead to a violation of surrogacy of PS_{EP} estimated without information on that covariate. Prior knowledge about the association of that covariate with the study outcome might be used as a warning sign in case outcome information is not available in the validation study. As in regression analyses, inclusion of a covariate unrelated to disease should be avoided. In PS analyses, not including such a covariate would lead to an increase in efficiency (23). Because PSC is used to adjust for unmeasured confounding, including covariates form a validation study thought to be unrelated to disease ($OR_{CD}=1$) would make no sense. Including only covariates from the validation study that truly are risk factors for the disease outcome would avoid biased results due to violation of surrogacy in two out of the four settings assessed where PSC was biased.

The assumption necessary for PSC to be valid can further be conceptualized in the framework of instrumental variables (27). One critical assumption of instrumental variable analysis is that the instrument is unrelated to the outcome given the exposure of interest (28). Similarly, PSC is valid if PS_{EP} is independent of disease given exposure of interest and PS_{GS}.

If the validation study contains data on disease outcome, fitting a model of the outcome as a function of these two propensity scores and exposure in the (internal) validation study allows one to test surrogacy before applying PSC. The proposed tests for surrogacy performed well in the scenarios assessed. The cut-points we used were chosen according to the scenarios assessed, however, and are arbitrary. The power of the likelihood ratio test to detect violations of surrogacy will depend on the size of the validation study. The percentage of variance in outcome explained by PS_{GS} might therefore be preferred in validation studies with few outcomes.

PSC based on regression calibration as proposed by Rosner et al. (15,16) had a tendency to ‘over-adjust’ even in scenarios where surrogacy was met. Furthermore, standard errors of the adjusted estimate obtained from regression calibration were consistently smaller than the empirical variance of the adjusted estimates across all simulations leading to non-nominal coverage of the confidence intervals (data not shown). We therefore implemented PSC as a single imputation according to Carroll et al. (17), which allowed matching on the imputed PS_{GS}. Matching on the imputed PS_{GS} and using bootstrapping to obtain a robust estimate of the variance solved both problems, resulting in nominal coverage probabilities for most scenarios assessed.

Coverage probabilities decreased with increasing incidence of the disease outcome and decreasing size of the validation sample. A rare disease outcome is a general assumption of regression calibration (16) and is likely to be exacerbated in PSC owing to the problem of the non-collapsibility of the odds ratio under exchangeability of exposed and unexposed given PS_{GS} (29). The lower coverage probabilities with smaller sizes of the validation study might

be an indication of problems due to model misspecification or non-convergence. Coverage probabilities are meaningless for biased estimators, since they would approach 0 with increasing study size. Despite this, we present coverage probabilities for all scenarios because scenarios with only small residual bias are likely to converge to unbiased ones with increasing study size and number of simulations.

Regression calibration approximations are known to fail when the measurement error is large (30,31), as when the correlation between the estimated error-prone and gold standard measurements is weak. Kuha observed that the performance of regression calibration degrades if the product of the squared estimate and its mean squared error exceeds 0.5 (30). In our scenarios, the median of this value ranged from 0.48 to 0.66, corresponding to a range where problems can be expected in a large proportion of simulated datasets. Since PS_{GS} captures all of the confounding in a single covariate, misspecification of its association with the outcome is likely to reduce its ability to control for confounding (32). The linear measurement error model of regression calibration is only one possible model to relate PS_{GS} to PS_{EP}.

Besides surrogacy, the validity of PSC is dependent on additional assumptions underlying all epidemiologic analyses. Even with validation data, it is unlikely that all confounders are measured with sufficient accuracy and therefore unmeasured confounding can never be completely ruled out. Residual or unmeasured confounding is only one aspect of uncertainty in epidemiologic studies (33) and considering multiple forms of biases is worthwhile (34,35).

We conclude that propensity score calibration to adjust for unmeasured confounding with validation data is a useful approach to reduce residual bias when the error-prone propensity score estimated in the main study is a surrogate for the true gold-standard propensity score. Like any method addressing unmeasured covariates or missing data, PSC is not a substitute for having all covariates adequately measured. If surrogacy is violated, PSC might increase rather than decrease bias. In the usual setting of validation studies without information on outcomes, PSC is likely but not guaranteed to improve estimates if an underlying theory about the confounding pattern is consistent with the necessary assumption. Adding measures of disease outcome to validation studies would allow epidemiologists to expand the application of PSC without relying on a strong surrogacy assumption.

ACKNOWLEDGEMENTS

This project was funded by a grant (RO1 023178) from the National Institute on Aging

REFERENCES

1. Cornfield J, Haenszel W, Hammond EC, et al. Smoking and lung cancer: Recent evidence and a discussion of some questions. J Natl Cancer Inst 1959;22:173–203. [PubMed: 13621204]
2. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. J R Stat Soc Ser B 1983;45:212–8.
3. Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika 1987;74:13–26.
4. Rosenbaum PR. Sensitivity analysis for matched case-control studies. Biometrics 1991;47:87–100. [PubMed: 2049516]
5. Greenland S. Basic methods for sensitivity analysis of bias. Int J Epidemiol 1996;25:1107–16. [PubMed: 9027513]
6. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 1998;54:948–63. [PubMed: 9750244]
7. Robins, JM.; Rotnitzky, A.; Scharfstein, DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran, ME.; Bery, DA., editors. Statistical Models in Epidemiology. Springer; New York: 1999.

8. Little, RJA.; Rubin, DA. Statistical Analysis with Missing Data. 2nd ed.. Wiley; New York: 2002.
9. Rosenbaum, P. Observational Studies. 2nd ed.. Springer; New York: 2002.
10. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. Annu Rev Publ Health 2000;21:121–45.
11. Schneeweiss S, Glynn RJ, Tsai EH, et al. Adjusting for unmeasured confounders in pharmacoepidemiologic claims data using external information: The example of COX2 inhibitions and myocardial infarction. Epidemiology 2005;16:17–24. [PubMed: 15613941]
12. MacLehose RF, Kaufman S, Kaufman JS, et al. Bounding causal effects under uncontrolled confounding using counterfactuals. Epidemiology 2005;16:548–55. [PubMed: 15951674]
13. Stürmer T, Schneeweiss S, Avorn J, et al. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am J Epidemiol 2005;162:279–89. [PubMed: 15987725]
14. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
15. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989;8:1051–69. [PubMed: 2799131]
16. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. Am J Epidemiol 1990;132:734–45. [PubMed: 2403114]
17. Carroll, RJ.; Ruppert, D.; Stefanski, LA. Measurement error in nonlinear models. Chapman/Hall; London: 1995.
18. Carroll RJ, Stefanski LA. Approximate quasi-likelihood estimation in models with surrogate predictors. J Am Stat Assoc 1990;85:652–63.
19. Stürmer T, Schneeweiss S, Brookhart MA, et al. Analytic Strategies to adjust confounding using Exposure Propensity Scores and Disease Risk Scores: Nonsteroidal Antiinflammatory Drugs and Short-Term Mortality in the Elderly. Am J Epidemiol 2005;161:891–8. [PubMed: 15840622]
20. Parsons, LS. Reducing bias in a propensity score matched-pair sample using greedy matching techniques. 2001. http://www2.sas.com/proceedings/sugi26/p214-26.pdf
21. Stürmer T, Joshi M, Glynn RJ, et al. A review of applications of propensity score methods showed increased use but infrequently different estimates compared with other methods. J Clin Epidemiol 2006;59:437–447. [PubMed: 16632131]
22. McFadden D. The measurement of urban travel demand. J Public Econom 1974;3:303–28.
23. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection in propensity score models: some insights from a simulation study. Am J Epidemiol 2006;163:1149–56. [PubMed: 16624967]
24. Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. J Am Stat Assoc 1994;89:1314–28.
25. Fung KY, Krewski D. Evaluation of regression calibration and SIMEX methods in logistic regression when one of the predictors is subject to additive measurement error. J Epidemiol Biostat 1999;4:65–74. [PubMed: 10619053]
26. Rothman KJ, Wentworth CE 3rd. Mortality of cystic fibrosis patients treated with tobramycin solution for inhalation. Epidemiology 2003;14:55–9. [PubMed: 12500046]
27. Greenland S. An introduction to instrumental variables for epidemiologists. Int J Epidemiol 2000;29:722–9. [PubMed: 10922351]
28. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc 1996;91:444–55.
29. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 1984;71:431–44.
30. Kuha J. Corrections for exposure measurement error in logistic regression models with an application to nutritional data. Stat Med 1994;13:1135–48. [PubMed: 8091040]
31. Stürmer T, Thürigen D, Spiegelman D, et al. The performance of methods for correcting measurement error in case-control studies. Epidemiology 2002;13:507–16. [PubMed: 12192219]

32. Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics 1973;29:185–203.

33. Maclure M, Schneeweiss S. Causation of bias: the episcope. Epidemiology 2001;12:114–22. [PubMed: 11138805]

34. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational epidemiologic data. Epidemiology 2003;14:451–8. [PubMed: 12843771]

35. Greenland S. Multiple-bias modeling for analysis of observational data. J R Stat Soc Ser A 2005;168:267–306.

Table 1

Parameters and their values in the basic and the alternative scenarios

Parameter				Parameter values	
Notation	**Simulation 1 (table 2)**	**Simulation 2 (table 3)**	**Meaning**	**Basic scenario**	**Alternative scenarios**
P_A			Prevalence of exposure of interest	0.20	-
I_Y			Cumulative incidence of disease	0.01	0.002, 0.01, 0.05, 0.10
OR_{AY}	$\exp(\eta_1)$ (eq. 5)	$\exp(\theta_1)$ (eq. 7)	Odds ratio of exposure - disease association	1.0	0.5, 1.0, 2.0
OR_{CA}	$\exp(\gamma_3)$ (eq. 3)	$\exp(\gamma_3)$ (eq. 3)	Odds ratio of confounder - exposure association (independent of X_1,X_2)	0.5	0.5, 1.0, 2.0
η_2	η_2 (eq. 5)	-	Log odds ratio of gold-standard PS – disease association (table 2)	− 9.0	−9.0, −5.0, −1.0, 1.0, 5.0, 9.0
OR_{CY}	-	$\text{Exp}(\theta_4)$ (eq. 7)	Odds ratio of confounder - disease association (table 3)	2.0	0.5, 1.0, 2.0
N_{MAIN}			Size of main study	10,000	1,000, 5,000, 10,000
$\%_{VAL}$			Size of validation study (in percent of main study)	10	2, 5, 10, 20, 50

Table 2

Odds ratio, percent bias reduction and coverage probability of 95 percent confidence interval of exposure-of-interest disease association OR_{AY} using propensity score calibration; median (interquartile range) from 1,000 datasets for each scenario (row) with disease simulated as a function of true propensity score and exposure; true OR_{AY}=1 in all scenarios except when OR_{AY} is the parameter varied

		Median OR_{AY}		PSC				
				OR_{AY}				
Parameter varied*	Value	Crude	Adjusted for X_1,X_2	Median	25th	75th	Bias reduction†	Coverage‡
I_Y	0.01	0.50	0.65	1.08	0.78	1.50	100	94.9
	0.05	0.49	0.63	1.04	0.89	1.22	109	89.9
	0.10	0.50	0.63	1.05	0.92	1.18	110	86.0
OR_{AY}	2.0	0.99	1.28	2.12	1.62	2.88	110	88.3
	1.0	0.50	0.65	1.08	0.78	1.50	100	94.9
	0.5	0.24	0.31	0.54	0.35	0.73	85	91.7
OR_{CA}	0.5	0.50	0.65	1.08	0.78	1.50	100	94.9
	1.0	0.69	1.00	1.00	0.77	1.30	–§	94.8
	2.0	1.12	1.08	1.00	0.81	1.25	–§	95.9
η_2	**−*9***	0.50	0.65	1.08	0.78	1.50	100	94.9
	−*5*	0.63	0.74	1.00	0.79	1.36	73	95.8
	−1	0.90	0.93	1.00	0.78	1.29	–§	95.4
	+1	1.12	1.08	1.00	0.79	1.26	–§	95.3
	+5	1.89	1.51	1.04	0.85	1.27	84	93.3
	+*9*	2.97	2.06	1.14	0.91	1.39	83	89.1
N_{MAIN}	10,000	0.50	0.65	1.08	0.78	1.50	100	94.9
	5,000	0.50	0.64	1.00	0.71	1.75	71	93.2
	1,000	0.46	0.59	1.00	0.50	3.46	30	98.0
$\%_{VAL}$	2	0.50	0.65	1.10	0.77	1.67	100	92.4
	5	0.50	0.65	1.08	0.77	1.50	100	93.6
	10	0.50	0.65	1.08	0.78	1.50	100	94.9
	20	0.50	0.65	1.08	0.78	1.43	100	95.4
	50	0.50	0.65	1.00	0.77	1.40	100	94.8

*For the definition and range of these parameters refer to table 1 (I_Y=0.002 not presented due to non-convergence); only one parameter (bold typeface) is varied at a time, whereas all other parameters are kept constant at the level of the basic scenario (italic typeface) presented in table 1; note that true OR_{AY}=1 in all scenarios except when OR_{AY} is the parameter varied

†Median percent bias reduction according to Cochran 1968: 100 χ {1 − (log(OR) PSC − η1[eq. 5])/(ß1[eq. 2] − η1[eq. 5])}; 0 percent equals no improvement over analysis controlling for error-prone PS, 100 percent equals no residual bias (truth)

‡Coverage probability of empirical 95 percent confidence interval obtained from 1,000 bootstrap samples

§Undefined or instable estimate, since expected value of denominator is either 0 (no bias controlling for X_1 and X_2) or close to 0

Table 3

Direction of confounding by observed (X_1, X_2) and unobserved (C) covariates, conformation with surrogacy assumption, and odds ratio, percent bias reduction, and coverage probability of 95 percent confidence interval of exposure-of-interest disease association OR_{AY} using propensity score calibration; median (interquartile range) from 1,000 datasets for each scenario (row) with disease simulated as a function of individual covariates X_1, X_2, C, and exposure; true OR_{AY}=1 in all scenarios except when OR_{AY} is the parameter varied

		Direction of confounding		Median OR_{AY}		Surrogacy		PSC				
								OR_{AY}				
Parameter varied*	Value	X_1,X_2	C	Crude	Adjusted for X_1,X_2	LRT†	%‡	Median	25th	75th	Bias reduction§	Coverage¶
	0.002	↓	↓	0.49	0.65	0.5	77.3	1.00	0.50	2.00	32	94.7
	0.01	↓	↓	0.49	0.65	0.5	90.9	1.00	0.75	1.37	86	96.5
I_Y	0.05	↓	↓	0.49	0.64	0.4	97.8	1.02	0.89	1.20	105	92.4
	0.10	↓	↓	0.50	0.65	0.4	98.5	1.02	0.92	1.14	106	89.9
	2.0	↓	↓	0.97	1.27	0.5	91.4	2.00	1.54	2.62	100	92.3
OR_{AY}	1.0	↓	↓	0.49	0.65	0.5	90.9	1.00	0.75	1.37	86	96.5
	0.5	↓	↓	0.25	0.32	0.5	90.8	0.50	0.33	0.70	74	92.5
	0.5	↓	↓	0.49	0.65	0.5	90.9	1.00	0.75	1.37	86	96.5
OR_{CA}	1.0	↓	→	0.74	1.00	0.05	43.6	1.00	0.77	1.25	–#	96.1
	2.0	↓	↑	1.13	1.54	0.01	8.6	2.09	1.60	2.78	−75	na
	0.5	↓	↑	1.14	1.54	0.01	8.2	2.12	1.67	2.80	−76	na
OR_{CY}	1.0	↓	→	0.75	1.00	0.05	39.7	1.50	1.11	2.00	–#	na
	2.0	↓	↓	0.49	0.65	0.5	90.9	1.00	0.75	1.37	86	96.5
	0.5	↓	↓	0.49	0.64	0.3	92.7	1.00	0.75	1.37	100	95.0
OR_{CY}	1.0	↓	→	0.75	1.00	0.2	45.5	1.50	1.12	2.00	–#	na
(OR_{CA}=2.0)	2.0	↓	↑	1.13	1.54	0.01	8.6	2.09	1.60	2.78	−75	na
	10,000	↓	↓	0.49	0.65	0.5	90.9	1.00	0.75	1.37	86	96.5
N_{MAIN}	5,000	↓	↓	0.47	0.62	0.5**	86.1**	1.00	0.67	1.60	100	95.5
	1,000	↓	↓	0.47	0.61	0.5**	67.3**	1.00	0.50	3.00	19	98.7
	2	↓	↓	0.49	0.64	0.5	73.8	1.03	0.75	1.50	100	92.9
	5	↓	↓	0.49	0.64	0.5	86.0	1.07	0.75	1.45	100	94.5
$\%_{VAL}$	10	↓	↓	0.49	0.65	0.5	90.9	1.00	0.75	1.37	86	96.5
	20	↓	↓	0.49	0.64	0.5	95.8	1.00	0.73	1.36	85	96.5
	50	↓	↓	0.49	0.64	0.4	98.2	1.00	0.73	1.27	85	95.4

*For the definition and range of these parameters refer to table 1; only one parameter (bold typeface) is varied at a time, whereas all other parameters are kept constant at the level of the basic scenario (italic typeface) presented in table 1 (with the exception of the second series of rows for OR_{CD}, where OR_{CE} is 2.0 instead of 0.5) ; note that true OR_{AY}=1 in all scenarios except when OR_{AY} is the parameter varied

†Median p-value for surrogacy assumption from likelihood-ratio-test comparing logistic regression model $\text{logit}(Y) = \nu_0 + \nu_1 A + \nu_2 PS_{GS} + \nu_3 PS_{EP}$ with logistic regression model $\text{logit}(Y) = \nu'_0 + \nu'_1 A + \nu'_2 PS_{GS}$, i.e. testing for the independent contribution of PS_{EP}

‡Percent of variance in Y explained by PS_{GS} and PS_{EP} which is due to PS_{GS}; calculated as ratio of likelihood-ratio comparing logistic regression model $\text{logit}(Y) = \nu'_0 + \nu'_1 A + \nu'_2 PS_{GS}$ with the nested logistic regression model $\text{logit}(Y) = \nu''_0 + \nu''_1 A$ and of the likelihood-ratio comparing the logistic regression model $\text{logit}(Y) = \nu_0 + \nu_1 A + \nu_2 PS_{GS} + \nu_3 PS_{EP}$ with the nested logistic regression model $\text{logit}(Y) = \nu''_0 + \nu''_1 A$ times 100; values close to the maximum possible value (100 percent) indicate that surrogacy holds

[§] Median percent bias reduction according to Cochran 1968: $100 \times \{1 - (\log(OR)\ PSC - \theta_1[\text{eq. 7}])/(ß_1[\text{eq. 2}] - \theta_1[\text{eq. 7}])\}$; 0 percent equals no improvement over analysis controlling for error-prone PS, 100 percent equals no residual bias (truth)

[¶] Coverage probability of empirical 95 percent confidence interval obtained from 1,000 bootstrap samples; na: not applicable since PSC estimator is biased due to violation of surrogacy

[#] Undefined since expected value of denominator is 0 (no bias controlling for X_1 and X_2)

[**] Based on 98.4% and 65.6% of all simulated studies for studies with N=5,000 and N=1,000, respectively, due to non-convergence of outcome models in validation study