American Journal of Epidemiology

Vol. 175, No. 12
DOI: 10.1093/aje/kwr463
Advance Access publication:
April 24, 2012

Practice of Epidemiology

Propensity Score Calibration in the Absence of Surrogacy

Mark Lunt, Robert J. Glynn, Kenneth J. Rothman, Jerry Avorn, and Til Stürmer*

* Correspondence to Dr. Til Stürmer, Department of Epidemiology, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill, McGavran-Greenberg Hall, CB 7435, Chapel Hill, NC 27599-7435 (e-mail: til.sturmer@post.harvard.edu).

Initially submitted February 10, 2011; accepted for publication November 14, 2011.

Propensity score calibration (PSC) can be used to adjust for unmeasured confounders using a cross-sectional validation study that lacks information on the disease outcome (Y), under a strong surrogacy assumption. Using directed acyclic graphs and path analysis, the authors developed a formula to predict the presence and magnitude of the bias of PSC in the simplest setting of a binary exposure (T) and 1 confounder (X) that are observed in the main study and 1 confounder (C) that is observed in the validation study only. PSC bias is predicted on the basis of parameters that can be estimated from the data and a single unidentifiable parameter, the relative risk (RR) associated with C (RR_{CY}). The authors simulated 1,000 cohort studies each with a Poisson-distributed outcome Y, varying parameter values over a wide range. When using the true parameter for RR_{CY}, the formula predicts PSC bias almost perfectly in this simple setting (correlation with observed bias over 24 scenarios assessed: $r = 0.998$). The authors conclude that the bias from PSC observed in certain scenarios can be estimated from the imbalance in C between treated and untreated persons, after adjustment for X, in the validation study and assuming a range of plausible values for the unidentifiable RR_{CY}.

bias (epidemiology); confounding factors (epidemiology); epidemiologic methods; path analysis; propensity score; propensity score calibration; research design

Abbreviations: CI, confidence interval; DAG, directed acyclic graph; EP, error-prone; GS, gold standard; IR, incidence rate; IRR, incidence rate ratio; NSAID, nonsteroidal antiinflammatory drug; OR, odds ratio; PSC, propensity score calibration.

Uncontrolled confounding can be a major source of bias in nonexperimental research. Investigators often lack measures of important potential confounders, such as smoking and body mass index in pharmacoepidemiologic studies that use claims data and laboratory values or blood pressure measurements in questionnaire-based studies. If individual data on confounders not measured in all study participants are available for a subsample or a validation study, these data can be used to adjust for joint confounding by the variables that are unobserved in the main study under reasonable assumptions about the selection of participants into the validation study (1). If the validation study contains information on the disease outcome of interest, the adjustment can be achieved using methods for missing data, such as multiple imputation, to adjust for the covariates that are missing in the main study (2, 3). If the validation study does not contain information on the disease outcome of interest, which is often the case, unmodified imputation techniques fail, because they assume availability of complete cases for imputation and/or estimation (4). For this specific setting, we have proposed propensity score calibration (PSC) (5, 6). PSC combines propensity scores to adjust for confounding in the main study (7) with methods for measurement error correction, specifically regression calibration (8, 9), to adjust for the error in the propensity score due to unmeasured covariates estimated in the cross-sectional validation study.

So far, we have established that the validity of PSC is dependent on a key assumption of regression calibration, called surrogacy (10). For PSC, this assumption means that the error-prone propensity score estimated in the main study is independent of the disease outcome given the gold standard propensity score estimated in the validation study. This assumption cannot be assessed in the cross-sectional validation study. So far, we have shown that surrogacy is violated

when the direction of the confounding by the variable(s) observed only in the validation study differs from the direction of confounding of the variable(s) observed in the main study. We have argued that the assumption of "unidirectionality" of confounding, also assumed in other methods used to adjust for uncontrolled confounding, might hold more often than not (6).

Using directed acyclic graphs (DAGs), Greenland has pointed out that surrogacy may not be as natural or credible for the propensity score as it is for measurement error (see the Acknowledgments). Here we present a framework for PSC using DAGs and develop a formula for predicting the presence and magnitude of the bias of PSC based on path analysis using regression coefficients (11). We evaluated the performance of this bias formula in simulations that address scenarios in which we have shown PSC to be unbiased and scenarios in which we have shown PSC to be biased, as well as additional scenarios not assessed so far. We also implemented it in an empirical example.

MATERIALS AND METHODS

DAGs and structural equations

A DAG consists of a set of points and a set of arcs joining the points, each arc being marked with an arrow to show its direction (12–14). A path is an ordered set of points on the graph such that 1) there is an arc between each point and the next (which may be forwards or backwards) and 2) no point appears in the path more than once.

DAGs can be used to display causality assumptions. For example, consider the DAG in Figure 1 (the labels on the arcs will be explained below). This graph tells us that the variables T and X both affect Y, that variable X affects T, that T has no effect on X, and that Y has no effect on T or X.

Traditionally, error terms (random variables with mean 0 that are uncorrelated with all other variables and error terms in the DAG) are not shown but are assumed to exist at each node. Hence, Figure 1 tells us that

$$Y = f_y(X, T) + \varepsilon_Y$$

$$T = f_T(X) + \varepsilon_T$$

but makes no assumptions about the form of the functions f_Y and f_T. If we assume that these functions are both linear, the DAG represents the structural equation model

$$X = \varepsilon_X$$

$$T = \lambda_3 X + \varepsilon_T$$

$$Y = \lambda_1 X + \lambda_2 T + \varepsilon_Y.$$

The parameters in the above model are causal: That is, if we change X by ΔX while holding T fixed, we would expect to see Y change by $\lambda_1 \Delta X$.

Given a set of data that we assume to be generated by a given structural equation model, Wright (11) gave a method

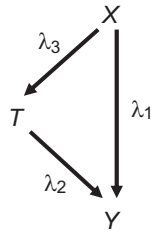

Figure 1. Directed acyclic graph for the effect of a treatment T on an outcome Y, with the variable X acting as a confounder.

of labeling each arc with a value, such that the observed correlation between any 2 variables is given by the sum of the values of all paths between the variables, where the value of a path is the product of the values of each arc in the path. The value of an arc $X \rightarrow Y$ is given by the correlation coefficient for the correlation between X and Y, after conditioning on a set of variables sufficient to block all other paths between X and Y.

In general, a path is blocked (12–14) if either 1) the path contains a collider (a node on the path at which 2 arrows meet), and we do not condition on the collider or any of its descendants, or 2) the path contains a chain $i \rightarrow z \rightarrow j$ or a fork $i \leftarrow z \rightarrow j$, and we condition on the variable z. Thus, in Figure 1, the path $X \rightarrow Y \leftarrow T$ is blocked by the variable Y, a collider, and Wright would assign the value ρ_{XT} to the path $X \rightarrow T$. To calculate the value of the arc $T \rightarrow Y$, however, we would need to block the path $T \leftarrow X \rightarrow Y$ by conditioning on X. The value of this arc would then be $\rho_{TY|X}$, where this notation stands for the correlation between T and Y conditional on X.

If we consider the DAG as representing a structural equation model, it is preferable to work with regression coefficients rather than correlation coefficients: What change in Y would we expect to see if X changes by ΔX? A similar approach can be used to calculate the regression coefficients for all predictors in a regression equation: The coefficient for a given predictor is the sum of the values of all paths between the predictor and the outcome which are not blocked by other predictors in the regression model, and the value of a path is the product of all of the arcs in that path. The regression coefficient for the regression of X on Y is not the same as that for the regression of Y on X, however. The value of a path going against the arrow is different from the value of a path going with the arrow. We can use the parameters of the structural equation model and Wright's method only if we are moving in the direction of the arrows. For example, we can obtain the observed conditional expectation of Y given X as

$$\begin{aligned} E[Y|X] &= \lambda_1 X + \lambda_2 E[T|X] + E[\varepsilon_Y] \\ &= \lambda_1 X + \lambda_2(\lambda_3 X + E[\varepsilon_T|X]) \\ &= (\lambda_1 + \lambda_2\lambda_3)X. \end{aligned}$$

This will not work for general paths, however. For example, consider the expectation of Y given T. There are 2 paths from T to Y: $T \rightarrow Y$ and $T \leftarrow X \rightarrow Y$. We need to calculate

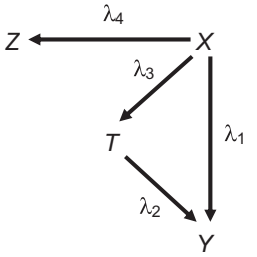

Figure 2. Directed acyclic graph for the effect of a treatment *T* on an outcome *Y*, in which the confounder *X* is measured with error.

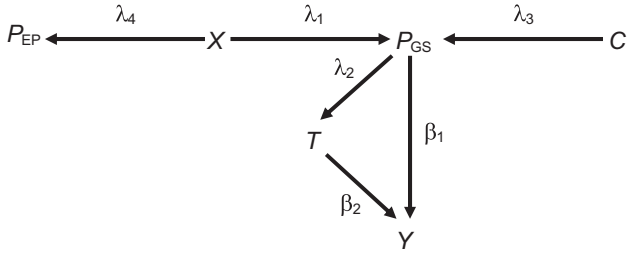

Figure 3. Directed acyclic graph illustrating the use of propensity score calibration to estimate the effect of treatment *T* on outcome *Y*, assuming surrogacy. (EP, error-prone; GS, gold standard).

the value of the arc $T \leftarrow X$, whereas what we know is the value of the arc $X \rightarrow T$. However, the value of a path pointing in the reverse direction can be calculated from the value of a path pointing in the forward direction and the variances of the variables that lie along the path. The details are given in Web Appendix 1, which is posted on the *Journal*'s website (http://aje.oxfordjournals.org/). If the value of the forward path is λ, then we will write the value of the reverse path as $r(\lambda)$.

Therefore, we can calculate the regression coefficient for the regression of *Y* on *T* as $\lambda_2 + r(\lambda_3)\lambda_1$. The causal effect of *T* on *Y* is λ_2, and the difference in *Y* due to confounding by *X* is $r(\lambda_3)\lambda_1$, with $r(\lambda_3)$ being the difference in *X* required to cause a difference of 1 in *T* and with λ_1 being the effect of a unit difference in *X* on *Y*. If we include *X* in the regression equation, this blocks the path $T \leftarrow X \rightarrow Y$, so the regression coefficient for *T* would then be λ_2, the true causal effect.

Regression calibration

Now suppose that we are unable to measure the confounder *X* perfectly, and our measurement is $Z = f(X) + \varepsilon_z$, where ε_z is a random disturbance from some distribution. Further, suppose that *Z* is a linear function of *X*, that is, $Z = \lambda_4 X + \varepsilon_z$. This situation is shown in Figure 2. The corresponding structural equation model is

$$X = \varepsilon_X$$

$$Z = \lambda_4 X + \varepsilon_Z$$

$$T = \lambda_3 X + \varepsilon_T$$

$$Y = \lambda_1 X + \lambda_2 T + \varepsilon_Y.$$

We are now unable to control for *X*, since it is not observed. We can adjust for *Z*, but this does not block the path $T \leftarrow X \rightarrow Y$. Nonetheless, adjusting for *Z* may improve our estimate of the treatment effect, depending on the parameters of the structural equation model. In the extremes, if $\lambda_4 = 0$, adjusting for *Z* will not affect the estimate, which will be $\lambda_2 + r(\lambda_3)\lambda_1$, while if $\varepsilon_z = 0$, adjusting for *Z* will remove the confounding completely.

If we write λ_{XY}^{Z} for the value of the arc $X \rightarrow Y$ when controlling for a set of variables *Z* and if we write λ_{XY} for the true causal effect of the arc, then $\lambda_{XY}^{Z} = \lambda_{XY}$ if and only if *Z* blocks all paths between *X* and *Y* except the direct path $X \rightarrow Y$. Then the estimated treatment effect after adjusting for *Z* is $\lambda_2 + r(\lambda_3^Z)\lambda_1$, and the regression equation for *Y* in terms of *Z* and *T* is

$$E[Y \mid Z, T] = (r(\lambda_4^T)\lambda_1)Z + (\lambda_2 + r(\lambda_3^Z)\lambda_1)T. \quad (1)$$

The bias in the coefficient for *T* is $r(\lambda_3^Z)\lambda_1$, which we cannot estimate without measuring *X*. If we perform a validation study in which we measure *X*, *Z*, and *T*, we can obtain estimates of $r(\lambda_3^Z)$ and λ_1 by fitting the regression of *X* on *Z* and *T*. The regression equation we obtain will be

$$E[X \mid Z, T] = r(\lambda_3^Z)T + r(\lambda_4^T)Z. \quad (2)$$

The coefficient of *T* gives an estimate of $r(\lambda_3^Z)$, and dividing the coefficient of *Z* in equation 1 by the coefficient of *Z* in equation 2 gives λ_1. We can therefore calculate $r(\lambda_3^Z)\lambda_1$, and subtracting this from the coefficient of *T* in equation 1 gives the true treatment effect, λ_2.

Regression calibration can also be used with a generalized linear model for *Y*, provided that the linear predictor of *Y* follows a normal distribution. For example, if *Y* were a count and Poisson regression were used to model it, the linear predictor would become

$$E[\log(Y) \mid Z, T] = (r(\lambda_4^T)\lambda_1)Z + (\lambda_2 + r(\lambda_3^Z)\lambda_1)T, \quad (3)$$

which is still a linear function of *Z* and *T*. To obtain the expected count, equation 3 can be exponentiated:

$$E[Y \mid Z, T] = \exp\{E[\log(Y) \mid Z, T]\}.$$

In what follows, *Y* will generally be a Poisson outcome variable, and we will use a generalized linear model to model it. For clarity, we will use λ for all paths modeled with linear regression and β for all paths modeled with another generalized linear model.

Propensity score calibration

At first sight, an error-prone propensity score model looks very similar to the above model. We have a group of variables *X*

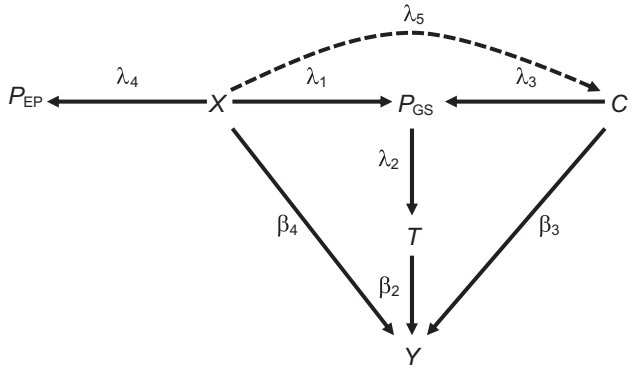

Figure 4. Directed acyclic graph illustrating the use of propensity score calibration to estimate the effect of treatment *T* on outcome *Y*, without assuming surrogacy. (EP, error-prone; GS, gold standard).

which are used to define the error-prone (EP) propensity score and a further group *C* measured in the validation sample used to define the gold standard (GS) propensity score. If we assume that *X* and *C* are uncorrelated and that neither affects *Y* directly, only through P_{GS}, we get Figure 3.

Then when we fit a Poisson regression model for *Y* in terms of P_{EP} and *T*, we again have 2 paths from *T* to *Y*: directly and via P_{GS}. Hence, the regression equations are

$$E[Y \mid P_{EP}, T] = \exp\{\alpha_{Y|P_{EP},T} + r(\lambda_4^T)\lambda_1^{T,P_{EP}}\beta_1 P_{EP} + (\beta_2 + r(\lambda_2^{P_{EP}})\beta_1)T\}. \quad (4)$$

$$E[P_{GS} \mid P_{EP}, T] = \alpha_{P_{GS}|P_{EP},T} + r(\lambda_4^T)\lambda_1^{T,P_{EP}} P_{EP} + r(\lambda_2^{P_{EP}})T. \quad (5)$$

Thus, we can again find $r(\lambda_2^{P_{EP}})$ and $r(\lambda_4^T)\lambda_1$ from a linear regression of P_{GS} on P_{EP} and *T*, and then subtract $r(\lambda_2^{P_{EP}})/(r(\lambda_4^T)\lambda_1^T)$ times the coefficient of P_{EP} from the coefficient of *T* to obtain an unbiased estimate of β_2. Bootstrapping can be used to obtain a confidence interval for β_2.

Surrogacy

Although Figure 3 looks very similar to Figure 2, it is less plausible as a causal model. This says that neither *X* nor *C* affects *Y* directly; they only affect it through the propensity score. For example, if we were concerned about age as a potential confounder, with mortality as the outcome, Figure 3 claims that age can only affect mortality through its influence on treatment, not directly. Note that the absence of these direct effects is not a condition for confounding control using propensity scores. A more plausible causal model that allows for direct effects of *X* and *C* on *Y* is shown in Figure 4.

Note the dotted arrow in Figure 4, labeled λ_5. This is required because P_{GS} is a collider, and if we adjust for P_{GS} (or one of its descendants, such as *T*), we will induce a correlation between *X* and *C*, although *X* and *C* are not correlated marginally. Therefore, the value of λ_5 is 0, but λ_5^T is nonzero.

If we assume this model, then P_{GS} does not act directly on *Y*, but *X* and *C* do. Suppose that

$$E[Y \mid X, C, T] = \exp\{\alpha_{XCT} + \beta_4 X + \beta_2 T + \beta_3 C\}. \quad (6)$$

We cannot estimate these parameters, since we do not have information about *Y* and *C* on the same subjects. If we fit a model for *Y* containing P_{EP}, our regression equations become

$$E[Y \mid P_{EP}, T] = \exp\{\alpha_{P_{EP},T} + r(\lambda_4^T)[\beta_4 + \lambda_5^T\beta_3]P_{EP} + (\beta_2 + r(\lambda_2^{P_{EP}})r(\lambda_3)\beta_3 + r(\lambda_2^{P_{EP}})r(\lambda_1^{P_{EP},T})\beta_4)T\}. \quad (7)$$

$$E[P_{GS} \mid P_{EP}, T] = r(\lambda_4^T)\lambda_1^T P_{EP} + r(\lambda_2^{P_{EP}})T. \quad (8)$$

In fact, $r(\lambda_1^{P_{EP},T}) = 0$, since $\sigma^2_{X|P_{EP}} = 0$, so the coefficient for *T* in the Poisson regression equation for *Y* is $\{\beta_2 + r(\lambda_2^{P_{EP}})r(\lambda_3)\beta_3\}$. The formula that we used previously to obtain an unbiased estimate of the effect of treatment (the coefficient of *T* in the Poisson regression of *Y* on EP and *T*, minus $r(\lambda_2^{P_{EP}})/r(\lambda_4^T)\lambda_1^T$ times the coefficient of EP) need no longer be unbiased. It becomes

$$\beta_2 + r(\lambda_2^{P_{EP}})r(\lambda_3^{P_{EP}})\beta_3 - \frac{r(\lambda_2^{P_{EP}})}{r(\lambda_4^T)\lambda_1^T} r(\lambda_4^T)[\beta_4 + \lambda_5^T\beta_3] = \beta_2 + r(\lambda_2^{P_{EP}})\left\{r(\lambda_3^{P_{EP}})\beta_3 - \frac{\beta_4 + \lambda_5^T\beta_3}{\lambda_1^T}\right\}. \quad (9)$$

The bias associated with this formula will be $r(\lambda_2^{P_{EP}})\left\{r(\lambda_3^{P_{EP}})\beta_3 - \frac{\beta_4+\lambda_5^T\beta_3}{\lambda_1^T}\right\}$, and the method of calculating the effect of treatment will be unbiased if either $r(\lambda_2^{P_{EP}}) = 0$ (i.e., the treatment assignment is independent of *C* given P_{EP}) or $r(\lambda_3^{P_{EP}})\beta_3 = \frac{\beta_4+\lambda_5^T\beta_3}{\lambda_1^T}$. This propensity-score-calibrated estimate of the treatment effect may be closer to the true value (if the signs of $r(\lambda_3^{P_{EP}})\beta_3$ and $\frac{\beta_4+\lambda_5^T\beta_3}{\lambda_1^T}$ are the same and if $|r(\lambda_3^{P_{EP}})\beta_3| > |\frac{\beta_4+\lambda_5^T\beta_3}{\lambda_1^T}|$).

Simulation study

To test the PSC bias predicted by the above formula against the empirical PSC bias, we conducted a simulation study with 2 independent continuous confounders *X* and *C*. We will discuss generalizations below. We describe the simulation study in Web Appendix 2 and present the simulation parameters and their values in Table 1.

Empirical example

To illustrate the implementation of PSC and of the proposed PSC bias estimation, we use the setting where PSC was first implemented (5), addressing the paradoxical inverse association between the use of nonsteroidal antiinflammatory drugs (NSAIDs) and short-term all-cause mortality in older adults (15). This can be found in Web Appendix 3.

Table 1. Parameters and Values Used in the Basic and Alternative Scenarios

Parameter		Parameter Value(s)	
Notation	Meaning	Basic Scenario	Alternative Scenarios
P_T	Prevalence of exposure of interest (T)	0.20	
IR_Y	Incidence rate of disease (Y)	0.01	0.005, 0.05, 0.10
IRR_{TY}	Incidence rate ratio for association between exposure (T) and disease (Y)	1.0	0.5, 2.0
OR_{XT}	Odds ratio for association between a measured confounder (X) and exposure (T) (independent of C)	0.5	
IRR_{XY}	Incidence rate ratio for association between a measured confounder (X) and disease (Y) (independent of T and C)	2.0	
OR_{CT}	Odds ratio for association between an unmeasured confounder (C) and exposure (T) (independent of X)	0.5	0.40, 0.67, 1.0, 1.5, 2.0, 2.5
IRR_{CY}	Incidence rate ratio for association between an unmeasured confounder (C) and disease (Y) (independent of T and X)	2.0	0.5, 0.67, 1.0, 1.5
N_{main}	Size of main study	10,000	
P_{val}	Proportion of observations in validation study	0.1	0.02, 0.05, 0.2, 0.5

Abbreviations: IR, incidence rate; IRR, incidence rate ratio; OR, odds ratio.

RESULTS

Simulation study

We varied the incidence rate (IR) of disease IR_Y, the incidence rate ratio (IRR) for the exposure-disease association IRR_{TY}, the odds ratio (OR) for the unobserved confounder-exposure association OR_{CT}, the IRR for the unobserved confounder-disease association IRR_{CY}, the size of the main study N_{main}, and the proportion of persons in the random validation sample P_{val} around the value of the basic scenario, while keeping all other parameters constant at the value of the basic scenario (see italic type in Table 2). The true IRR_{TY} is 1.0 in all scenarios, except for the 2 rows with $IRR_{TY} = 2$ and $IRR_{TY} = 0.5$. We present the median crude IRR_{TY} and the median IRR_{TY} adjusting for the observed covariate X only.

For PSC, we present the median observed bias and the median predicted bias according to equation 9 using study-specific estimates for all parameters, with the exception of β_3 ($\log IRR_{CY}$), where we substituted the true parameter. We also present the median value and interquartile range for IRR_{TY}, the median percentage of bias reduction, the percentage of studies in which the 95% confidence interval covers the true IRR_{TY}, and a measure of surrogacy (S), with values close to 100 indicating that surrogacy holds.

In the basic scenario, the crude IRR_{TY} is 0.42. Compared with the IRR_{TY} of 0.62 after control for confounding by the observed covariate X, adjusting for the unmeasured confounder C using PSC leads to an almost unbiased median IRR_{TY} of 1.09 and therefore a bias reduction of 100%. The 95% confidence interval has nominal coverage, and a value of 88% for our assessment of surrogacy (percentage of variance in Y explained by P_{GS} and P_{EP} which is due to P_{GS}) indicates that surrogacy might be a reasonable assumption.

When we vary one parameter at a time, PSC is clearly biased in some scenarios. The bias can be predicted almost perfectly by the bias formula derived above as equation 9, using the path analysis framework. The correlation between the median predicted bias and the median observed bias over the scenarios assessed is 0.998.

We recently characterized scenarios in which PSC is unbiased on the basis of surrogacy. The results presented provide a more refined view. Although obviously the inverse correlation between bias and our measure of surrogacy is strong (-0.89 for the median values over the scenarios assessed), the magnitude of bias is not well predicted by our measure of surrogacy, and some scenarios show larger bias despite less pronounced violation of surrogacy than others.

Empirical example

The crude hazard ratio for NSAID use and all-cause mortality is 0.68 (95% confidence interval (CI): 0.66, 0.71). Adjusting for age removes some of the confounding (hazard ratio = 0.72, 95% CI: 0.70, 0.75). Using PSC to additionally control for not having had an influenza shot (as a proxy for frailty) results in a hazard ratio of 0.76 (95% CI: 0.73, 0.80). The results of the PSC bias prediction are summarized in Table 3. Assuming, for example, a mortality hazard ratio of 2.0 for not having had a flu shot, the predicted bias for the log hazard ratio was 0.040 (Table 3), resulting in a bias-adjusted PSC estimate of the hazard ratio equal to 0.73.

DISCUSSION

Here we have presented a novel framework for PSC that allows us to predict the presence and magnitude of bias and thus correct for it when applying PSC in the simplest possible

Table 2. Crude and Adjusted Incidence Rate Ratios for the Exposure-Outcome (*T*-*Y*) Association (Crude, Adjusted for *X* Only, and Adjusted for *X* and *C* Using Propensity Score Calibration)[a]

Parameter[b] and Value	IRR_{TY}		Propensity Score Calibration					
	Crude	*X*[c]	Predicted Bias[d]	Observed Bias	Bias Reduction, %[e]	IRR_{TY}	Coverage Probability, %[f]	Surrogacy, %[g]
IR_Y								
0.005	0.42	0.62	0.05	0.00	75	1.00	95	79
0.01	*0.42*	*0.62*	*0.06*	*0.09*	*100*	*1.09*	*95*	*88*
0.05	0.42	0.63	0.07	0.06	113	1.07	90	98
0.10	0.43	0.64	0.07	0.06	114	1.07	86	99
IRR_{TY}								
2.0	0.85	1.27	0.06	0.09	118	2.18	93	95
1.0	*0.42*	*0.62*	*0.06*	*0.09*	*100*	*1.09*	*95*	*88*
0.5	0.21	0.31	0.06	0.07	79	0.54	93	95
OR_{CT}								
0.4	0.37	0.55	0.30	0.29	140	1.33	91	91
0.5	*0.42*	*0.62*	*0.06*	*0.09*	*100*	*1.09*	*95*	*88*
0.67	0.49	0.76	−0.09	−0.08	42	0.92	94	93
1.0	0.63	0.99	0.00	0.00	—[h]	1.00	96	54
1.5	0.82	1.29	0.47	0.51	−57	1.67	75	10
2.0	0.99	1.55	1.07	1.10	−135	3.00	25	6
2.5	1.13	1.74	1.67	1.54	−189	4.69	6	6
IRR_{CY}								
0.5	0.99	1.54	1.09	1.04	−132	2.83	27	6
0.67	0.82	1.27	0.87	0.83	−156	2.29	55	14
1.0	0.64	0.98	0.57	0.54	—	1.71	81	50
1.5	0.49	0.76	0.27	0.29	100	1.33	90	85
2.0	*0.42*	*0.62*	*0.06*	*0.09*	*100*	*1.09*	*95*	*88*
IRR_{CY} ($OR_{CT} = 2.0$)								
0.5	0.41	0.62	0.06	0.00	100	1.00	95	93
0.67	0.50	0.75	0.27	0.26	119	1.30	93	83
1.0	0.64	0.98	0.56	0.56	—	1.75	83	48
1.5	0.84	1.31	0.86	0.86	−139	2.38	53	12
2.0	0.99	1.55	1.07	1.10	−135	3.00	25	6
P_{val}								
0.02	0.42	0.63	0.08	0.10	100	1.11	91	81
0.05	0.42	0.63	0.06	0.09	100	1.09	94	91
0.1	*0.42*	*0.62*	*0.06*	*0.09*	*100*	*1.09*	*95*	*88*
0.2	0.42	0.63	0.06	0.10	100	1.10	94	97
0.5	0.42	0.63	0.06	0.09	100	1.09	94	98

Abbreviations: EP, error-prone; GS, gold standard; IR, incidence rate; IRR, incidence rate ratio; OR, odds ratio.

[a] Median values from 1,000 simulations for each scenario (row); true $IRR_{TY} = 1$ in all scenarios except when IRR_{TY} is the parameter varied.

[b] For the definition and range of these parameters, refer to Table 1. Only 1 parameter 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 IRR_{CY}, where OR_{CT} is 2.0 instead of 0.5 to show the importance of the direction of confounding by the unmeasured confounder *C*).

[c] Controlling for measured covariate *X* only.

[d] Using equation 9 and the true parameters for β_3 (log IRR_{CY}); all other parameters are estimated from the actual data.

[e] Median percentage of bias reduction according to Cochran (29): 0% equals no improvement over the analysis controlling for measured covariate *X* only, and 100% equals no residual bias (truth).

[f] Coverage probability of the empirical 95% confidence interval (2.5th–97.5th percentiles) obtained from 1,000 bootstrap samples.

[g] Assessment of the surrogacy assumption for propensity score calibration: percentage of variance in *Y* explained by PS_{GS} and PS_{EP} which is due to PS_{GS}; calculated as the ratio of the likelihood ratio comparing the logistic regression model $\text{logit}(Y) = v_0' + v_1'T + v_2'PS_{GS}$ with the nested logistic regression model $\text{logit}(Y) = v_0'' + v_1''T$ to the likelihood ratio comparing the logistic regression model $\text{logit}(Y) = v_0 + v_1T + v_2PS_{GS} + v_3PS_{EP}$ with the nested logistic regression model $\text{logit}(Y) = v_0'' + v_1''T$ times 100. Values close to the maximum possible value (100%) indicate surrogacy.

[h] Undefined, since the expected value of the denominator is 0 (no bias controlling for observed covariate *X* only).

Table 3. Models, Parameters, and Estimated Values for Propensity Score Calibration Bias Prediction[a] in the Empirical Example

Model	Study	Parameter of Interest	Notation (Equation 9)	Estimated Value
$E(P_{GS}) = \beta_0 + \beta_T T + \beta_{Pep} P_{EP}$	Validation	β_T	$r(\lambda_2^{P_{EP}})$	0.00395
$E(C) = \beta_0 + \beta_{Pgs} P_{GS} + \beta_{Pep} P_{EP}$	Validation	β_{Pgs}	$r(\lambda_3^{P_{EP}})$	21.88586
$\text{Log}(\lambda(t\|T,X)/\lambda_0(t)) = \beta_0 + \beta_T T + \beta_C C$	Not applicable[b]	β_C	β_3	−0.693[b]
$E(C) = \beta_0 + \beta_T T + \beta_X X$	Validation	β_X	λ_5^T	0.00268
$\text{Log}(\lambda(t\|T,X)/\lambda_0(t)) = \beta_0 + \beta_T T + \beta_X X$	Main	β_X	β_4	0.04564
$E(P_{GS}) = \beta_0 + \beta_T T + \beta_X X$	Validation	β_X	λ_1^T	−0.00174

Abbreviations: EP, error-prone; GS, gold standard.

[a] Predicted propensity score calibration bias on the log scale $r(\lambda_2^{P_{EP}})\left\{ r(\lambda_3^{P_{EP}})\beta_3 - \frac{\beta_4 + \lambda_5^T \beta_3}{\lambda_1^T} \right\}$ (equation 9): $0.00395\{21.88586 \times 0.693 - (0.04564 + 0.00268 \times -0.693)/-0.00174\} = 0.040$.

[b] Assumed value for the association between C (receipt of a flu shot) and all-cause mortality from the literature (hazard ratio = 0.5). This value cannot be estimated from the data because C is unobserved in the main study and Y is unobserved in the validation study.

setting with a single observed confounder and a single unobserved confounder. The bias estimation is based on parameters that can be estimated in the main study and the (external) cross-sectional validation study and a single unobservable parameter IRR_{CY}—that is, the association of the unmeasured confounder with the outcome, which is estimable only in a validation study with adequate numbers of disease outcomes. This bias equation advances the applicability of PSC as a sensitivity analysis in the setting of a cross-sectional validation study, as shown in our empirical example. Alternative sensitivity methods include performing maximum likelihood or multiple imputation with any parameters needed for identification (e.g., IRR_{CY}) set to various values, or Bayesian analyses with various priors on these parameters.

Any method that tries to address uncontrolled confounding without observing the joint distribution of the confounder with exposure and disease will need to rely on strong assumptions or prior distributions. Following the assumptions for regression calibration, we evaluated surrogacy as a good candidate and were able to show that a measure of surrogacy indeed helped to separate scenarios in which PSC was biased from scenarios in which PSC was unbiased in our simulation studies. The scenarios where PSC was unbiased had in common the fact that the direction of confounding of the observed confounder(s) X was the same as the direction of confounding by the unobserved confounder C (5, 6). Here we present a more refined view of bias encountered when applying PSC based on a balance of the magnitudes of observed and unobserved confounding rather than just the directions.

In our empirical example in Web Appendix 3, we assume that NSAID use has no protective effect on mortality and that any inverse association is thus attributable to confounding (16). Jackson et al. (17) observed an implausible mortality reduction of over 60% (relative risk = 0.39) associated with receipt of a flu shot prior to the flu season in older adults. The most plausible explanation for both findings is unmeasured confounding by frailty: Patients close to death are less likely to receive certain treatments, especially preventive treatments (15). Thus, we used not having had a flu shot as a proxy for frailty in our example. Not having had a flu shot was associated with an approximately 50% lower prevalence of NSAID use in the validation study, and our best guess is that it is associated with increased mortality. Based on this assumption and information on flu shots available only in the validation data, PSC improves our estimate of the NSAID-mortality association.

The major strength of the proposed bias formula is that, at least under the assumption of independence of observed and unobserved confounders, only 1 parameter needs to be substituted. All others can be estimated from the data. Prior knowledge on the relative risk associated with the unmeasured confounder (IRR_{CY}) might be easier to obtain than prior knowledge on the association of the unobserved confounder with exposure (OR_{CT}). This can be seen in our empirical example. Results from several nonexperimental studies indicate that older adults not receiving the flu shot have approximately twice the short-term mortality as those who receive a flu shot (e.g., see Jackson et al. (17)). Data on the association between NSAID use and having had a flu shot, however, are sparse or nonexistent. Note that most of the flu shot-mortality association can be observed prior to the flu season and thus is probably due to unmeasured confounding by frailty; that is, it is not causal (17, 18). We thus use the reported flu shot-mortality hazard ratio that is confounded by frailty as our best guess to estimate the PSC bias.

Unless a factor unobserved in the main study is supposed to be a risk factor for disease, information on such a factor will rarely be sought using a validation study. An intuitive advantage of PSC is that it uses all of the information available in the cross-sectional validation study to estimate the joint association between the observed and unobserved covariates and exposure (P_{GS}). By sampling from prior distributions for the relative risk associated with the unmeasured confounder, PSC can be employed to create uncertainty intervals for the target effect and thus become part of bias analysis (19–22).

The major limitation of the proposed bias formula is that it is based on the simplest setting of 2 covariates, one measured (X) and one unmeasured (C). Because linear combinations of multivariate-normal variables are normal, our results should

apply to settings in which *X*, *C*, and their combination are multivariate-normal. Settings with nonnormal (e.g., binary) confounders will be more complicated and possibly intractable analytically. Nonetheless, experience in other problems (e.g., effects of case-control matching (23), prediction for multiple imputation (4)) suggests that normal cases can be reasonable guides to scenarios involving categorical variables. Given the rough nature of external bias adjustments and sensitivity analysis, we would expect similar practical conclusions to hold for PSC. The major limitation of PSC for sensitivity analysis, however, will be the lack of availability of cross-sectional validation studies that are representative of the main study, and more specifically, possible nontransportability of models.

The use of DAGs for propensity scores is nonstandard. Furthermore, there may be different DAGs with which to describe the PSC setting. The DAG we propose for PSC allowed us to predict bias from PSC, however, and thus has empirical validity, at least over the scenarios assessed in the simulations. The primary drawback of path analysis in this context is its assumption of linear associations among all variables in the model. Furthermore, interactions between variables cannot be incorporated perfectly. The use of linear regression for dichotomous exposures is also found in instrumental-variable analysis (24).

Other criticisms of path analysis (25, 26) do not apply to this context. We assume that we know the causal model apart from the exposure effects; it is not deduced from the path analysis. Path analysis using correlation coefficients depends on the standard deviation of the variables concerned, but here we are using regression coefficients, which are not affected by the standard deviation (27, 28). Our primary assumption then is that these effects are the same (i.e., the models are transportable).

Conclusions from simulations are by definition restricted to the scenarios assessed. Our previous simulations did not consider scenarios with considerable asymmetry in strength of confounding between observed and unobserved variables which we have evaluated here.

We found that the magnitude and direction of bias encountered when applying PSC to control for a single confounder measured only in a cross-sectional validation study may often be small and can be predicted on the basis of a bias equation derived from DAGs and path analysis. To do so, we need to assume a plausible value for the single unobservable parameter, the relative risk associated with the unobserved confounder. This is an important step towards the understanding and implementation of PSC as a sensitivity analysis. Its utility will depend on extension of the bias formula to the multivariable setting.

ACKNOWLEDGMENTS

Author affiliations: Arthritis Research UK Epidemiology Unit, School of Translational Medicine, University of Manchester, Manchester, United Kingdom (Mark Lunt); Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts (Robert J. Glynn, Jerry Avorn); RTI Health Solutions, Research Triangle Park, North Carolina (Kenneth J. Rothman); and Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (Til Stürmer).

This project was funded by grants (R01 AG023178 and R01 AG018833) from the National Institute on Aging and a grant (17552) from Arthritis Research UK.

The authors thank Dr. Sander Greenland for his critique of propensity score calibration based on DAGs following their presentation of the method during the 2007 Joint Statistical Meetings of the American Statistical Association, the International Biometric Society, the Institute of Mathematical Statistics, and the Statistical Society of Canada (Salt Lake City, Utah, July 29–August 2, 2007), as well as his valuable comments on an early version of the manuscript.

Conflict of interest: none declared.

REFERENCES

1. Stürmer T, Glynn RJ, Rothman KJ, et al. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. *Med Care*. 2007; 45(10 suppl 2):S158–S165.
2. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons, Inc; 1987.
3. Little RJ. Regression with missing X's: a review. *J Am Stat Assoc.* 1992;87(420):1227–1237.
4. Little RJA, Rubin DB. *Statistical Analysis With Missing Data*. 2nd ed. New York, NY: John Wiley & Sons, Inc; 2002.
5. 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(3):279–289.
6. Stürmer T, Schneeweiss S, Rothman KJ, et al. Performance of propensity score calibration—a simulation study. *Am J Epidemiol*. 2007;165(10):1110–1118.
7. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
8. 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(9):1051–1069.
9. 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(4): 734–745.
10. Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models*. London, United Kingdom: Chapman & Hall Ltd; 1995.
11. Wright S. Correlation and causation. *J Agric Res.* 1921; 20:557–585.
12. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology.* 1999;10(1):37–48.
13. Pearl J. *Causality.* 2nd ed. New York, NY: Cambridge University Press; 2009.
14. Glymour MM, Greenland S. Causal diagrams. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology.* 3rd ed.

Philadelphia, PA: Lippincott Williams & Wilkins; 2008: 183–212.
15. Glynn RJ, Knight EL, Levin R, et al. Paradoxical relations of drug treatment with mortality in older persons. *Epidemiology.* 2001;12(6):682–689.
16. 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(9):891–898.
17. Jackson LA, Jackson ML, Nelson JC, et al. Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *Int J Epidemiol*. 2006;35(2):337–344.
18. McGrath LJ, Kshirsagar AV, Cole SR, et al. Influenza vaccine effectiveness in patients on hemodialysis: an analysis of a natural experiment. *Arch Internal Med.* 2012;172(7):548–554.
19. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology.* 2003;14(4):451–458.
20. Greenland S. Multiple-bias modelling for analysis of observational data. *J R Stat Soc Ser A*. 2005;168(2):267–306.
21. Schneeweiss S, Glynn RJ, Tsai EH, et al. Adjusting for unmeasured confounders in pharmacoepidemiologic claims data using external information: the example of COX2 inhibitors and myocardial infarction. *Epidemiology.* 2005;16(1): 17–24.
22. Greenland S, Lash TL. Bias analysis. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology.* 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008: 345–380.
23. Thomas DC, Greenland S. The relative efficiencies of matched and independent sample designs for case-control studies. *J Chronic Dis*. 1983;36(10):685–697.
24. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*. 2000;29(4):722–729.
25. Karlin S, Cameron EC, Chakraborty R. Path analysis in genetic epidemiology: a critique. *Am J Hum Genet.* 1983;35(4):695–732.
26. Freedman DA. As others see us: a case study in path analysis. *J Educ Stat*. 1987;12:101–128.
27. Greenland S, Schlesselman JJ, Criqui MH. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *Am J Epidemiol*. 1986;123(2):203–208.
28. Greenland S, Maclure M, Schlesselman JJ, et al. Standardized regression coefficients: a further critique and review of some alternatives. *Epidemiology.* 1991;2(5):387–392.
29. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 1968;24(2):295–313.