Adjusting Effect Estimates for Unmeasured Confounding with Validation Data using Propensity Score Calibration

Til Stürmer1,2, Sebastian Schneeweiss1,3, Jerry Avorn1, and Robert J. Glynn1,2,4

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 Biostatistics, Harvard School of Public Health, Boston, MA.

Received for publication December 12, 2003; accepted for publication April 20, 2005.

Often, data on important confounders are not available in cohort studies. Sensitivity analyses based on the relation of single, but not multiple, unmeasured confounders with an exposure of interest in a separate validation study have been proposed. In this paper, the authors controlled for measured confounding in the main cohort using propensity scores (PS’s) and addressed unmeasured confounding by estimating two additional PS’s in a validation study. The “error-prone” PS exclusively used information available in the main cohort. The “gold standard” PS additionally included data on covariates available only in the validation study. Based on these two PS’s in the validation study, regression calibration was applied to adjust regression coefficients. This propensity score calibration (PSC) adjusts for unmeasured confounding in cohort studies with validation data under certain, usually untestable, assumptions. The authors used PSC to assess the relation between nonsteroidal antiinflammatory drugs (NSAIDs) and 1-year mortality in a large cohort of elderly persons. “Traditional” adjustment resulted in a hazard ratio for NSAID users of 0.80 (95% confidence interval (CI): 0.77, 0.83) as compared with an unadjusted hazard ratio of 0.68 (95% CI: 0.66, 0.71). Application of PSC resulted in a more plausible hazard ratio of 1.06 (95% CI: 1.00, 1.12). Until the validity and limitations of PSC have been assessed in different settings, the method should be seen as a sensitivity analysis.

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

Abbreviations: CI, confidence interval; MCBS, Medicare Current Beneficiary Survey; NSAID, nonsteroidal antiinflammatory drug; PS, propensity score; PSC, propensity score calibration.

Unmeasured confounding can be a major source of bias in observational research. Cohort studies often lack measures of important potential confounders, such as smoking and body mass index in pharmacoepidemiologic studies using claims data, or laboratory or blood pressure measurements in questionnaire-based studies. Various methods have been proposed for assessing the sensitivity of observed associations to the possible effect of unobserved confounders (1–6). Recent studies on the side effects of medications have taken this research one step further by considering external data describing the distributions of confounders in a validation sample and their association with disease based on the medical literature. This information can be used to adjust estimates of the association observed in the main study (7, 8). However, these approaches do not take the joint distribution of multiple unobserved confounders into account. The latter methods also treat the external distributions of confounders as known when in fact they are estimates, usually made from small to moderate-sized samples.

Our objective here was to propose a method of adjusting for multiple unmeasured confounders in a cohort study. The suggested method incorporates information from an

Correspondence to Dr. Til Stürmer, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women’s Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120 (e-mail: til.sturmer@post.harvard.edu).
external validation sample to calibrate the propensity score (PS) used to adjust for confounding. After a brief introduction to background information on PS’s and regression calibration, we develop the rationale for our proposed propensity score calibration (PSC) approach and illustrate its application in a cohort study of the association between nonsteroidal antiinflammatory drugs (NSAIDs) and 1-year mortality in the elderly.

**BACKGROUND**

**Propensity score**

PS is defined as the conditional probability of exposure to a drug or other potential risk factor given observed covariates (9). Each subject has a vector of observed covariates, \( \mathbf{X} \), and an indicator of exposure (or treatment), \( E \), where \( E = 1 \) if exposed and \( E = 0 \) if unexposed. The PS, \( e(\mathbf{X}) \), is the probability of exposure for a person with covariates \( \mathbf{X} \), that is,

\[
e(\mathbf{X}) = \text{prob}(E = 1|\mathbf{X}).
\]

This PS is usually estimated from the data at hand using multivariable logistic regression. Persons with the same estimated PS are then thought to have the same chance of being exposed, although they may have very different \( \mathbf{X} \)’s. As a group, however, treated and untreated subjects paired on the same PS will have similar distributions of \( \mathbf{X} \) (10).

Once a PS has been estimated, it can be used in different ways to control for confounding in cohort studies. Applications include matching on the estimated PS, multivariate adjustment by subclassification on the estimated PS, or modeling of the estimated PS-outcome association, as well as combinations of these methods with “traditional” multivariable outcome modeling (11). In the current context, we will consider modeling of the estimated PS-outcome association in a Cox proportional hazards model

\[
h(t|E,e(\mathbf{X})) = h_0(t)\exp\{\beta_E E + \beta_X e(\mathbf{X})\}.
\]

This model is used to allow the application of the proposed PSC, although we do not suggest that such a model would be preferable to a model including all covariates or other ways of using the estimated PS. In theory, conditioning on the estimated PS should lead to exchangeability of exposed and unexposed subjects and thus yield unconfounded treatment estimates in any relative risk or hazard model, as long as there is no unmeasured confounding. Use of the estimated PS in a Cox proportional hazards model has been described by D’Agostino (12) and previously applied by Muller et al. (13).

**Regression calibration**

In the context of generally sparse use of methods of correcting effect estimates for measurement error in epidemiology, regression calibration is the most widely used approach (14–17). It estimates a linear measurement error model of the true or “gold standard” variable for one measured with error in a validation study, and uses the resulting regression parameters \( \hat{\lambda} \) to correct the “naive” regression estimates \( \hat{\beta} \) obtained from error-prone covariate data in the main study.

**Propensity score calibration**

A PS estimated when additional confounders are unobserved can be viewed as a variable measured with error, either because of the lack of information on important predictors of exposure (unmeasured variables) or because of imperfectly measured predictors.

A validation study usually includes a gold standard measure of the variable measured with error in the main study, in addition to the error-prone measurement (already available in the main study). Since the PS is estimated without information concerning outcomes, a cross-sectional validation study is sufficient for assessing the error in the main study PS. This can be achieved by directly comparing a PS containing the same information as the main study PS (i.e., the error-prone PS) with the validation study PS, which includes additional important determinants of the exposure of interest that are unobserved in the main study. The latter estimated PS can be seen as the gold standard PS, whereas the former is the error-prone PS, identical in its determinants to the error-prone PS in the main study.

Once the error in the main study PS is estimated in the validation study, regression calibration can be used to adjust the main study PS for that error. This will also adjust estimates of the effect of the exposure of interest on disease risk for bias due to residual confounding that results from the mismeasured PS.

Following the general notation introduced for the PS, we would define

\[
e(\mathbf{X}_{EP}) = \text{prob}(E = 1|\mathbf{X}_{EP})
\]

as the error-prone (EP) variable and

\[
e(\mathbf{X}_{GS}) = \text{prob}(E = 1|\mathbf{X}_{GS})
\]

as the gold standard (GS), with \( \mathbf{X}_{GS} \) generally being an expanded set of covariates including \( \mathbf{X}_{EP} \). Under the assumption that \( e(\mathbf{X}_{EP}) \) is a surrogate for \( e(\mathbf{X}_{GS}) \)—that is, that \( e(\mathbf{X}_{EP}) \) is independent of the outcome of interest given \( \mathbf{X}_{GS} \) and exposure—the measurement error model then is

\[
E[e(\mathbf{X}_{GS})|E,e(\mathbf{X}_{EP})] = \lambda_0 + \lambda_E E + \lambda_X e(\mathbf{X}_{EP}).
\]

If only one covariate is present, the estimate adjusted for measurement error is very easily obtained by dividing the “naive” regression estimate in the main study by the regression parameter from the measurement error model in the validation sample (\( \hat{\beta}/\hat{\lambda} \)). For two independent variables in the Cox proportional hazards outcome model, one measured with error (the PS) and one without (the exposure), the target model would be

\[
h(t|E,e(\mathbf{X}_{GS})) = h_0(t)\exp\{\beta_E E + \beta_X e(\mathbf{X}_{GS})\}
\]

if information on all covariates were available in the main cohort. The regression calibration-adjusted (15) estimator for the effect of \( E \) then is

\[
\hat{\beta}_E = \hat{\beta}_E - \hat{\lambda}_E \hat{\beta}_X/\hat{\lambda}_X.
\]
and the adjusted estimator for \( \hat{e}(\mathbf{X}) \), the effect of the true PS, is

\[
\hat{\beta}_X^{*} = \hat{\beta}_X / \hat{\pi}_X.
\]  

(8)

Adjusted estimates for the variances can then be calculated, accounting for the additional uncertainty caused by the estimation of \( \lambda \) in the validation study (15). The method has been developed for cohort studies and produces approximately consistent estimates if the measurement error variance is not too large. In the case of an “alloyed” gold standard, nondifferential error is assumed. The resulting estimates are still consistent if the “alloyed” gold standard is unbiased for the true gold standard (18). Regression calibration leads to approximate consistency in Cox proportional hazards models and can therefore be used to remove most of the bias due to measurement error (16, 19).

The method is easily applicable using a SAS macro that can be obtained from Dr. Donna Spiegelman at the Harvard School of Public Health (http://www.hsph.harvard.edu/faculty/spiegelman/blinplus.html).

**METHODS AND RESULTS**

**Study populations**

**Main study population.** To test this approach, we identified a main study population assembled for an analysis of pain medication use in the elderly. It consisted of all community-dwelling New Jersey residents aged 65 years or older who filled prescriptions within the Medicaid program or the Pharmaceutical Assistance to the Aged and Disabled program and who were hospitalized at any time between January 1, 1995, and December 31, 1997. Eligible persons were those who filled a prescription for any drug within 120 days before hospitalization and another prescription more than 365 days before hospitalization, since covariates were assessed during that time period.

For all subjects, we extracted data on the following variables: age, sex, race, all prescriptions filled within 120 days before the index date, all diagnoses assigned within 365 days before the index date, number of hospitalizations, and number of physician visits. Time to death or 365 days follow-up (whichever came first) was assessed starting from the date of hospital admission, on the basis of linkage to Medicare files (20).

**Validation study.** The Medicare Current Beneficiary Survey (MCBS) is conducted in a sample of beneficiaries selected each year to be representative of the current Medicare population, including both aged and disabled beneficiaries living in the community or in institutions. Data, including data on medication use over the past 4 months (verified by inspection of medication containers), are obtained from face-to-face interviews and linked to Medicare claims data. The survey has a high response rate (85–95 percent) and very high data completeness (21–23).

The MCBS data used for the validation study in this analysis were drawn from a list of all persons enrolled in Medicare on January 1, 1999. As in our main study, the validation study population was restricted to persons aged 65 years or older who were living in the community (\( n = 10,446 \)). To make the validation study population more comparable to that of the main study, we randomly selected MCBS enrollees so that their age (three categories) and sex distribution matched the one observed in the main cohort (frequency matching). This resulted in our having 5,108 MCBS subjects for all subsequent analyses.

Table 1 describes the main study population of 103,133 and the validation study population of 5,108. Because of matching, the age and sex distributions were very similar. The main study population was more diverse with respect to race than the validation study. The main differences were observed with respect to comorbidity, which was much more prevalent in the main study than in the validation study. Accordingly, persons in the main study population also had more physician visits and hospitalizations. NSAID use was more prevalent in the main study (17.8 percent) than in the validation study (12.1 percent).

**Propensity of NSAID use**

**Main study.** We first estimated the propensity of NSAID use during the previous 4 months (yes/no) in the main study (claims data set) using logistic regression. Since over 18,000 persons used NSAIDs, no variable selection was conducted; instead, all covariates presented in table 1 were entered into the model.

**Validation study.** We then estimated two different PS’s in the validation study, using similar logistic regression models. The first model used covariates identical to those in the main study, and the second used these covariates plus additional information available only in the validation study. The first model corresponds to the error-prone PS model that lacks important confounders and the second to the gold standard PS. For the gold standard PS, we addressed unmeasured confounding and imperfectly measured confounding in two ways. To address unmeasured confounding, we used the same variables as in the error-prone PS in combination with variables not available in the main study (i.e., smoking, body mass index, activities of daily living, education, and income). To additionally address residual confounding due to imperfectly measured confounders (24), we used self-reports (“Have you ever been told that you had . . .”) of lifetime rheumatoid arthritis or osteoarthritis in addition to the claims data diagnostic codes for arthritis. We hypothesized that self-reports would be more predictive of NSAID use than claims data alone. The predictive value of each PS model was estimated using the area under the receiver operating characteristic curve (c statistic) (25).

**Determinants of NSAID use.** The results obtained from these four different PS models are presented in table 2. On the left side of the table, we present the logistic regression results from the main study using claims data only. The three columns on the right describe the results of different models for the propensity of NSAID use in the validation study. The PS model called “error-prone” follows the PS model of the main study, using the same variables. The PS model called “unmeasured” uses the same variables along with added variables not available in the main data set. Finally, the PS model called “un- and imperfectly measured”
TABLE 1. Description of the study populations in the main study and the external validation study

<table>
<thead>
<tr>
<th></th>
<th>Main study (n = 103,133)</th>
<th>External validation study (n = 5,108)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>%</td>
</tr>
<tr>
<td>Mean age (years)</td>
<td>78.8</td>
<td>(7.6)</td>
</tr>
<tr>
<td>Female gender</td>
<td>76,782</td>
<td>75</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>82,039</td>
<td>80</td>
</tr>
<tr>
<td>Black</td>
<td>13,720</td>
<td>13</td>
</tr>
<tr>
<td>Other</td>
<td>7,374</td>
<td>7</td>
</tr>
<tr>
<td>Diagnoses based on claims data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>10,319</td>
<td>10</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>34,466</td>
<td>33</td>
</tr>
<tr>
<td>Diabetes</td>
<td>31,164</td>
<td>30</td>
</tr>
<tr>
<td>Cancer</td>
<td>17,364</td>
<td>17</td>
</tr>
<tr>
<td>Arthritis (RA† or OA‡)</td>
<td>4,846</td>
<td>5</td>
</tr>
<tr>
<td>Diagnoses based on self-reports</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arthritis (RA or OA)</td>
<td>3,504</td>
<td>69</td>
</tr>
</tbody>
</table>

Use of the health-care system

<table>
<thead>
<tr>
<th>No. of physician visits</th>
<th>Main study</th>
<th>External validation study</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–5</td>
<td>38,924</td>
<td>4,312</td>
</tr>
<tr>
<td>6–11</td>
<td>33,450</td>
<td>513</td>
</tr>
<tr>
<td>≥12</td>
<td>30,759</td>
<td>283</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No. of hospitalizations</th>
<th>Main study</th>
<th>External validation study</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>64,703</td>
<td>4,157</td>
</tr>
<tr>
<td>1</td>
<td>22,293</td>
<td>603</td>
</tr>
<tr>
<td>≥2</td>
<td>16,137</td>
<td>348</td>
</tr>
</tbody>
</table>

Use of medications

<table>
<thead>
<tr>
<th>Nonsteroidal antiinflammatory drugs</th>
<th>Main study</th>
<th>External validation study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18,326</td>
<td>620</td>
</tr>
</tbody>
</table>

| Thiazides                          | 6,377      | 108                       |
| Steroids                           | 7,848      | 490                       |
| Anticoagulants                     | 7,613      | 321                       |

* Numbers in parentheses, standard deviation.
† RA, rheumatoid arthritis; OA, osteoarthritis.

contains the same variables as the PS model called “unmeasured” plus self-reported arthritis in addition to the claims data diagnosis of arthritis.

Predictive value for exposure. The area under the receiver operating characteristic curve of the PS in the validation study using claims data only (0.60) is similar to the one from the main study (0.63), indicating only a modest capacity to predict NSAID use. Adding information on body mass index and performance of activities of daily living (as well as education, income, and smoking status, which are less pronounced predictors of NSAID use) increases the predictive value to 0.66. Adding self-reported arthritis, which is a strong predictor of NSAID use (odds ratio = 4.1, 95 percent confidence interval (CI): 3.1, 5.5) independently of the claims data diagnosis of arthritis, further increases the predictive value of the PS model (area under the receiver operating characteristic curve = 0.71).

Unadjusted association between NSAID use and mortality

During the follow-up period of 1 year, 21,928 (21.3 percent) hospitalized patients died, either during hospitalization or afterwards. Without any control for confounding, NSAID use appears to be associated with a 32 percent (95 percent CI: 29, 34) reduction in mortality risk (table 3). There is no known biologic reason to expect that NSAID use would cause a reduction in the risk of death. Instead, the observed association is likely to be due to selection bias—that is, the fact that physicians are more likely to treat symptomatic pain with narcotic agents rather than NSAIDs in patients who are moribund.

Control for observed confounding

Table 3 also describes the association between NSAID use and 1-year mortality from Cox proportional hazards models using various approaches to control for observed confounding. Controlling for age and gender in the “traditional” outcome model, we observe a risk reduction that is somewhat closer to the null (26 percent; 95 percent CI: 23, 29) in comparison with the unadjusted result. Controlling for all variables presented in table 1 in the outcome model, we observe a risk reduction of 20 percent (95 percent CI: 17, 23). Essentially the same amount of risk reduction (19 percent; 95 percent CI: 16, 22) is observed when PS methods are applied (i.e., modeling the incidence of disease as a function of the estimated probability of exposure as a continuous variable and unique confounder together with the exposure).

Propensity score calibration

Implementation. We implemented the PSC approach by using regression calibration to correct for measurement error in the PS of the main study. The SAS macro “%blinplus” uses the validation study containing data on the two estimated PS’s (error-prone and gold standard) as well as the parameter estimates, their standard errors, and covariances from the Cox proportional hazards model in the main study (i.e., adjusted for the error-prone PS; see appendix 1). The macro output provides the adjusted relative hazard rate estimates, including 95 percent confidence intervals that are adjusted for additional uncertainty from the estimation of the error model in the validation study. We present the corresponding programming steps and a sample output in appendix 1.

Application using MCBS data for the external validation study. Using the MCBS as an external validation study, the application of the PSC approach results in estimates for the association between NSAID use and 1-year mortality that are closer to the null and even slightly beyond the null (table 3). When we use the “unmeasured” confounding model including the claims data variables plus interview data as the gold standard, the risk reduction with NSAIDs...
diminishes to just 8 percent (95 percent CI: 4, 12). Finally, when we use the self-reports of arthritis in addition to the claims data diagnostic code in the gold standard model to address confounding due to unmeasured and imperfectly measured covariates, NSAID use is associated with a 6 percent increase (95 percent CI: 0, 12) in mortality risk.

**TABLE 2. Propensity of use of nonsteroidal antiinflammatory drugs in the main study population and the external validation study ("error-prone" and "gold standard")**

<table>
<thead>
<tr>
<th></th>
<th>Main study</th>
<th></th>
<th>External validation study</th>
<th></th>
<th></th>
<th>Unmeasured</th>
<th>Un- and imperfectly measured</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR</td>
<td>95% CI</td>
<td>OR</td>
<td>95% CI</td>
<td>OR</td>
<td>95% CI</td>
<td>OR</td>
</tr>
<tr>
<td>Age (per year)</td>
<td>0.98</td>
<td>0.98, 0.99</td>
<td>0.98</td>
<td>0.97, 1.00</td>
<td>0.99</td>
<td>0.98, 1.00</td>
<td>0.98</td>
</tr>
<tr>
<td>Female gender</td>
<td>1.2</td>
<td>1.2, 1.3</td>
<td>1.2</td>
<td>1.0, 1.5</td>
<td>1.2</td>
<td>1.0, 1.5</td>
<td>1.1</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>1.6</td>
<td>1.5, 1.7</td>
<td>1.4</td>
<td>1.1, 1.9</td>
<td>1.3</td>
<td>1.0, 1.7</td>
<td>1.2</td>
</tr>
<tr>
<td>Other</td>
<td>2.0</td>
<td>1.9, 2.2</td>
<td>1.5</td>
<td>1.0, 2.2</td>
<td>1.5</td>
<td>1.0, 2.3</td>
<td>1.6</td>
</tr>
<tr>
<td>Diagnoses based on claims data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>0.9</td>
<td>0.8, 0.9</td>
<td>1.1</td>
<td>0.8, 1.5</td>
<td>1.1</td>
<td>0.8, 1.6</td>
<td>1.1</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>0.9</td>
<td>0.9, 1.0</td>
<td>0.9</td>
<td>0.6, 1.3</td>
<td>0.8</td>
<td>0.5, 1.2</td>
<td>0.8</td>
</tr>
<tr>
<td>Diabetes</td>
<td>1.0</td>
<td>1.0, 1.0</td>
<td>0.9</td>
<td>0.6, 1.3</td>
<td>0.7</td>
<td>0.5, 1.0</td>
<td>0.7</td>
</tr>
<tr>
<td>Cancer</td>
<td>0.8</td>
<td>0.8, 0.8</td>
<td>0.6</td>
<td>0.4, 0.9</td>
<td>0.6</td>
<td>0.4, 1.0</td>
<td>0.6</td>
</tr>
<tr>
<td>Arthritis (RA† or OA‡)</td>
<td>2.1</td>
<td>2.0, 2.2</td>
<td>2.4</td>
<td>1.7, 3.4</td>
<td>2.1</td>
<td>1.5, 3.0</td>
<td>1.8</td>
</tr>
<tr>
<td>Diagnoses based on self-reports</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arthritis (RA or OA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.1</td>
<td>3.1, 5.5</td>
<td></td>
</tr>
<tr>
<td>Use of the health-care system†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of physician visits</td>
<td>1.3</td>
<td>1.3, 1.3</td>
<td>1.1</td>
<td>1.0, 1.4</td>
<td>1.1</td>
<td>0.9, 1.2</td>
<td>1.1</td>
</tr>
<tr>
<td>No. of hospitalizations</td>
<td>0.9</td>
<td>0.9, 0.9</td>
<td>1.1</td>
<td>0.9, 1.3</td>
<td>1.0</td>
<td>0.9, 1.2</td>
<td>1.0</td>
</tr>
<tr>
<td>Use of medications</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thiazides</td>
<td>1.3</td>
<td>1.2, 1.3</td>
<td>1.6</td>
<td>0.9, 2.5</td>
<td>1.5</td>
<td>0.9, 2.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Steroids</td>
<td>1.0</td>
<td>0.9, 1.0</td>
<td>1.5</td>
<td>1.2, 2.0</td>
<td>1.4</td>
<td>1.1, 1.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Anticoagulants</td>
<td>0.5</td>
<td>0.5, 0.6</td>
<td>0.5</td>
<td>0.3, 0.8</td>
<td>0.5</td>
<td>0.3, 0.7</td>
<td>0.5</td>
</tr>
<tr>
<td>Body mass index§ (per kg/m²)</td>
<td>1.06</td>
<td>1.04, 1.07</td>
<td>1.05</td>
<td>1.03, 1.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education¶</td>
<td>1.0</td>
<td>0.8, 1.1</td>
<td>1.0</td>
<td>0.8, 1.1</td>
<td>1.0</td>
<td>0.8, 1.1</td>
<td>1.0</td>
</tr>
<tr>
<td>Income#</td>
<td>1.1</td>
<td>1.0, 1.2</td>
<td>1.1</td>
<td>1.0, 1.2</td>
<td>1.1</td>
<td>1.0, 1.2</td>
<td>1.1</td>
</tr>
<tr>
<td>Smoking</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current smoking</td>
<td>1.0</td>
<td>0.7, 1.4</td>
<td>1.0</td>
<td>0.7, 1.4</td>
<td>1.0</td>
<td>0.7, 1.4</td>
<td>1.0</td>
</tr>
<tr>
<td>Past smoking</td>
<td>1.1</td>
<td>0.9, 1.3</td>
<td>1.1</td>
<td>0.9, 1.3</td>
<td>1.1</td>
<td>0.9, 1.3</td>
<td>1.1</td>
</tr>
<tr>
<td>Activities of daily living**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulties with</td>
<td>1.3</td>
<td>1.2, 1.4</td>
<td>1.2</td>
<td>1.1, 1.3</td>
<td>1.2</td>
<td>1.1, 1.3</td>
<td>1.2</td>
</tr>
<tr>
<td>Inability to perform</td>
<td>1.2</td>
<td>1.1, 1.4</td>
<td>1.2</td>
<td>1.0, 1.3</td>
<td>1.2</td>
<td>1.0, 1.3</td>
<td>1.2</td>
</tr>
<tr>
<td>Area under the receiver operating characteristic curve</td>
<td>0.63</td>
<td>0.60</td>
<td>0.66</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Odds ratios and 95% confidence intervals from four separate multivariable logistic regression models including all variables for which estimates are presented.
† OR, odds ratio; CI, confidence interval; RA, rheumatoid arthritis; OA, osteoarthritis.
‡ Per increase in category of number of physician visits (0–5, 6–11, ≥12) or hospitalizations (0, 1, ≥2).
§ Weight (kg)/height (m)².
¶ Per increase in category of education (less than high school, high school, associate degree, or more).
# Per increase in category of annual income ($<10,000, >$10,000–$20,000, >$20,000–$40,000, >$40,000).
** Number of activities of daily living that were reported to be difficult and impossible to perform, respectively.

Am J Epidemiol 2005;162:279–289
TABLE 3. Association between use of nonsteroidal antiinflammatory drugs and 1-year mortality in a population-based cohort of 103,133 elderly persons

<table>
<thead>
<tr>
<th>Model</th>
<th>Hazard ratio</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted</td>
<td>0.68</td>
<td>0.66, 0.71</td>
</tr>
<tr>
<td>“Traditional” multivariate outcome model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age- and gender-adjusted</td>
<td>0.74</td>
<td>0.71, 0.77</td>
</tr>
<tr>
<td>Fully adjusted†</td>
<td>0.80</td>
<td>0.77, 0.83</td>
</tr>
<tr>
<td>Propensity score (main study)-adjusted</td>
<td>0.81</td>
<td>0.78, 0.84</td>
</tr>
<tr>
<td>PSC-adjusted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unmeasured confounding</td>
<td>0.92</td>
<td>0.88, 0.96</td>
</tr>
<tr>
<td>Un- and imperfectly measured confounding</td>
<td>1.06</td>
<td>1.00, 1.12</td>
</tr>
</tbody>
</table>

* Propensity score calibration (PSC) adjustment was based on data from 5,108 participants in the Medicare Current Beneficiary Survey, used as an external cross-sectional validation study. Hazard ratios and 95% confidence intervals were estimated using Cox proportional hazards regression. PSC-adjusted models include uncertainty due to the estimation of the error model.

† Adjusted for all variables used in the estimation of the propensity score in the main study (i.e., all variables presented in table 1).

Simulation study. To assess the performance of PSC, we conducted a simulation study. Although simulating a confounder requires information on the outcome, PSC simulation results reported ignore this hypothetical outcome and therefore apply to the external validation study design. Starting from the observed cohort, we simulated 1,000 studies of 103,133 participants each, adding a single dichotomous confounder that was parameterized to be inversely associated with the exposure and a risk factor for mortality. For this simulated confounder, we assumed the following expected prevalences: unexposed and alive—0.5; unexposed and dead—0.75; exposed and alive—0.25; and exposed and dead—0.5. These parameters resulted in an overall prevalence of the confounder of 51 percent, an exposure-confounder odds ratio of 0.33, and a hazard ratio for mortality (independent of the exposure) of 2.7. In each study, we then randomly sampled participants into an internal validation study with a selection probability of 5 percent. For confounding, including the simulated confounder, by adjusting for the estimated gold standard PS in the Cox proportional hazards model to obtain an estimate of the unobservable truth. Then we applied PSC using only the 5 percent internal validation study to estimate the error in the error-prone PS and to apply regression calibration to the whole cohort estimates based on the error-prone PS. The median estimate using PSC (hazard ratio = 1.04; see table 4) was virtually identical to the median estimate from the unobservable truth (hazard ratio = 1.03). The increased width of the empirical 95 percent confidence interval of the PSC reflects the additional uncertainty introduced by the 5 percent validation study. The asymmetry pointing towards higher values indicates a slight tendency towards overadjustment.

TABLE 4. Association between use of nonsteroidal antiinflammatory drugs (NSAIDs) and 1-year mortality in a population-based cohort of 103,133 elderly persons: results from a simulation study of propensity score calibration (PSC)

* Hazard ratios and 95% empirical confidence intervals for the association between NSAID use and 1-year mortality from 1,000 simulations, estimated using Cox proportional hazards regression.
† Assumed “truth” including all observed variables and the simulated confounder “unobserved” in the whole cohort in the estimation of the propensity score used to control for confounding; confounder simulated as a dichotomous risk factor for mortality (hazard ratio = 2.7) inversely associated with NSAID use (odds ratio = 0.33) and a marginal prevalence of 51%.
‡ Using observed information on the simulated confounder from the 5% validation study only.

DISCUSSION

To our knowledge, we have combined for the first time two existing epidemiologic methods, PS’s and regression calibration, to adjust for unmeasured confounding in cohort studies using validation data. By taking the joint distribution of unmeasured confounders into account, the PSC extends prior work using sensitivity analyses and adjustments based on single confounders (1–8). In our example, this novel approach appeared effective in adjusting for unmeasured confounding by virtually eliminating the presumably spurious “protective” effect of NSAIDs on mortality. In addition, the amount of residual confounding due to unmeasured and poorly measured covariates was important enough to qualitatively change the likely spurious association between NSAID use and all-cause mortality. The proposed strategy is easy to conceptualize and implement using user-friendly, well-documented, and available software.

PSC needs to be compared with other methods to address unobserved confounding in nonexperimental research (1–8). Others (e.g., Schneeweiss et al. (8)) have proposed sensitivity analyses addressing the effect of several confounders by using external information on the relations of unobserved confounders with the exposure and disease of interest. To our knowledge, none of these methods can sufficiently address the joint effect of several unobserved confounders. Another important advantage of PSC is that it is not dependent on outcome information in the validation study; that is, it can be applied using internal or external (i.e., cross-sectional) validation studies without the need to specify the confounder-disease association.

Like regression calibration, PSC is also dependent on the assumption that the error-prone variable is a surrogate for the gold standard variable—that is, that the error-prone PS is independent of disease given the gold standard PS and exposure (26, 27) (see appendix 2). This strong assumption cannot be tested without information on the outcome (e.g.,
in an external validation study). If this assumption is violated, PSC can be far less useful and even counterproductive, in that it can increase rather than decrease bias.

Unfortunately, little is known about components of PS’s and their interaction with respect to the surrogacy assumption. Omitting a single covariate from a PS can clearly lead to a violation of surrogacy, but outside knowledge on the association of the covariate with disease could be used to detect such a violation. In the setting of multiple covariates, such an assessment might not be possible, but the surrogacy assumption might be more plausible if an underlying framework for confounding (e.g., frailty) can be put forward and the covariates observed in the main study are assumed to be an unbiased selection of the covariates observed in the validation study. Intuitively, one might hope that a more refined PS, based on additional covariates plus alternative measures with less error, will contain all of the relevant information on propensity of exposure captured in an error-prone PS.

If information on the outcome is available in the validation study (e.g., in an internal validation study), one can check the independence of the error-prone PS and disease given the gold standard PS and exposure. If this is the case, the assumption may be considered valid in the context of the main study as well. In such a setting, unobserved confounding in the main study can also be seen as a missing covariate problem, and methods of addressing missing data, including multiple imputation (28) and Bayesian methods (29, 30), may be applied.

The design of validation studies has been extensively addressed (31–35), and these issues are likely to apply to PSC as well. External validation studies are limited by their lack of information on the outcome, as well as by limited comparability with the main study (26, 36). It is important to note that medication use was assessed by prescription claims data in the main study and by interview in the validation study. Ideally, medication use should be assessed by prescription data in the validation study as well, in order to exactly reproduce the error-prone PS in the validation study. Depending on the relative quality of these sources of information on medication use (37), the error in the error-prone PS might have been under- or overestimated somewhat in our example. The observed differences in the prevalence of medication use per se do not invalidate the approach (38, 39). The sizes of the main and validation studies will have a big impact on the precision of the resulting estimates (38, 40). Because the validation study was very large in our example, there was only a minor increase in the width of the confidence interval for the exposure of interest using PSC.

We present a single example to illustrate the applicability of the analytical strategy and to encourage further research into its validity and limitations in different settings. Many of the issues regarding measurement error have been addressed previously and are likely to apply to PSC as well. Accordingly, the proposed method is approximately unbiased, with the amount of residual bias depending on established theoretical and analytical properties of PS’s and regression calibration. However, the combination of these methods could lead to analysis of parameter constellations not previously assessed (17, 18, 41). Regression calibration approximations break down when the measurement error is large, that is, when the correlation between the estimated error-prone PS and the gold standard PS is weak. In our example, the correlations with the error-prone PS’s were moderate (0.56–0.68).

Our approach is easy to implement (see appendix 1). PS’s are increasingly used in medical research (42), although many issues concerning their optimal implementation are still unresolved. Regression calibration is the approach most widely applied to address measurement error in multivariate epidemiologic analyses. Its advantages include the ability to adjust for confounding in the primary regression by continuous, discrete, and ordinal variables, which are assumed to be perfectly measured, and the availability of an easy-to-implement macro (see appendix 1) providing adjusted effect estimates and standard errors for a variety of multivariate models.

We chose the example of NSAIDs and all-cause mortality, since NSAID use is unlikely to lead to reduced mortality in this elderly population (43). Glynn et al. (43) argued that in an elderly population, selected classes of drugs, including NSAIDs, are more likely to be prescribed to healthier subjects less close to death. NSAID use in the elderly is associated with several adverse outcomes, including increased risk of gastrointestinal hemorrhage (44–47), impaired kidney function (48–50), and hypertension (51, 52). Therefore, no association with mortality or a slightly increased risk of mortality seems biologically more plausible than any amount of reduced risk.

We used the PS as a continuous variable in our disease and error models; this depends on the assumption of a (log-)linear association with the dependent variable. This assumption can be tested (e.g., by inclusion of categories or quadratic and cubic terms) in the measurement error model (equation 5), and it was essentially met in our study. The assumption that the gold standard PS has a linear association with the logit of disease given exposure (9) cannot be tested for the target disease model (equation 6). Since the gold standard PS captures all of the confounding in a single covariate, misspecification of its association with the outcome is likely to have pronounced effects on its ability to control for confounding (53).

We did not assess the role of variable selection procedures when estimating the PS. With large data sets, it might be best to err on the side of including apparently unimportant variables and also interactions between variables (12), a topic which is beyond the scope of this paper. Age was the only continuous variable in our main study data set, and results were virtually unchanged by including age in categories instead of age as a continuous variable in the PS models. The number of exposed persons in the validation data set will limit the number of variables that can be used to estimate the PS (and therefore the application of the method) in smaller studies (54, 55). The confidence intervals of PSC take the uncertainty in the estimation of the error model into account but not the uncertainty due to model misspecification or the use of different definitions of some of the covariates, including NSAID use. Therefore, confidence intervals should be interpreted only as a rough guide, and a minimum estimate, of the inherent uncertainty (56).

We conclude that the PSC approach, that is, the combination of PS’s and regression calibration, can substantially...
improve control for confounding by unmeasured and imperfectly measured confounders in cohort studies when internal or external validation data are available. The approach might be especially advantageous for pharmacoepidemiologic research based on claims data. Once the advantages and limitations of the approach have been assessed in a variety of settings and parameter constellations, it might help address one of the main problems in pharmacoepidemiologic research using large-scale claims databases—that is, the lack of information on important confounders—and could lead to improvement in the quality of research in this field. In the meantime, to avoid obtaining misleading results, investigators should not use this approach as a "black box" or a standard procedure. Until the validity and limitations of PSC have been assessed in different settings, the method should be seen rather as a sensitivity analysis.

ACKNOWLEDGMENTS

This project was funded by a grant (RO1 AG023178) from the National Institute on Aging. The authors thank Dr. Kenneth J. Rothman for helpful discussions and his valuable suggestions. Conflict of interest: none declared.

REFERENCES

APPENDIX 1

For notation, see methods: e = exposure of interest (e = 1(0) if the exposure of interest is present (absent)), x = error-prone propensity score, xgs = ‘gold-standard’ propensity score.

Program code

```sas
proc logistic data=main descending; * PS in main study *;
  model e = age sex black other etc.;
  output out=main predicted=x;
proc phreg data=main covout outset=bvarb;
  model pdays*death(0) = x e; * outcome model adjusting for error-prone PS *;
proc logistic data=val descending; * PS in validation study (error-prone) *;
  model e = age sex black other etc.;
  output out=val2 predicted=x;
proc logistic data=val2 descending; * PS in validation study (gold-standard) *;
  model e = age sex black other etc.
    bmi edu inc cs mk psmk adlssdiff adlunable; * covariates not observed in main study *;
```

Am J Epidemiol 2005;162:279–289
output out=val3 predicted=xgs;
%include ‘......\blinplus.sas’;
data wts;* weights needed for calibration *;
input name $ weight;
cards;
x 1.0
e 1.0
;
%blinplus (
   type = PHREG, /* type can be LOGISTIC, REG, or PHREG */
   valid = val3, /* Dataset with Validation Study data */
   main_est = bvarb, /* Dataset containing est.s from main study regress. */
   weights = wts, /* Dataset with weights */
   err_var = x e, /* List of the variables measured with error */
   true_var = xgs e, /* List of the variables measured without error */
   depend = pdays, /* Name of the dependent variable */
   change = , /* Percent Change Scale: T/F (need type=REG for T) */
   internal = f, /* If validation study is Internal, do you wish to */
                /* combine estimates? T/F (if T, need to provide */
                /* specify val_est argument */
   val_est = /* Dataset containing Internal validation Est.s */
);

Program output

Main study regression coefficients: Uncorrected

<table>
<thead>
<tr>
<th>WT</th>
<th>B</th>
<th>SE</th>
<th>OR</th>
<th>95% CI</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1.00</td>
<td>-6.01139</td>
<td>0.12320</td>
<td>0.00245</td>
<td>0.00192 – 0.00312</td>
</tr>
<tr>
<td>e</td>
<td>1.00</td>
<td>0.21433</td>
<td>0.02010</td>
<td>0.80708</td>
<td>0.77590 – 0.83952</td>
</tr>
</tbody>
</table>

Main study regression coefficients: Corrected

<table>
<thead>
<tr>
<th>WT</th>
<th>B</th>
<th>SE</th>
<th>OR</th>
<th>95% CI</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1.00</td>
<td>-6.27872</td>
<td>0.19011</td>
<td>0.00188</td>
<td>0.00129 – 0.00272</td>
</tr>
<tr>
<td>e</td>
<td>1.00</td>
<td>0.05416</td>
<td>0.02923</td>
<td>1.05566</td>
<td>0.99687 – 1.11791</td>
</tr>
</tbody>
</table>

APPENDIX 2

Propensity score calibration can be proven in the case of a normally distributed outcome $Y$ according to Carroll et al. (26). Carroll and Stefanski (27) also provide extensions to generalized linear models that are not presented here for the sake of clarity.

Suppose we have measures of the continuous outcome $Y$, the dichotomous exposure of interest $E$, and a subset of covariates $X_{EP}$ available for the entire population, and the entire set of covariates $X_{GS}$ available for a subgroup. Our target model would be

$$E[Y|E,e(X_{GS})] = \beta_{0} + \beta_{E}E + \beta_{X}e(X_{GS}),$$

where $(X_{GS})$ is the gold standard (GS) propensity score.

According to the property of conditional expectation, the expected value of $Y$ given the error-prone (EP) propensity score $e(X_{EP})$ can be written as

$$E[Y|E,e(X_{EP})] = E\{E[Y|E,e(X_{GS}),e(X_{EP})]|E,e(X_{EP})\}.$$
Under the assumption that \( e(X_{EP}) \) is independent of the outcome given \( e(X_{GS}) \) and \( E \) (i.e., is a surrogate of \( e(X_{GS}) \)),

\[
E[Y|E,e(X_{EP})] = E\{\beta_0 + \beta_E E + \beta_X e(X_{GS})|E,e(X_{EP})\},
\]

which leads to

\[
E[Y|E,e(X_{EP})] = \beta_0 + \beta_E E + \beta_X e(X_{GS})|E,e(X_{EP})].
\]

Thus, regression calibration, that is, the single imputation \( E\{e(X_{GS})|E,e(X_{EP})\} \) to control for confounding of the exposure-disease association in the outcome model, is valid under the assumptions that disease is a linear function of \( e(X_{GS}) \) and \( E \) and that \( e(X_{EP}) \) is a surrogate for \( e(X_{GS}) \) (i.e., is independent of \( Y \) given \( e(X_{GS}) \) and \( E \)).