Practice of Epidemiology

Adjustment for Missing Confounders in Studies Based on Observational Databases: 2-Stage Calibration Combining Propensity Scores From Primary and Validation Data

Hui-Wen Lin* and Yi-Hau Chen*

* Correspondence to Dr. Hui-Wen Lin, Department of Mathematics, Soochow University, 70 Linhsi Road, Shihlin, Taipei 11102, Taiwan (e-mail: hwlin@scu.edu.tw); or Dr. Yi-Hau Chen, Institute of Statistical Science, Academia Sinica, 128 Section 2 Academia Road, Taipei 11529, Taiwan (e-mail: yhchen@stat.sinica.edu.tw).

Initially submitted January 13, 2014; accepted for publication April 29, 2014.

Bias caused by missing or incomplete information on confounding factors constitutes an important challenge in observational studies. The incorporation of external data on more detailed confounding information into the main study data may help remove confounding bias. This work was motivated by a study of the association between chronic obstructive pulmonary disease and herpes zoster. Analyses were based on administrative databases in which information on important confounders—cigarette smoking and alcohol consumption—was lacking. We consider adjusting for the confounding bias arising from missing confounders by incorporating a validation sample with data on smoking and alcohol consumption obtained from a small-scale National Health Interview Survey study. We propose a 2-stage calibration (TSC) method, which summarizes the confounding information through propensity scores and combines the analysis results from the main and the validation study samples, where the propensity score adjustment from the main sample is crude and that from the validation sample is more precise. Unlike the existing methods, the validity of the TSC approach does not rely on any specific measurement error model. When applying the TSC method to the motivating study above, the odds ratio of herpes zoster associated with chronic obstructive pulmonary disease is 1.91 (95% confidence interval: 1.62, 2.26) after adjustment for cumulative smoking and alcohol consumption.

calibration; chronic obstructive pulmonary disease; herpes zoster; missing confounders; propensity score

Abbreviations: COPD, chronic obstructive pulmonary disease; HZ, herpes zoster; LHID, Longitudinal Health Insurance Database; NHIS, National Health Interview Survey; RC, regression calibration; TSC, 2-stage calibration.

Adjustment for confounding variables is a key step toward establishing causal inference on the treatment/exposure effects in observational studies. In practice, however, precise data on potential confounding variables may not be easy to obtain, leading to the confounding bias problem caused by missing confounding variables. Although the causal treatment/exposure effects may be assessed through sensitivity analysis (1–3) when the confounding variables are not observed, such a strategy is most applicable only when the missing confounding variables are categorical and of few dimensions, because it is, in general, prohibitive to enumerate all possible set-ups for confounding variables to which the sensitivity of the analysis is to be assessed.

In certain situations, more detailed confounding information may be ascertained for another set of study subjects external to the main study. Such supplemental information can thus be used as “validation data” to correct for confounding bias resulting from incomplete confounder information in the main study data. This strategy is especially useful when the main study is based on data extracted from large-scale observational databases such as administrative health databases or electronic health record systems. For example, Stürmer et al. (4) considered adjusting for the bias of estimation from a large health care database, from which only crude confounding information was available, by using precise confounding information obtained from a small-scale survey study and then...
applying the regression calibration (RC) method, a widely used method for correcting measurement error bias (5). Under a similar setting, McCandless et al. (6) proposed to incorporate external confounding information through propensity scores using an elegant Bayesian framework.

The success of the above methods relies on assumptions regarding the relationship between the actual and crude confounding information obtained from the validation and the main study data, respectively. For example, the RC method of Stürmer et al. (4) hinges crucially on the linear measurement model, assuming a linear relationship between the actual and crude confounding information (in terms of propensity score). It also requires the surrogacy assumption that the measurement errors are nondifferential (i.e., independent of the outcome variable given the true underlying confounders (7)). The method of McCandless et al. (6) does not require the surrogacy assumption, but still needs to assume the true and crude confounders to be independent of each other or to marginally follow some posited relationship. Violation of these measurement error assumptions, which is not uncommon in practice, can result in bias in any direction for the treatment/exposure effects (8).

In this study, we focus on the study design similar to that considered by Stürmer et al. (4) and McCandless et al. (6). Specifically, we consider the study consisting of a large-scale main study with data on outcome and treatment/exposure status and with incomplete confounding information, as well as a small-scale validation study with data on complete confounding information, in addition to the variables already available in the main study. We examine an analysis strategy to combine data from the main and validation studies, namely the 2-stage calibration (TSC) method first proposed by Chen and Chen (9). The TSC method requires that the validation data are gathered from a sample, possibly outside the main study, that is representative of the main study population. This is, in fact, also an assumption required in the methods of Stürmer et al. (4) and McCandless et al. (6). The main advantage of TSC is that its validity relies only on the representative assumption mentioned above and not on any specific measurement error model. Therefore, TSC is more widely applicable than the existing methods.

This study was motivated by a previous study conducted by Yang et al. (10) on the association between chronic obstructive pulmonary disease (COPD) and herpes zoster (HZ). COPD is an autoimmune disease with chronic systemic inflammation involving more than just airways and lungs. Given that various immune-mediated diseases, such as rheumatoid arthritis and inflammatory bowel disease, have been reported to be associated with an increased risk of HZ, it is reasonable to hypothesize that the immune dysregulation found in COPD may put patients at higher risk of developing HZ (10). However, smoking and alcohol consumption may provoke inflammation in the lung. In a study by Yang et al. (10), analysis was based on data from 42,430 subjects from an administrative database in which information about cigarette smoking and alcohol consumption was lacking, although information on other confounders, including age, sex, diabetes mellitus, hypertension, coronary artery disease, chronic liver disease, autoimmune disease, and cancer, was available. To further adjust for missing confounders (i.e., cigarette smoking and alcohol consumption) in this study, we selected a validation sample of 1,148 subjects who were comparable to the main study sample from a survey study, which contained data on subjects’ COPD and HZ status and all the confounders to be adjusted for, including smoking and alcohol consumption. We then applied the TSC method to combine the results from the main and validation studies to yield an estimate for the association between COPD and HZ corrected for confounding bias in the main study because of failure to adjust for smoking and alcohol consumption. Similarly to Stürmer et al. (4), we adopted the propensity score method (11, 12) to accommodate the high-dimensional confounding variables. In addition to this application examining the association between COPD and HZ adjusted for smoking and alcohol consumption, we also performed a series of simulation studies to examine the performance of the TSC method.

METHODS

Study populations

The study motivating this research consists of a large-scale main study lacking precise confounding information (i.e., smoking and alcohol consumption) and a small-scale validation study performed on the same population as the main study and collecting data on precise confounding information, in addition to variables collected in the main study.

**Main study population.** The 2005 Longitudinal Health Insurance Database (LHID) contains all medical claims data for 1 million beneficiaries randomly sampled from 25.68 million enrollees in Taiwan. As in the study by Yang et al. (10), we identified from the LHID database 8,486 patients with diagnosis of COPD between January 1, 2004, and December 31, 2006, and without diagnosis of HZ before 2004. These subjects served as the exposed cohort (to COPD) in our main study. The unexposed cohort consisted of 33,944 subjects randomly selected from the remaining subjects in LHID, who had no diagnosis of HZ before 2004 and no diagnosis of COPD between 2004 and 2006. The 33,944 subjects were age- and sex-matched (4 controls for every patient) to the exposed cohort. All subjects in the main study were followed from the date of cohort entry (January 1, 2004) until they developed HZ or December 31, 2006, whichever was earlier. The aim was to assess the exposure effect of COPD on the development of HZ. Information on confounders collected in the main study included age (in years), sex (male/female), diabetes mellitus (yes/no), hypertension (yes/no), coronary artery disease (yes/no), chronic liver disease (yes/no), autoimmune disease (yes/no), and cancer (yes/no). Descriptive statistics for these variables are shown in Table 1.

**Validation study population.** The 2005 National Health Interview Survey (NHIS) was conducted by the National Health Research Institute and the Bureau of Health Promotion in Taiwan to gather data on medical claims, health behaviors, and quality of life for approximately 26,658 Taiwan residents in 2005. Using the same inclusion criteria as in the main study for both the COPD and non-COPD cohorts, we identified 244 patients with COPD and 904 age- and sex-matched non-COPD subjects (approximately 4 non-COPD
subjects for each COPD patient) from the NHIS database. All subjects in the validation study sample were followed from the date of cohort entry (January 1, 2004) until they developed HZ or December 31, 2006, whichever was earlier. Confounder variables collected in the validation study include all those already considered in the main study, as well as information on cumulative smoking and alcohol consumption, which are important confounders that were missing in the main study. The 2 variables were measured by questionnaire with ordinal categories as shown in Table 1. In later analysis, cumulative smoking was coded as a continuous variable ranging from 0 to 3 (0: no cigarettes, 1: few cigarettes, 2: less than 100 cigarettes, 3: more than 100 cigarettes). Similarly, the variable for alcohol consumption was continuous, ranging from 0 to 5 (0: never/seldom, 1: less than once a month, 2: twice a month, 3: weekly, 4: once every 2–3 days, 5: once every day).

### Propensity score

Similarly to Stürmer et al. (4), we obtained propensity scores $PS_C$ and $PS_P$ in the main and validation studies, respectively, to summarize the confounding information collected in the main and validation studies. $PS_C = \Pr(T = 1|C)$ is the crude propensity score based on the incomplete confounding information $C$, where $T = 1$ or 0 denotes the presence or absence of exposure (i.e., COPD), and $C$ is the vector of the covariates observed in both the main and validation studies (i.e., age, sex, diabetes mellitus, hypertension, coronary artery disease, chronic liver disease, rheumatic disease, and cancer); $PS_C$ can be obtained by fitting a multivariate logistic regression model to data in the main study. The $PS_P = \Pr(T = 1|C, U)$ is the precise propensity score based on the complete confounding information $C$ and $U$, where $U$ is the vector for covariates (smoking and alcohol consumption) observed additionally in the validation study but missing in the main study, and $C$ is defined as above; $PS_P$ can be obtained by fitting a multivariate logistic regression model to data in the validation study.

### 2-Stage calibration

Extending the idea of Chen and Chen (9) to the current setting, we propose a TSC method to combine analysis results from the main and validation studies and to correct for the confounding bias in the main study. In the following, we use the notation defined in the above subsection. In addition, $D = 1$ or 0 denotes the presence or absence of the disease outcome (i.e., HZ), and $n_M$ and $n_V$ are the sizes of the main and validation study samples, respectively.

In the first stage, data on variables $(D,T,C)$ that are commonly observed in the main and validation study samples are pooled into a single “first-stage” sample of size $N = n_M + n_V$, and the following logistic regression is fitted to the first-stage sample to assess the association between COPD and HZ, adjusting for the crude propensity score on the basis of the incomplete confounding information $C$:

$$\log \left( \frac{\Pr(D = 1|T, C)}{\Pr(D = 0|T, C)} \right) = \delta + \gamma T + \theta f(PS_C),$$

where $PS_C$ is the propensity score obtained from the logistic regression fit of $T$ to $C$ in the first-stage sample, and $f$ is a set of suitable transformation functions for $PS_C$. We consider 2 choices of $f$: the identity function (i.e., $f(PS_C) = PS_C$) and the natural spline functions (6, 13). The parameter $\theta$ is the coefficient for $f(PS_C)$, and a prime denotes transposition of the vector. The estimate $\gamma$ of $\gamma$ provides an estimate for the
association (i.e., log odds ratio) between COPD and HZ. However, because of incompleteness of confounding information, $\gamma$ is subject to residual confounding bias.

In the second stage, the logistic regression (equation 1) is fitted again but to the validation study sample, and the estimate $\hat{\gamma}$ of $\gamma$ is obtained. Additionally, the following logistic regression is fitted to data on $(D, T, C, U)$ in the validation study sample:

$$\log \left( \frac{P(D = 1|T, C, U)}{P(D = 0|T, C, U)} \right) = \alpha + \beta T + g(PS_U),$$

where $PS_U$ is the propensity score obtained from the logistic regression fit of $T$ to $(C, U)$ in the validation study sample, and $g$ is a set of transformation functions. Again, we consider $g$ to be an identity function or a set of natural spline functions (6, 13). The estimate $\hat{\beta}$ of $\beta$ provides an estimate for the association between COPD and HZ, which is free of confounding bias assuming complete confounding information has been incorporated in $(C, U)$.

Although $\hat{\beta}$ is a valid estimate for the association between COPD and HZ, its construct is based simply on the validation sample (6, 13). The estimate $\hat{\beta}$ can be based on the Wald test statistic where

$$PSP = \frac{\hat{\beta} - \lambda \sqrt{\hat{\gamma} - \bar{\gamma}}}{\sqrt{\hat{\gamma} - \bar{\gamma}}}$$

where $\lambda$ is the covariance between $\hat{\beta}$ and $(\hat{\gamma} - \bar{\gamma})$, and $\nu$ is the variance of $(\hat{\gamma} - \bar{\gamma})$. In Web Appendix 1, available at http://aje.oxfordjournals.org/, we provide formulae for estimating these variance/covariance quantities, which are based on the well-known sandwich-type estimation of variance/covariance.

Assuming that the validation study sample is comparable to the main study sample, such that the estimates $\hat{\gamma}$ and $\bar{\gamma}$ for the association parameter $\gamma$ have the same limiting value (as sample sizes tend to infinity), then the TSC estimate $\bar{\beta}$ will have the same limiting value as the validation estimate $\hat{\beta}$, namely the true association parameter value $\beta$ in the model adjusted for the precise propensity score (equation 2). In fact, $\bar{\beta}$ is asymptotically normal with mean equal to the true $\beta$ value and variance given by

$$\text{var}(\bar{\beta}) = \frac{\lambda^2}{\nu},$$

which can be estimated by simply plugging estimates for $\text{var}(\hat{\beta})$, $\lambda$, and $\nu$ (Web Appendix 1). When using information from both the main and validation studies, the TSC estimate $\bar{\beta}$ has smaller variance and, hence, is more efficient than the validation sample–only estimate $\hat{\beta}$. The confidence interval and hypothesis testing regarding the association parameter $\beta$ can thus be performed according to this asymptotic normality theory. For example, the testing of the null hypothesis $H_0$: $\beta = 0$ can be based on the Wald test statistic $\frac{\hat{\beta}^2}{\text{var}(\hat{\beta})}$, which is asymptotically $1 - df \chi^2$ distribution under $H_0$. Inference based on $\beta$ will be more efficient and powerful than that based on $\bar{\beta}$ and less biased than that based on $\bar{\gamma}$, the first-stage estimate without adjustment for complete confounding information. These properties will be confirmed by a series simulation studies presented below.

**Extensions and summary**

In many situations, such as for rare diseases, the case-control sampling design is easier to implement than the prospective sampling design. Suppose that the main and/or validation study samples are drawn by case-control sampling from comparable populations, and the estimates $\hat{\gamma}$, $\bar{\gamma}$, $\hat{\beta}$, and $\bar{\beta}$ are obtained from the main and validation samples as above by logistic regression models. Then, by the well-known result from the equivalence of prospective and retrospective logistic regression analysis (14), we can see that the validity of the TSC analysis remains.

The TSC analysis presented above for binary outcomes under logistic regression can be readily extended to other types of outcomes, including continuous, count, and time-to-event outcomes, as well as to general regression models such as the generalized linear model and the Cox proportional hazards model (9, 15). For the association parameter estimates $\hat{\gamma}$, $\bar{\gamma}$, and $\hat{\beta}$ obtained by general regression models with data from validation and/or pooled samples, we still calculate the TSC estimate $\bar{\beta}$ using equation 3, with the parameters $\lambda$ and $\nu$ obtained by the general formulae given in Web Appendix 1. Also, the extension to multivariate exposure is straightforward, as shown in Web Appendix 1.

In summary, the TSC procedure amounts to the following general algorithm:

1. **Step 1**: estimate the treatment effect $\gamma$, adjusted for crude propensity scores, by $\bar{\gamma}$ in the first-stage (primary + validation) data
2. **Step 2**: estimate $\gamma$ by $\hat{\gamma}$ with the validation data
3. **Step 3**: estimate the treatment effect $\hat{\beta}$, adjusted for precise propensity scores, by $\bar{\beta}$ in the validation data
4. **Step 4**: obtain the TSC estimate $\bar{\beta}$ by equation 3 (or more generally, equation 3 in Web Appendix 1) and its variance by equation 4 (or more generally, equation 4 in Web Appendix 1)

Detailed mathematical expressions are given in Web Appendix 1. The TSC analysis for linear, logistic, and Cox regression models has been implemented in SAS software (SAS Institute, Inc., Cary, North Carolina) macros that can be obtained from the second author’s website (http://www.stat.sinica.edu.tw/yhchen/download.htm).

**Simulation study**

We perform simulations to examine finite sample performance of the proposed method, and to compare the proposed method with the propensity score RC method by Stürmer et al. (4). All of the simulations are based on 1,000 replications.

In the simulations, the exposure of interest, $T$, and the outcome of interest, $D$, are both dichotomous variables. The crude and precise confounder variables $C = (C_1, C_2)$ and $U$ are all independent standard normal random variables. The conditional probability that $T = 1$ given $(C_1, C_2, U)$, which

corresponds to the propensity score \(PS_p\), is given by
\[
PS_p = Pr(T = 1|C_1, C_2, U) = \frac{\exp(b_0 + b_1C_1 + b_2C_2 + b_3U)}{1 + \exp(b_0 + b_1C_1 + b_2C_2 + b_3U)}.
\]

The conditional probability that \(D = 1\) given \(T\) and \((C_1, C_2, U)\) is given by
\[
Pr(D = 1|T, C_1, C_2, U) = Pr(D = 1|T, PS_p) = \frac{\exp(\alpha + \beta T + \theta PS_p)}{1 + \exp(\alpha + \beta T + \theta PS_p)}.
\]

We also consider a second disease model to allow violation of the propensity score assumption made in equation 6, where the conditional probability for \(D = 1\) given \(T, C_1, C_2, U\) is given by
\[
Pr(D = 1|T, C_1, C_2, U) = \frac{\exp(\alpha + \beta T + \theta_1C_1 + \theta_2C_2 + \theta_3U + \theta_4C_1U)}{1 + \exp(\alpha + \beta T + \theta_1C_1 + \theta_2C_2 + \theta_3U + \theta_4C_1U)}.
\]

Data on \((D, T, C_1, C_2, U)\) are generated under prospective sampling of \(n_M\) subjects in the main study and \(n_V\) subjects in the validation study. In analysis, data on \(U\) in the main study are omitted.

Throughout the simulations, \(n_M = 10,000\), and \(n_V = 1,000\). In equation 5, \(b_0 = 0.5, b_1 = b_2 = 0.405\), and \(OR_{TU} = \exp(b_3)\) is varied. In the first disease model (equation 6), \(\theta = -6\). In the second disease model (equation 7), \(\theta_1 = 0.3, \theta_2 = -0.1, \theta_3 = 0.6,\) and \(OR_{MU} = \exp(\theta_4)\) is varied. In both disease models, the intercept parameter \(\alpha\) is set such that \(P_D = Pr(D = 1) = 0.05\). We first perform simulations under the null scenario where there is no association between \(T\) and \(D\) (i.e., \(\beta = 0\)). We then consider the simulation scenario where there exists association between \(T\) and \(D\), such that \(\beta = 0.4\) or \(-0.4\). When applying TSC, no matter whether the true disease model is given by equation 6 or equation 7, \(\gamma\) and \(\beta\) are always estimated with the simulated data from the working models given in equations 1 and 2 with the identity or natural cubic spline transformation for propensity scores. The crude and precise propensity scores \(PS_C\) and \(PS_p\) used in these models are estimated by the logistic regression fit of \(T\) to \((C_1, C_2)\) and \((C_1, C_2, U)\), respectively. We compare results from the first-stage analysis adjusting crude propensity score \(PS_C\), the second-stage analysis adjusting precise propensity score \(PS_p\), the TSC, and the propensity score RC by Stürmer et al. (4).

**RESULTS**

**Simulation results**

Table 1 shows simulation results when the disease model is given in equation 6 for the null scenario (no association between \(T\) and \(D, \beta = 0\)) and the alternative scenario (nonnull association between \(T\) and \(D, \beta = 0.4\)). Results of different estimates of \(\hat{\beta}\) in equation 2 under identity transformation of \(PS_p\) are shown, including the estimate obtained from the first-stage (pooled) sample adjusting the crude propensity score \(PS_C\) (5). The estimate from the second-stage (validation) sample adjusting the precise propensity score \(PS_p\) \(\hat{\beta}\), the TSC estimate \(\hat{\beta}_T\), and the RC estimate \(\hat{\beta}_{RC}\). We can see from Table 1 that the TSC method provides accurate estimation and correct type I error rate close to the nominal value of 0.05 under all simulation set-ups. In contrast, the first-stage estimate \(\hat{\beta}_T\) using incomplete confounder information is biased, and the associated type I error rate is out of control; the bias is more serious when the association between \(T\) and \(U\) \((OR_{TU})\) is stronger. Similarly, the RC method also has inflated type I error, except in the trivial case where \(T\) is independent of \(U\) \((OR_{TU} = 1)\). The estimate \(\hat{\beta}\) based on the validation (second-stage) sample is unbiased, and the association test based on it has correct type I error rate. However, \(\beta\) has larger standard errors than the TSC estimate. Table 1 also shows the power of the Wald test for testing the disease-exposure association when such association indeed exists under the disease model given in equation 6. It is seen that the power of the TSC method is much higher than that of the validation sample–only method adjusting \(PS_p\). The power advantage of the TSC method is owing to the smaller standard error of \(\hat{\beta}\) when compared with the validation sample–only estimate \(\hat{\beta}\). The first-stage and RC methods can have high power. However, they are subject to uncontrolled type I error rate, as we have noted in the simulation under the null scenario.

Tables 2 and 3 show simulation results comparing the performance of the validation sample–only method adjusting the \(PS_p\), the TSC, and the RC methods under the null (Table 2) and alternative (Table 3) hypotheses of the disease-exposure association, respectively, when the underlying disease model is given in equation 7. The association parameters \(\gamma\) and \(\beta\) are estimated by working disease models (equations 1 and 2). Unless \(T\) is unrelated to \(U\) \((OR_{TU} = 1)\), the type I error rate of the RC method is beyond the nominal level of 0.05, because the linear measurement error model assumption, a requirement for the RC method, is no longer satisfied in such settings. In contrast, the TSC method still provides good control of type I error rate in these settings because it is free of any measurement error model assumption. This is true even when TSC is performed with the working disease models 1 and 2 with identity transformation of the propensity scores, which are obviously wrong disease models here. We can see from Table 3 that the power of the TSC approach, based on either propensity scores or their natural cubic spline functions, is much higher than that of the validation sample–only method adjusting \(PS_p\) in all scenarios. We do not show the power of the RC method here because it is subject to loss of control of type I error rate (Table 2).

**Analysis of the relationship between COPD and HZ based on the LHID and NHIS databases**

Table 4 compares some demographic and baseline clinical characteristics between the main and validation study samples, where the main study sample comes from the LHID database, and the validation study sample comes from the NHIS database. The age, sex, and comorbidity distributions are
very similar for the 2 samples. The only difference occurs in autoimmune disease, which is slightly more prevalent in the COPD cases of the main study than in the COPD cases of the validation study. Cumulative smoking and alcohol consumption are confounders observed only in the validation study and missing in the main study. As expected, the COPD patients tended to have higher cumulative smoking rates than the non-COPD subjects in the validation study. The difference in distributions of alcohol consumption between COPD and non-COPD subjects in the validation study sample is of no statistical significance.

We implement the TSC approach to assess the odds ratio of HZ associated with COPD adjusted for cumulative smoking and alcohol consumption through the crude propensity score \( PSC \) and the precise propensity score \( PSP \), which are described in the “Propensity score” subsection above. We also obtain the estimate of the odds ratio of HZ associated with COPD adjusted for \( PSC \) from the data pooling the main and validation study samples, as well as the estimate of the odds ratio of HZ associated with COPD adjusted for \( PSP \) from the validation study sample. In the analysis, the propensity scores enter into the working disease models (equations 1 and 2) through either the identity transformation (i.e., PS) or the natural cubic spline functions with 3 knots. As can be seen from Table 5, analysis results for the 2 variants are quite similar for all methods; hence, we focus on results based on PS. The estimate of the odds ratio of HZ associated with COPD adjusted for \( PSC \) from the pooled sample is 1.71 (95% confidence interval: 1.50, 1.95), which is lower than the estimate of 2.84 (95% confidence interval: 0.96, 8.39) adjusted for the precise propensity score \( PSP \) from the validation study sample. In contrast, the TSC estimate is 1.91 (95% confidence interval: 1.62, 2.26), which is larger in magnitude than the pooled-sample estimate adjusted for \( PSC \) and clearly more accurate in estimation and more powerful in significance testing than the validation-sample estimate (\( P < 0.001 \) for TSC vs. \( P = 0.06 \) for the validation estimate). We also obtain the bootstrap confidence intervals for each method (based on 1,000 bootstrap samples), which are close to those based on asymptotic theory.

Because we indeed have data on the onset time of HZ or the censoring (follow-up) time for subjects in the main and validation studies, we further performed analyses parallel to those mentioned above, where each working outcome model is given by the Cox regression model with the same covariate variables as in the original logistic regression model. For each method, the estimate of the hazard ratio for incidence of HZ-associated COPD is similar to the corresponding odds

<table>
<thead>
<tr>
<th>Setting Values</th>
<th>( \beta )</th>
<th>SE(( \beta ))</th>
<th>Size, %</th>
<th>( \beta )</th>
<th>SE(( \beta ))</th>
<th>Size, %</th>
<th>( \beta )</th>
<th>SE(( \beta ))</th>
<th>Size, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( PS^* ) NCS(^b )</td>
<td>( PS^* ) NCS(^b )</td>
<td>( PS^* ) NCS(^b )</td>
<td>( PS^* ) NCS(^b )</td>
<td>( PS^* ) NCS(^b )</td>
<td>( PS^* ) NCS(^b )</td>
<td>( PS^* ) NCS(^b )</td>
<td>( PS^* ) NCS(^b )</td>
<td>( PS^* ) NCS(^b )</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.02</td>
<td>0.03</td>
<td>0.37</td>
<td>0.39</td>
<td>3.6</td>
<td>4.5</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.03</td>
<td>0.03</td>
<td>0.38</td>
<td>0.39</td>
<td>3.5</td>
<td>3.9</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.05</td>
<td>0.05</td>
<td>0.33</td>
<td>0.33</td>
<td>3.6</td>
<td>3.7</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.01</td>
<td>0.00</td>
<td>0.33</td>
<td>0.33</td>
<td>4.3</td>
<td>4.3</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.35</td>
<td>0.35</td>
<td>4.5</td>
<td>4.0</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.5</td>
<td>4</td>
<td>0.01</td>
<td>0.03</td>
<td>0.33</td>
<td>0.30</td>
<td>3.2</td>
<td>5.7</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.02</td>
<td>0.01</td>
<td>0.34</td>
<td>0.27</td>
<td>2.4</td>
<td>4.6</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.02</td>
<td>0.01</td>
<td>0.33</td>
<td>0.25</td>
<td>3.5</td>
<td>4.3</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.01</td>
<td>0.01</td>
<td>0.33</td>
<td>0.26</td>
<td>3.8</td>
<td>4.8</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.01</td>
<td>0.01</td>
<td>0.35</td>
<td>0.27</td>
<td>5.3</td>
<td>4.5</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Abbreviations: NCS, natural cubic spline; OR, odds ratio; PS, propensity score; RC, regression calibration; SE, standard error; TSC, 2-stage calibration.

\(^a\) Adjusted for the propensity score itself.

\(^b\) Adjusted for natural cubic spline functions of the propensity score.

ratio estimate because the follow-up durations are relatively short (i.e., 2 years). Again, we observe that the TSC estimate is closer to the estimate from the validation data than the estimate from the pooled sample and has a much narrower confidence interval and more powerful testing result than the estimate from the validation data ($P < 0.001$ for TSC vs. $P = 0.06$ for the validation estimate).

**DISCUSSION**

Analysis of large administrative databases has become popular in medical research in recent years because researchers can readily and inexpensively gather longitudinal data from such databases for a large group of subjects without needing to conduct lengthy prospective studies. Data from such databases also bypass recall bias when compared with self-reported data in common retrospective studies. However, a major limitation for administrative databases is that they often lack data on more detailed confounding information such as cigarette smoking, alcohol consumption, dietary intake, and occupational exposure. Naive analysis with data of this type is subject to confounding bias. A way to circumvent this drawback is to further gather data on more precise confounding information from certain specialized research or survey databases that contain such data. Data collected in this way then serve as the validation data, which provide a basis for correcting confounding bias in the main study based on administrative databases.

Stürmer et al. (4) have successfully implemented the idea of using external validation data on confounding variables with the aid of the techniques of RC and propensity score.

---

**Table 3.** Parameter Estimate for $\beta = \log(\text{OR}_{2})$, Standard Error, and Power for the Significance Test of $\beta = 0$ Based on the Second-Stage Analysis Adjusting the Precise Propensity Score $\text{PS}_p$ and the TSC With the Relationship Between Disease and $\text{PS}_p$ Misspecified

<table>
<thead>
<tr>
<th>Setting Values</th>
<th>Adjusted for $\text{PS}_p$</th>
<th>$\beta$</th>
<th>SE($\beta$)</th>
<th>Power, %</th>
<th>Adjusted for $\text{TSC}$</th>
<th>$\beta$</th>
<th>SE($\beta$)</th>
<th>Power, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PS$^a$ NCS$^b$ PS$^a$ NCS$^b$</td>
<td>PS$^a$ NCS$^b$</td>
<td></td>
<td></td>
<td>PS$^a$ NCS$^b$ PS$^a$ NCS$^b$</td>
<td>PS$^a$ NCS$^b$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta = 0.4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{OR}<em>{2}$ by $\text{OR}</em>{T}$ 2</td>
<td>2</td>
<td>0.40</td>
<td>0.36</td>
<td>0.37</td>
<td>14.5</td>
<td>16.3</td>
<td>0.35</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.39</td>
<td>0.32</td>
<td>0.32</td>
<td>19.0</td>
<td>19.1</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.38</td>
<td>0.31</td>
<td>0.30</td>
<td>21.6</td>
<td>20.7</td>
<td>0.37</td>
<td>0.35</td>
</tr>
<tr>
<td>$\beta = -0.4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{OR}<em>{2}$ by $\text{OR}</em>{T}$ 2</td>
<td>2</td>
<td>-0.33</td>
<td>-0.36</td>
<td>0.35</td>
<td>0.37</td>
<td>16.6</td>
<td>18.6</td>
<td>-0.35</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-0.34</td>
<td>0.32</td>
<td>0.32</td>
<td>20.2</td>
<td>20.0</td>
<td>-0.35</td>
<td>-0.36</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>-0.36</td>
<td>0.32</td>
<td>0.32</td>
<td>20.5</td>
<td>19.6</td>
<td>-0.36</td>
<td>-0.35</td>
</tr>
<tr>
<td>$\beta = 0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{OR}<em>{2}$ by $\text{OR}</em>{T}$ 2</td>
<td>2</td>
<td>0.41</td>
<td>0.31</td>
<td>0.31</td>
<td>23.5</td>
<td>23.0</td>
<td>0.38</td>
<td>0.37</td>
</tr>
<tr>
<td>$\beta = -0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{OR}<em>{2}$ by $\text{OR}</em>{T}$ 2</td>
<td>2</td>
<td>-0.36</td>
<td>-0.38</td>
<td>0.37</td>
<td>0.39</td>
<td>15.9</td>
<td>18.4</td>
<td>-0.36</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-0.37</td>
<td>0.35</td>
<td>0.35</td>
<td>17.7</td>
<td>18.2</td>
<td>-0.38</td>
<td>-0.38</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>-0.39</td>
<td>0.36</td>
<td>0.36</td>
<td>17.9</td>
<td>17.3</td>
<td>-0.39</td>
<td>-0.38</td>
</tr>
</tbody>
</table>

Abbreviations: NCS, natural cubic spline; OR, odds ratio; PS, propensity score; SE, standard error; TSC, 2-stage calibration.

$^a$ Adjusted for the propensity score itself.

$^b$ Adjusted for natural cubic spline functions of the propensity score.
McCandless et al. (6) proposed an alternative Bayesian propensity score approach, which enjoys good estimation efficiency and avoids the surrogacy assumption needed in the method of Stürmer et al., although modeling assumptions on the marginal relationship between measured and unmeasured confounders are still needed. Faries et al. (16) considered a similar Bayesian approach with propensity score adjustments, also requiring assumptions for the distribution of the unmeasured confounder given the measured ones. Schneeweiss et al. (3) used external information on prevalence of confounding variables, the confounder-exposure association, and the confounder-disease association, to perform sensitivity analysis on residual confounding bias for the exposure-disease association arising from unadjusted confounders.


Table 4. Demographic Characteristics and Comorbid Disorders for COPD and Non-COPD Cohorts in the Main and Validation Studies From the Longitudinal Health Insurance Database and the National Health Interview Survey, Taiwan, 2005

<table>
<thead>
<tr>
<th>Baseline Variable</th>
<th>Main Study</th>
<th>Validation Study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COPD (n=8,486)</td>
<td>Non-COPD (n=33,944)</td>
</tr>
<tr>
<td></td>
<td>No.</td>
<td>%</td>
</tr>
<tr>
<td>Follow up, years</td>
<td>2.3 (0.63)a</td>
<td>2.5 (0.52)a</td>
</tr>
<tr>
<td>Age, years</td>
<td>71.2 (9.74)a</td>
<td>70.3 (9.55)a</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>5,858</td>
<td>69.0</td>
</tr>
<tr>
<td>Female</td>
<td>2,628</td>
<td>31.0</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1,507</td>
<td>17.8</td>
</tr>
<tr>
<td>No</td>
<td>6,979</td>
<td>82.2</td>
</tr>
<tr>
<td>Hypertension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>4,290</td>
<td>50.6</td>
</tr>
<tr>
<td>No</td>
<td>4,196</td>
<td>49.4</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>4,868</td>
<td>57.4</td>
</tr>
<tr>
<td>No</td>
<td>3,618</td>
<td>42.6</td>
</tr>
<tr>
<td>Chronic liver disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>930</td>
<td>11.0</td>
</tr>
<tr>
<td>No</td>
<td>7,556</td>
<td>89.0</td>
</tr>
<tr>
<td>Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>503</td>
<td>5.9</td>
</tr>
<tr>
<td>No</td>
<td>7,983</td>
<td>94.1</td>
</tr>
<tr>
<td>Autoimmune disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>204</td>
<td>2.4</td>
</tr>
<tr>
<td>No</td>
<td>8,282</td>
<td>97.6</td>
</tr>
<tr>
<td>Cumulative smoking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>138</td>
<td>56.6</td>
</tr>
<tr>
<td>Few</td>
<td>2</td>
<td>0.8</td>
</tr>
<tr>
<td>&lt;100 cigarettes</td>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td>≥100 cigarettes</td>
<td>101</td>
<td>41.4</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never/seldom</td>
<td>207</td>
<td>84.8</td>
</tr>
<tr>
<td>&lt;Once a month</td>
<td>10</td>
<td>4.1</td>
</tr>
<tr>
<td>Twice a month</td>
<td>6</td>
<td>2.5</td>
</tr>
<tr>
<td>Weekly</td>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td>Once every 2–3 days</td>
<td>4</td>
<td>1.6</td>
</tr>
<tr>
<td>Once every day</td>
<td>14</td>
<td>5.7</td>
</tr>
</tbody>
</table>

Abbreviation: COPD, chronic obstructive pulmonary disease.

a Value expressed as mean (standard deviation).
Table 5. Various Estimates for the Odds Ratios and Hazard Ratios of Herpes Zoster Associated With Chronic Obstructive Pulmonary Disease, Longitudinal Health Interview Database and National Health Interview Survey, Taiwan, 2005

<table>
<thead>
<tr>
<th>Model</th>
<th>OR/HR</th>
<th>95% CI</th>
<th>Bootstrap 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted for PS&lt;sub&gt;C&lt;/sub&gt;&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.71</td>
<td>1.50, 1.95</td>
<td>1.50, 1.94</td>
</tr>
<tr>
<td>Adjusted for PS&lt;sub&gt;p&lt;/sub&gt;&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2.84</td>
<td>0.96, 8.39</td>
<td>0.79, 9.20</td>
</tr>
<tr>
<td>NCS&lt;sup&gt;d&lt;/sup&gt;</td>
<td>1.91</td>
<td>1.62, 2.26</td>
<td>1.59, 2.25</td>
</tr>
<tr>
<td>Adjusted for PS&lt;sub&gt;C&lt;/sub&gt;</td>
<td>1.71</td>
<td>1.50, 1.95</td>
<td>1.49, 1.97</td>
</tr>
<tr>
<td>Adjusted for PS&lt;sub&gt;p&lt;/sub&gt;</td>
<td>2.82</td>
<td>0.96, 8.31</td>
<td>0.69, 8.50</td>
</tr>
<tr>
<td>TSC</td>
<td>1.97</td>
<td>1.60, 2.43</td>
<td>1.54, 2.50</td>
</tr>
<tr>
<td>Cox Regression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted for PS&lt;sub&gt;C&lt;/sub&gt;</td>
<td>1.96</td>
<td>1.72, 2.24</td>
<td>1.72, 2.23</td>
</tr>
<tr>
<td>Adjusted for PS&lt;sub&gt;p&lt;/sub&gt;</td>
<td>2.89</td>
<td>0.96, 8.71</td>
<td>0.62, 9.85</td>
</tr>
<tr>
<td>TSC</td>
<td>2.21</td>
<td>1.79, 2.72</td>
<td>1.82, 2.65</td>
</tr>
<tr>
<td>NCS&lt;sup&gt;d&lt;/sup&gt;</td>
<td>1.96</td>
<td>1.72, 2.24</td>
<td>1.72, 2.23</td>
</tr>
<tr>
<td>Adjusted for PS&lt;sub&gt;p&lt;/sub&gt;</td>
<td>2.86</td>
<td>0.96, 8.57</td>
<td>0.68, 9.53</td>
</tr>
<tr>
<td>TSC</td>
<td>2.28</td>
<td>1.80, 2.90</td>
<td>1.77, 2.89</td>
</tr>
</tbody>
</table>

Abbreviations: CI, confidence interval; HR, hazard ratio; NCS, natural cubic spline; OR, odds ratio; PS, propensity score; TSC, 2-stage calibration.

<sup>a</sup> Adjusted for the propensity score itself.

<sup>b</sup> Crude propensity score predicted from confounding factors including age, sex, diabetes mellitus, hypertension, coronary artery disease, chronic liver disease, autoimmune disease, and cancer.

<sup>c</sup> Precise propensity score predicted from the same confounding factors as in the crude propensity score PS<sub>C</sub>, plus cumulative smoking and alcohol consumption.

<sup>d</sup> Adjusted for natural cubic spline functions of the propensity score.

The focus of their method is on sensitivity evaluation rather than formal inference of the exposure effect.

In this study, we propose the TSC approach to adjust for confounding variables unobserved in the main study using validation data. In contrast to the RC method (4) and the Bayes propensity score method (6, 16), the TSC approach does not depend on the surrogacy assumption and the modeling assumption on the relationship between measured and unmeasured confounders. We believe this relaxation is important because it renders our proposal wider applicability even in settings with complex measurement error structures, such as case-control studies with differential measurement errors.

As in other methods combining a main study with an external study, such as those of Stürmer et al. (4) and McCandless et al. (6), our proposal requires that all the variables are measured in the same fashion between the 2 studies combined. Therefore, a necessary criterion for choosing an external data sample is the consistency in the definition and measurement instrument of the variables between the external and the main studies. In our motivating study, because the outcome (HZ status), the exposure (COPD status), and the measured confounders (comorbid disorders) in the validation and the primary studies are defined by the same criteria on the basis of *International Classification of Diseases, Ninth Revision, Clinical Modification*, codes, the consistency issue is minimal. In fact, as is evident from equation 3, as long as the association parameters γ and β measuring the association between outcome and exposure adjusted for crude and precise confounder measurements, respectively) are virtually consistent between the main and the validation studies, which should be the exact condition under which combining 2 studies is valid for inference on the association, the proposed TSC estimator will be virtually unbiased, even though the marginal distributions of individual variables are different between the 2 studies.

Our approach requires that the outcome variable is also observed in the validation sample, unlike the RC method by Stürmer et al. (4), in which the validation sample may not contain outcome data. This implies that our approach is akin to using internal validation data (9, 15). Methods by McCandless et al. (6) and Faries et al. (16) are similarly applicable in studies with internal validation data. Our method is easier to implement than the latter ones based on the Bayesian analysis, and, as discussed above, our method requires fewer modeling assumptions.

The focus of the paper is the adjustment for confounders using propensity scores. Our method can also be applied to adjust for missing confounders directly. This can be seen by replacing equations 1 and 2 with

\[
\log\left(\frac{\Pr(D = 1|T, C)}{\Pr(D = 0|T, C)}\right) = \delta + \gamma T + \theta C,
\]

and then using the same procedure listed in the subsection, “Extensions and summary”; see also section 5.2 of the article by Chen and Chen (9). It suggests a general methodology to adjust for several missing confounders, without having to make a multiple imputation model for missing data.

In the application of the TSC method to the LHID and NHIS databases, with the former used as the main study and the latter used as the validation study, the analysis result reveals that patients with COPD are at higher risk of developing HZ compared with the general population after controlling for other HZ risk factors including cigarette smoking and alcohol consumption. This result improves that obtained by Yang et al. (10), in the sense that important confounders, cigarette smoking and alcohol consumption, have been adjusted for in the analysis. The increased risk of HZ in patients with COPD deserves physicians’ attention. Postherpetic neuralgia, the substantial morbidity caused by HZ, can cause depression and anxiety, which in turn can negatively affect daily functioning (17) and result in further deterioration of quality of life in patients who already suffer from chronic lung disease.
ACKNOWLEDGMENTS

Author affiliations: Department of Mathematics, Soochow University, Taipei, Taiwan (Hui-Wen Lin); Institute of Statistical Science, Academia Sinica, Taipei, Taiwan (Yi-Hau Chen); and Institute of Public Health, National Yang-Ming University, Taipei, Taiwan (Yi-Hau Chen).

This study is based in part on data from the National Health Insurance Research Database provided by the Bureau of National Health Insurance, Department of Health, managed by National Health Research Institutes in Taiwan, and supported by the National Science Council of Taiwan.

Conflict of interest: none declared.

REFERENCES