Published in final edited form as:
Med Care. 2007 October ; 45(10 SUPL): S158–S165.

Adjustments for Unmeasured Confounders in Pharmacoepidemiologic Database Studies Using External Information

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

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

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

3 *Department of Biostatistics, Harvard School of Public Health, Boston, MA*

4 *RTI Health Solutions, Research Triangle Park, NC*

5 *Department of Epidemiology, Boston University School of Public Health, Boston, MA*

6 *Department of Epidemiology, Harvard School of Public Health, Boston, MA*

Abstract

Background—Non-experimental studies of drug effects in large automated databases can provide timely assessment of real-life drug use, but are prone to confounding by variables that are not contained in these databases and thus cannot be controlled.

Objectives— To describe how information on additional confounders from validation studies can help address the problem of unmeasured confounding in the main study.

Research Design—Review types of validation studies that allow adjustment for unmeasured confounding and illustrate these with an example.

Subjects: Main study—New Jersey residents 65 years or older hospitalized between 1995 and 1997, who filled prescriptions within Medicaid or a pharmaceutical assistance program. Validation study: representative sample of Medicare beneficiaries.

Measures—Association between nonsteroidal anti-inflammatory drugs (NSAIDs) and mortality.

Author for correspondence and reprints: Til Stürmer, MD, MPH, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120, Phone: +1 617 278 0627, Fax: +1 617 232 8602, Email: til.sturmer@post.harvard.edu.
Til Stürmer, MD, MPH, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120, Phone: +1 617 278 0627, Fax: +1 617 232 8602, Email: til.sturmer@post.harvard.edu
Robert J Glynn, ScD, PhD, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120, Phone: +1 617 278 0972, Fax: +1 617 232 8602, Email: rglynn@rics.bwh.harvard.edu
Kenneth J Rothman, DrPH RTI Health Solutions PO Box 12194, Research Triangle Park, NC 27709, Phone: +1 617 964 0977, Email: krothman@rti.org
Jerry Avorn, MD, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120, Phone: +1 617 278 0930 ,Fax: +1 617 232 8602, Email: javorn@partners.org
Sebastian Schneeweiss, MD, ScD, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120, Phone: +1 617 278 0937, Fax: +1 617 232 8602, Email: schneeweiss@post.harvard.edu

Results—Validation studies are categorized as internal (ie, additional information is collected on participants of the main study) or external. Availability of information on disease outcome will affect choice of analytic strategies. Using an external validation study without data on disease outcome to adjust for unmeasured confounding, propensity score calibration (PSC) leads to a plausible estimate of the association between NSAIDs and mortality in the elderly, if the biases caused by measured and unmeasured confounders go in the same direction.

Conclusions—Estimates of drug effects can be adjusted for confounders that are not available in the main but can be measured in a validation study. PSC uses validation data without information on disease outcome under a strong assumption. The collection and integration of validation data in pharmacoepidemiology should be encouraged.

Keywords

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

Introduction

Randomized controlled trials are often regarded as the most accurate method to assess treatment effects. Random assignment of a large number of subjects into treatment groups usually leads to a good balance of observed and unobserved risk factors in all groups. Nevertheless, randomized controlled trials have major limitations when they are used to assess the role of medications in the etiology and management of chronic diseases. The main limitations stem from selection of participants into trials who are healthier and want to pursue a more healthful life style, the long time required from trial design to completion, the relatively short duration of exposure, and high cost. In addition, frail elderly patients, who use the most drugs and have the highest adverse event rates, are often underrepresented in trials. Generalization from trials can be erroneous because effect sizes, baseline risks, and comorbidity have been shown to differ between trial populations and the broader population that is not represented in trials.[1]

In observational studies it is possible to include a wide variety of participants and assess long-term exposures, including medications, in a timely manner. Without treatment allocation by chance, however, bias due to different baseline risks for disease in users and non-users of drugs cannot be ruled out completely. With respect to prescribed medications, this bias is called confounding by indication.[2] It stems from risk factors for disease that influence the treatment choices of physicians and patients, including the decision to start or stay on a drug. Some of these factors, like attitudes towards health,[3] prevention,[4] and frailty,[5,6] are subtle and hard to measure. The potential for confounding is probably larger in observational studies assessing medications than in studies assessing lifestyle factors. There is increasing evidence (eg, 7) that confounding by indication may not have been completely controlled for in observational studies on pharmacoprevention of chronic diseases.[8]

The problem can be more pronounced in pharmacoepidemiologic studies based on large health care utilization databases, since these data are collected for reasons unrelated to the research hypothesis and thus rarely contain sufficient information on all important confounders. These databases are, however, often the best source of information on the association between drugs and diseases, because they allow timely assessment of infrequently used drugs and rare outcomes in large, representative populations under usual care conditions.

To address unmeasured confounding in pharmacoepidemiologic database studies, several approaches have been proposed.[9] They can be classified according to whether they use additional data not contained in the original administrative dataset or not. Examples of methods that do not rely on additional data collection include instrumental variable analyses,[10,11]

methods that calculate bounds for causal effects under increasingly restrictive assumptions about the unmeasured confounder and the effect of the exposure,[12–14] as well as case-crossover studies[15]. Here, we will focus on analytic methods that control for bias due to confounding that is measured not in the main study, but in an outside body of data that can serve as a validation study.

The aims of the paper are to describe the use of external data to adjust for unmeasured confounding, propose a framework to categorize different uses of outside data, present an application to adjust for unmeasured confounding based on a method developed by our group, and advocate the use of and research into external information in pharmacoepidemiologic studies.

Framework

There is no uniform framework for the use of external information to adjust for unmeasured confounding. So far, most research has focused on the validation of exposures[16] and disease outcomes[17] to reduce bias due to misclassification or measurement error. Although many of these issues apply to external information on confounders, the problem is distinct. Apart from misclassification and measurement error of the confounder leading to residual confounding, confounding is defined by the joint distribution of the confounder(s) with both exposure and outcome. Therefore both associations need to be considered when deciding on the use of external information. Here we will focus on the assessment of confounding. In any setting, researchers may want to make optimal use of the additionally collected data by combining external information on confounding with validation of exposures and outcomes to quantify more than 1 source of bias.[18,19]

Aggregated data: Sensitivity analyses of individual confounders

Simple sensitivity analyses can be performed without use of additional data (eg, by making structural assumptions and selecting parameters for a single confounder that would explain the observed study finding under an alternative, usually null, hypothesis)[12,20]. They can be extended to use parameter values from published studies or outside data sources (aggregated or individual data) to adjust observed estimates based on the distribution of specific confounders (eg, smoking) in the population and their association with exposure and with disease outcomes.[21] Such analyses are advantageous because they can be easily performed based on information from the literature or expert opinion on the distribution of a specific unmeasured confounder and its associations with exposure and disease outcome. No additional data need be collected, and these analyses provide bounds for possible confounding by individual factors. Their main disadvantage is that they do not address joint confounding by several unmeasured covariates.

To address confounding by multiple confounders, an extension of these simple sensitivity analyses has been applied.[22] The approach is based on separate estimates of the prevalence of external dichotomous confounders, their association with the exposure of interest, and their association with disease outcome. Using a weighted average of the expected confounding effect of several covariates, this ad hoc method allows researchers to approximate the magnitude and direction of confounding by several covariates. The method treats the unobserved confounders as separate covariates without taking into account their joint distributions with the exposure and the disease outcome.

Using data from a cross-sectional validation study, Schneeweiss et al. could show that selective COX2 inhibitor users were less likely to be smokers (8% versus 10%) than nonselective NSAID users, while the prevalence of obesity (24% vs. 24%) was comparable. Failure to adjust for five potential confounders not measured in the administrative claims data (smoking, obesity,

aspirin use, education, and income) would lead to only a small underestimation of the association between selective COX2 inhibitors and myocardial infarction and is thus unlikely to introduce substantial bias in the study based on claims data.[22]

Although this approach is a significant improvement over analyses based on a single confounder, it does not allow researchers to address possible joint confounding by these variables. The method can still be applied based on distributions and associations taken from the literature. If it is based on information contained in outside study data on individuals, however, it does not make optimal use of the joint distribution of the covariates contained in the validation study.

Individual-level data: External adjustments addressing the joint distribution of several confounders

To address joint confounding by multiple covariates, data on the joint distribution of these covariates are necessary. This procedure requires individual-level data from validation studies. Such studies can be categorized as external or internal depending on whether data are collected from individuals outside or within the main study population.

Internal validation studies

Internal validation studies can be defined as studies based on additional data obtained for a subset of participants in the main study population. This sub-sample can then be conceptualized as “complete subjects,” in contrast to the other participants, who are regarded as incomplete subjects because they have missing values for all confounders assessed only in the validation study. The main study participants on whom additional data are collected in the validation study can be a random sub-sample of the baseline cohort. In that case, a method such as multiple imputation[23] can be used to fill in the missing values of the confounder(s) in the main dataset to control for unmeasured confounding.

Garshick et al. recently used multiple imputation in a retrospective cohort study on lung cancer in railroad workers exposed to diesel exhaust without information on smoking.[24] Without adjustment for smoking, the relative risk of lung cancer in those exposed to diesel exhaust was 1.35. Smoking history was obtained using data from a companion case-control study. Smoking histories of 5 random persons of the validation study with the same job category (the exposure of interest), age, birth cohort, and outcome of interest (whether the person died of lung cancer or not) were imputed for each person of the main study. The resulting 5 different datasets were then analyzed separately and results combined to obtain mean estimates and their standard errors. This multiple imputation led to an estimate of the relative risk for lung cancer and exposure to diesel exhausts of 1.22. Based on data on smoking from the validation study, the authors conclude that small differences in smoking behavior between diesel exposed and unexposed workers do explain some but not all of the elevated risk.[24] Although the companion study is not an internal validation study in the strict sense of the term, the study clearly contains enough detail on the exposure of interest, measured confounders, and the outcome of interest (a prerequisite for multiple imputation to be applicable)[25] to be used as an example of an internal validation study.

Instead of imputing actual values from random draws of persons from the validation study, multiple imputation is often based on fitting a linear model in the validation study with the covariate that is missing in the main study included as the dependent variable. Imputations are then based on multiple draws from the posterior distribution of the parameter estimates of this regression model plus an additional random error term.

To increase efficiency, selection into the validation study can be based on available information about exposure or disease outcome of interest or both (ie, non-random selection). This approach is referred to as anamorphic design[26] or 2-stage design.[27,28] The data can then be analyzed using a method that takes into account the sampling for the internal validation study. Examples of such a method are maximum likelihood, the method developed by Breslow and Cain, and estimating equations.[29]

The main advantages of internal validation studies over external validation studies are 1) better representativeness of the main study and 2) possible efficiency gains based on over-sampling on exposure or disease outcome of interest. Most published work on efficiency gains by different non-random sampling strategies (eg, counter-matching,[30] flexible matching[31]) is based on estimation of main exposure effects and interactions. The general principles of non-random selection to increase efficiency also hold when considering confounding.[32]

Examples of 2-stage designs include an early cohort study on vasectomy and non-fatal myocardial infarction[33] and a nested case-control study on nonsteroidal anti-inflammatory drugs and breast cancer[34]. Both use claims data to define the study population, the drug exposure of interest, and the outcome. In the study by Walker et al., information on additional confounders for matched sets of exposed and unexposed was collected for sets that included at least one outcome (myocardial infarction).[33] Chart review of these sets allowed the authors to efficiently validate both exposure and outcome and abstract information on additional potential confounders including obesity, hypertension, diabetes, and smoking history. Because the relative risk for vasectomy and myocardial infarction was essentially the same before and after control for these additional covariates in the stage 2 sample (1.25 vs. 1.22), the authors concluded that there was little evidence for confounding by these factors and reported the stage 1 estimate as their primary finding (RR=1.0).[33]

Similarly, Sharpe et al. report the stage 2 estimate based on information on additional confounders obtained by telephone interview in selected cases and controls separately from the stage 1 estimate (not adjusting for these potential confounders).[34] Based on the similarity of these two estimates the authors conclude that the protective effect (of nonsteroidal anti-inflammatory drugs on breast cancer) could not be attributed to confounding by other determinants.[34]

Possibly because of the absence of joint confounding by the additional covariates, both studies use the information from stage 2 qualitatively. If confounding by the additional covariates is detected in stage 2, however, the estimates from stages 1 and 2 can be quantitatively combined into an adjusted overall estimate.[35]

The main disadvantage of an internal validation study based on non-random sampling is that it is specific to a certain exposure-outcome association, and therefore less efficient for addressing unmeasured confounding for other associations. Internal validation studies depend on the feasibility of collecting additional data from study participants. Contacting individuals whose data are routinely collected for administrative purposes may not be possible because of privacy concerns or laws. Temporality and (differential) recall-bias are further concerns that need to be considered.

External validation studies

External validation studies are usually cross-sectional because of cost and time constraints in collecting information on infrequent disease outcomes. Their main advantages are that data are often already collected and can be re-used for several main studies addressing multiple hypotheses. The main disadvantage of external validation studies is that they often do not contain exactly the same measures used in the main study, that they are not perfectly

representative of the main study, and that they usually lack information about the disease outcome of interest. All of these challenge the assessment of the joint distribution of the unmeasured confounders with the disease outcome of interest.

Since external validation studies are not specific to a certain hypothesis, many such studies are available, including public domain data or data that can be purchased for a small fee. Their use in pharmacoepidemiology is often limited, however, because data on medication use from external sources is often cross-sectional (eg, current use is assessed based on self-administered questionnaires or interviews) with varying potential for misclassification and a limited number of incident users. To provide enough information on rarely used drugs, including newly marketed ones, validation studies need to be large and current. We present a detailed example of an external validation study in a following section.

In Table 1 we summarize advantages and limitations of internal and external validation studies, and list possible analytic techniques to use data from these studies to adjust for confounding unmeasured in the main study.

Application example

In the following sections we illustrate how external validation data without information on the disease outcome of interest can be used to adjust estimates of an exposure-outcome association based on published work. This application example has been previously described in detail[36] and is only summarized here.

There is no known biological reason to expect that nonsteroidal anti-inflammatory drugs (NSAIDs) would cause a reduction in the risk of death (indeed, there is some evidence for the contrary). Glynn et al. observed that NSAID were associated with a strong reduction in risk for short-term mortality (relative risk = 0.74) in elderly hospitalized patients, however, even after a wide variety of health indicators available in claims data were controlled for.[6] This association is likely to be due to selection bias leading to strong unmeasured confounding: Physicians are less likely to prescribe NSAIDs (eg, compared with narcotics) in frail old adults as well as in patients with advanced cancer and a variety of other comorbidities, including renal disease, that are associated with a high mortality. Some of these, however, are measured in claims data and a single unobserved confounder would need to be strongly associated with avoidance of NSAIDs and mortality as well as be prevalent to explain the strong inverse association between NSAIDs and mortality. Although a single confounder explaining the strong inverse association might be implausible, joint confounding by a variety of confounders, each of which is only moderately associated with NSAID use and mortality and not very prevalent, but all acting in the same direction might nevertheless be plausible.

To incorporate information on joint confounding by unmeasured covariates using data from a cross-sectional external validation study, we combined propensity scores (PS)[37] and regression calibration[38] into propensity score calibration (PSC)[36].

The PS is defined as the conditional probability of exposure (to a drug) given observed covariates. It is usually estimated from the data at hand using multivariable logistic regression. Individuals with the same estimated PS are then thought to have the same chance of being exposed. As a group, treated and untreated subjects paired on the same PS will have similar distributions of and thus comparisons be unconfounded by observed covariates.[37]

Regression calibration is a method to correct effect estimates for measurement error.[38] In the context of generally sparse use of methods to correct for measurement error in epidemiology, regression calibration is the most widely used approach. It is based on data from a validation study that includes the “error-prone” measure of the variable used in the main study and an

additional “gold-standard” measure of the same variable. Within the validation study, one estimates a linear measurement error model with the “true” or “gold-standard” variable as dependent variable and the “error-prone” variable and variables measured without error as independent covariates. Under the main assumption that the “error-prone” variable contains no information on the outcome beyond the “gold-standard” variable (surrogacy),[39] regression calibration then uses the regression estimates from this measurement error model to correct the 'naive' regression estimates obtained from error-prone variable in the main study.[36] Regression calibration is used mainly to correct associations between continuous exposures (eg, blood pressure, nutrients) and outcomes for measurement error in the exposure of interest.

To apply regression calibration to adjust for the joint confounding of multiple confounders unobserved in the main study, we first combine all confounders into a single score, the PS, and assume that the PS estimated in the main study based on a subset of important confounders is measured with error. This error can be estimated in an external validation study using data on additional confounders. One can then adjust for unmeasured confounding due to that measurement error using regression calibration.

To implement PSC, we first controlled for measured confounding in the main cohort using the “error-prone” PS estimated in the main study. We then estimated 2 additional PSs in the external validation study: the “error-prone” PS based on information available in the main cohort, and the “gold-standard” PS that included covariates available only in the validation study (see Table 2). Based on these 2 PSs in the validation study, we applied regression calibration to correct regression coefficients in the main cohort.[36]

Main study

To test this approach, we identified a main study population assembled for an analysis of pain medication use in elderly patients.[40] It comprised all community dwelling New Jersey residents who were aged 65 years or older, filled prescriptions within Medicaid or the Pharmaceutical Assistance to the Aged and Disabled program, and were hospitalized between January 1, 1995 and December 31, 1997. Eligible individuals were those who filled a prescription for any drug within 120 days before hospitalization and another prescription more than 365 days before hospitalization. Covariates were assessed during the 365 days before hospitalization.

For all 103,133 eligible subjects we extracted the following variables: age, sex, race, all prescriptions filled within 120 days before the date of hospital admission, all diagnoses assigned, number of hospitalizations, and number of physician visits within 365 days before that date. The time until death or 365 days of follow-up (whichever came first) was assessed starting from the date of hospital admission, based on linkage to Medicare files.[41]

External 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 medication use over the last 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 (between 85% and 95%) and very high data completeness.[42]

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 living in the community (10,446 persons). To make the validation study population more comparable with the main study, we

randomly selected MCBS individuals according to the age (3 categories) and sex distribution in the main cohort (frequency matching). This resulted in 5,108 MCBS subjects used for all subsequent analyses.

Control for observed confounding

During a follow-up period of 1 year, 21,928 (21.3%) of the main study population died during or after hospitalization. Without any control for confounding, NSAID use appeared to be associated with a 32% (95% CI: 29 – 36%) mortality risk reduction (Table 3). This observed association is likely to be due to selection bias (ie, the fact that physicians are less likely to treat pain with NSAIDs in frail old adults). Controlling for just age and gender in the conventional outcome model, we observed a smaller estimate of decreased risk (26%; 95% CI: 23 – 29%) compared with the unadjusted result. Controlling for a wide variety of health indicators available in claims data (for a list, see Table 3 footnote) in the outcome model, we observed a risk reduction of 20% (95% CI: 17 – 23%). We observed essentially the same amount of risk reduction (19%; 95% CI: 16 – 22%) when we controlled for confounding using a PS that was estimated based on the same claims data covariates as were used in the outcome model and modeling mortality as a function of the estimated PS together with the exposure.

Control for unobserved confounding with Propensity Score Calibration

We implemented the PSC approach by using regression calibration to correct for measurement error in the PS of the main study. Regression calibration was based on the estimation of the 2 PSs ("error-prone" and "gold-standard") in the external validation study (Table 4). Better prediction of exposure by inclusion of the survey information in the "gold-standard" PS resulted in an increased c-statistic or area under the receiver operating characteristic curve (AUC) of 0.66, compared with 0.60 in the "error-prone" PS. None of the additional variables entered in the "gold-standard" PS is strongly related to exposure to NSAIDs. Nevertheless, the prediction of NSAIDs exposure is substantially improved as quantified by the c-statistic, giving plausibility to our hypothesis that the strong unmeasured confounding in the main study is not due to a single dominant confounder but rather due to the joint effect or combination of multiple modest confounders. In general, however, the c-statistic of the PS has only limited value when assessing its performance to control for confounding.[43]

We implemented regression calibration using a linear measurement error model. The "gold-standard" PS that predicted NSAID exposure based on claims data combined with interview data was the dependent variable; the "error-prone" PS that predicted NSAID exposure based on claims data only was one independent variable, and NSAID exposure served as the second independent variable. As noted above, the analysis of the exposure-outcome association controlling for the "error-prone" PS in the main study indicated a 19% risk reduction. When we adjusted this estimate for the measurement error in the "error-prone" PS (as estimated by comparison with the "gold-standard" PS in the validation study), we found NSAID use to be associated with a very small 6% (95% CI: 0 to 11%) increased mortality risk after PSC (Table 3). Thus, adding information on additional confounders using PSC resulted in a more plausible estimate for the association between NSAID use and all-cause short-term mortality in the elderly.[36]

Like regression calibration, PSC is dependent on the surrogacy assumption that the "error-prone" PS does not contain any information on the disease outcome given the "gold-standard" PS and exposure.[39,44] In simulations over a wide range of parameters we found that PSC is valid if surrogacy holds, but that it can increase rather than decreases bias in situations where surrogacy is violated.[45] Surrogacy holds when the directions of confounding by the measured and the unmeasured covariates are the same. In the NSAIDs example, it is plausible that the underlying frailty leading both to a lower propensity for NSAID use and to higher mortality is

only partly captured in the main study and is better captured when additional information from the validation study (eg, data on activities of daily living) is added. Thus, surrogacy might be assumed in this setting. The "gold-standard" PS also performs as an approximate instrument under assumptions similar to surrogacy because it hopefully better approaches the true, but unknown, propensity of treatment than the "error-prone" one.[39,46]

Conclusions

The advantages of health care utilization data for the assessment of intended and unintended drug effects have a price: the relatively narrow set of variables available to characterize patients. This limitation makes the data from such analyses vulnerable to confounding bias, such as confounding by indication. Of the possible sources of additional data to adjust for unmeasured confounding, cross-sectional external validation studies are most widely available. Internal validation studies offer specific advantages since they are more representative and allow application of methods that do not require assumptions about the direction of confounding. Such validation studies should therefore be considered whenever feasible. Using PSC to incorporate information on confounders not available in health care utilization data from external validation studies can help researchers to adjust for unmeasured confounding, albeit under a strong surrogacy assumption.

Acknowledgements

The manuscript was prepared for presentation at the Agency for Healthcare Research and Quality meeting on "Comparative Effectiveness and Safety: Emerging Methods Symposium", June 19–20, 2006, in Rockville, MD. This project was funded by a grant (RO1 023178) from the National Institute on Aging.

Sources of support: Grant (RO1 023178) from the National Institute on Aging Running head: Adjustments for Unmeasured Confounders

References

1. Strom, BL., editor. Pharmacoepidemiology. 4. Sussex: John Wiley & Sons Ltd; 2005.
2. Walker AM. Confounding by indication. Epidemiology 1996;7:335–336. [PubMed: 8793355]
3. Stürmer T, Hasselbach P, Amelang M. Personality, lifestyle, and risk of cardiovascular disease and cancer: follow-up of population based cohort. Br Med J 2006;332:1359–1362. [PubMed: 16687457]
4. Glynn RJ, Schneeweiss S, Wang PS, et al. Selective prescribing led to overestimation of the benefits of lipid-lowering drugs. J Clin Epidemiol 2006;59:819–828. [PubMed: 16828675]
5. Redelmeier DA, Tan SH, Booth GL. The treatment of unrelated disorders in patients with chronic medical diseases. New Engl J Med 1998;338:1516–1520. [PubMed: 9593791]
6. Glynn RJ, Knight EL, Levin R, et al. Paradoxical relations of drug treatment with mortality in older persons. Epidemiology 2001;12:682–689. [PubMed: 11679797]
7. Writing group for the Women's Health Initiative investigators. Risk and benefits of estrogen plus progestin in healthy postmenopausal women. Principal results from the Women's Health Initiative randomized controlled trial. J Am Med Assoc 2002;288:321–333.
8. Stürmer T, Buring JE, Lee IM, et al. Colorectal cancer after start of nonsteroidal anti-inflammatory drug use. Am J Med 2006;119:494–502. [PubMed: 16750963]
9. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol Drug Saf 2006;15:291–303. [PubMed: 16447304]
10. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc 1996;91:444–455.
11. Brookhart MA, Wang PS, Solomon DH, et al. Evaluating short-term drug effects in claims databases using physician-specific prescribing preferences as an instrumental variable. Epidemiology 2006;17:268–275. [PubMed: 16617275]

12. Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika 1987;74:13–26.
13. Brumback BA, Hernan MA, Haneuse SJPA, et al. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. Stat Med 2004;23:749–767. [PubMed: 14981673]
14. MacLehose RF, Kaufman S, Kaufman JS, et al. Bounding causal effects under uncontrolled confounding using counterfactuals. Epidemiology 2005;16:548–555. [PubMed: 15951674]
15. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol 1991;133:144–153. [PubMed: 1985444]
16. Fraser GE. A search for truth in dietary epidemiology. Am J Clin Nutr 2003;78:521S–525S. [PubMed: 12936944]
17. Ray WA, Griffin MR, Fought RL, et al. Identification of fractures from computerized Medicare files. J Clin Epidemiol 1992;45:703–714. [PubMed: 1619449]
18. Phillips CV. Quantifying and reporting uncertainty from systematic errors. Epidemiology 2003;14:459–466. [PubMed: 12843772]
19. Greeenland S. Multiple-bias modelling for analysis of observational data. J R Stat Soc A 2005;168:267–306.
20. Walker, AM. Observation and Inference: An Introduction to the Methods of Epidemiology. Newton: Epidemiology Resources Inc; 1991. p. 120-124.
21. Bross IDJ. Spurious effects from an extraneous variable. J Chronic Dis 1966;19:637–647. [PubMed: 5966011]
22. Schneeweiss S, Glynn RJ, Tsai EH, et al. Adjusting for unmeasured confounders in pharmacoepidemiologic claims data using external information: the example of COX2 inhibitors and myocardial infarction. Epidemiology 2005;16:17–24. [PubMed: 15613941]
23. Little, RJA.; Rubin, DB. Statistical Analysis With Missing Data. 2. New York: John Wiley & Sons; 2002.
24. Garshick E, Laden F, Hart JE, et al. Smoking imputation and lung cancer in railroad workers exposed to diesel exhaust. Am J Ind Med 2006;49:709–718. [PubMed: 16767725]
25. Moons KG, Donders RA, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006;59:1092–1101. [PubMed: 16980150]
26. Walker AM. Anamorphic analysis. Sampling and estimation for covariate effects when both exposure and disease are known. Biometrics 1982;38:1025–1032. [PubMed: 7168792]
27. Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. Am J Epidemiol 1988;128:1198–1206. [PubMed: 3195561]
28. Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. Stat Med 1991;10:739–747. [PubMed: 2068427]
29. Zhao LP, Lipsitz S. Designs and analysis of two-stage studies. Stat Med 1992;11:769–782. [PubMed: 1594816]
30. Langholz B, Clayton D. Sampling strategies in nested case-control studies. Environ Health Perspect 1994;102(suppl 8):47–51. [PubMed: 7851330]
31. Stürmer T, Brenner H. Flexible matching strategies to increase power and efficiency to detect and estimate gene-environment interactions in case-control studies. Am J Epidemiol 2002;155:593–602. [PubMed: 11914186]
32. Hanley JA, Csizmadi I, Collet JP. Two-stage case-control studies: precision of parameter estimates and considerations in selecting sample size. Am J Epidemiol 2005;162:1225–1234. [PubMed: 16269581]
33. Walker AM, Jick H, Hunter JR, et al. Vasectomy and non-fatal myocardial infarction. Lancet 1981;317:13–15. [PubMed: 6109049]
34. Sharpe CR, Collet JP, McNutt M, et al. Nested case-control study of the effects of non-steroidal anti-inflammatory drugs on breast cancer risk and stage. Br J Cancer 2000;83:112–120. [PubMed: 10883678]

35. Collet JP, Schaubel D, Hanley J, et al. Controlling confounding when studying large pharmacoepidemiologic databases: a case study of the two-stage sampling design. Epidemiology 1998;9:309–315. [PubMed: 9583424]

36. Stürmer T, Schneeweiss S, Avorn J, et al. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am J Epidemiol 2005;162:279–289. [PubMed: 15987725]

37. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.

38. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. Am J Epidemiol 1990;132:734–45. [PubMed: 2403114]

39. Carroll, RJ.; Ruppert, D.; Stefanski, LA. Measurement Error in Nonlinear Models. London: Chapman & Hall; 1995.

40. Stürmer T, Schneeweiss S, Brookhart MA, et al. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: Nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. Am J Epidemiol 2005;161:891–898. [PubMed: 15840622]

41. Yuan Z, Cooper GS, Einstadter D, et al. The association between hospital type and mortality and length of stay. Med Care 2000;38:231–245. [PubMed: 10659696]

42. Eppig FJ, Chulis GS. Matching MCBS and Medicare data: The best of both worlds. Health Care Financing Review 1997;18:211–229. [PubMed: 10170350]

43. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. Am J Epidemiol 2006;163:1149–1156. [PubMed: 16624967]

44. Carroll RJ, Stefanski LA. Approximate quasi-likelihood estimation in models with surrogate predictors. J Am Stat Assoc 1990;85:652–63.

45. Stürmer T, Schneeweiss S, Rothman KJ, et al. Performance of propensity score calibration – a simulation study. Am J Epidemiol 2007;165:1110–18. [PubMed: 17395595]

46. Stürmer T, Schneeweiss S, Rothman KJ, et al. Stürmer et al. respond to “Advancing propensity score methods in epidemiology”. Am J Epidemiol 2007;165:1122–23.

Table 1

Classification of validation studies for external adjustment of confounding unmeasured in the main study and general notions about their availability, possible use of different analytic strategies, and possible use for multiple associations

	Validation study		
	Internal	External	
Information on disease outcome	Yes	Yes	No
Availability	Rare*	Rare	Frequent
Analytic methods			
Multiple imputation	Yes†	Yes†	No
2-stage sampling	Yes	No	No
Propensity score calibration	Yes	Yes	Yes
Validation study can be used to adjust multiple associations	Yes (if random sample)	Yes	Yes
Transportability of parameters‡	Yes	?	?

*Availability does not equal feasibility – internal validation studies might be rarely available but easily feasible in specific settings.

†Internal validation studies contain information on disease outcome of interest by design. Any particular validation study might not have sufficient information on the disease outcome of interest due to small size and/or low incidence, however, for multiple imputation to be applied.

‡Models and their relevant parameters from the validation study can be applied to the main study - this is usually given in internal validation studies but needs to be assumed (i.e., is questionable) in external validation studies.

Table 2

Concept of propensity score calibration

Study	Main Study	Validation Study	
Propensity Score Variables	**Error-Prone Claims Data**	**Error-Prone Claims Data**	**Gold-Standard Claims & Survey**
Exposure	X	X	X
Demographics	X	X	X
Diagnoses	X	X	X
Procedures	X	X	X
Prescriptions	X	X	X
Visits	X	X	X
Smoking			X
Aspirin			X
Body Mass Index			X
Education			X
Activities of Daily Living			X

Table 3

Association between non-steroidal anti-inflammatory drug use and 1-year mortality in a population-based cohort of 103,133 elderly - propensity score calibration adjustment based on data from 5,108 participants of the Medicare Current Beneficiary Survey as external cross-sectional validation study (modified from reference 29)

	Hazard Ratio*	(95% Confidence Interval)*
Unadjusted model	0.68	(0.66 – 0.71)
Conventional multivariate outcome model		
Age and gender adjusted	0.74	(0.71 – 0.77)
Fully adjusted†	0.80	(0.77 – 0.83)
Propensity score (main study) adjusted†	0.81	(0.78 – 0.84)
Propensity score calibration adjusted	1.06	(1.00 – 1.12)

*Hazard ratios and their 95% confidence intervals estimated using Cox proportional hazards regression; propensity score calibration adjusted models include uncertainty due to the estimation of the error model.

†Adjusted for age (continuous), sex, race (white, black, other), myocardial infarction, congestive heart failure, diabetes, cancer, arthritis (RA or OA), number of physician visits (0–5, 6–11, 12+), number of hospitalizations (0, 1, 2+), and use of thiazides, steroids, and anticoagulants.

Table 4

Propensity of non-steroidal anti-inflammatory drug use in the main study population and the external validation study (modified from reference 29)

	Main Study		Validation Study			
	"Error Prone" Propensity Score		"Error-Prone" Propensity Score		"Gold-Standard" Propensity Score	
	OR*	(95% CI*)	OR*	(95% CI*)	OR*	(95% CI*)
Age (1 year)	0.98	(0.98 – 0.99)	0.98	(0.97 – 1.00)	0.98	(0.97 – 1.00)
Female gender	1.2	(1.2 – 1.3)	1.2	(1.0 – 1.5)	1.1	(0.9 – 1.4)
Race						
Black	1.6	(1.5 – 1.7)	1.4	(1.1 – 1.9)	1.2	(0.9 – 1.7)
Other	2.0	(1.9 – 2.2)	1.5	(1.0 – 2.2)	1.6	(1.1 – 2.5)
Diagnoses based on claims data						
Myocardial infarction	0.9	(0.8 – 0.9)	1.1	(0.8 – 1.5)	1.1	(0.7 – 1.5)
Congestive heart failure	0.9	(0.9 – 1.0)	0.9	(0.6 – 1.3)	0.8	(0.6 – 1.3)
Diabetes	1.0	(1.0 – 1.0)	0.9	(0.6 – 1.3)	0.7	(0.5 – 1.0)
Cancer	0.8	(0.8 – 0.8)	0.6	(0.4 – 0.9)	0.6	(0.4 – 1.0)
Arthritis (RA or OA)	2.1	(2.0 – 2.2)	2.4	(1.7 – 3.4)	1.8	(1.3 – 2.5)
Diagnoses based on self report						
Arthritis (RA or OA)	-	-	-	-	4.1	(3.1 – 5.5)
Health care system use						
Number of physician visits†	1.3	(1.3 – 1.3)	1.1	(1.0 – 1.4)	1.1	(0.9 – 1.3)
Number of hospitalizations†	0.9	(0.9 – 0.9)	1.1	(0.9 – 1.3)	1.0	(0.9 – 1.2)
Medications						
Thiazides	1.3	(1.2 – 1.3)	1.6	(0.9 – 2.5)	1.5	(0.9 – 2.5)
Steroids	1.0	(0.9 – 1.0)	1.5	(1.2 – 2.0)	1.3	(1.0 – 1.8)
Anticoagulants	0.5	(0.5 – 0.6)	0.5	(0.3 – 0.8)	0.5	(0.3 – 0.7)
Body mass index (1 kg/m^2)	-	-	-	-	1.05	(1.03 – 1.06)
Education‡	-	-	-	-	1.0	(0.8 – 1.1)
Income§	-	-	-	-	1.1	(1.0 – 1.2)
Smoking						
Current	-	-	-	-	1.0	(0.7 – 1.4)
Past	-	-	-	-	1.1	(0.9 – 1.3)
Activities of daily living						
Difficulties with¶	-	-	-	-	1.2	(1.1 – 1.3)
Unable to perform¶	-	-	-	-	1.2	(1.0 – 1.3)
AUC	0.63		0.60		0.66	

*Odds ratios and their 95% confidence intervals from 3 separate multivariable logistic regression models including all variables with presented values

†1 unit increase in category of number of physician visits (0 – 5, 6 – 11, 12+) and hospitalizations (0, 1, 2+)

‡1 unit increase in category of education (up to 12th grade, high school, associate degree or more)

§1 unit increase in category of income per year (up to 10k, 10 – 20k, 20 – 40k, more than 40k US$)

¶Number of activities of daily living that were reported to be difficult and impossible to perform, respectively

OR indicates odd ratio; CI, confidence interval; AUC, Area under the receiver operating characteristic curve (c-statistic).