Published in final edited form as:
Am J Epidemiol. 2005 May 1; 161(9): 891–898.

Analytic Strategies to Adjust Confounding Using Exposure Propensity Scores and Disease Risk Scores: Nonsteroidal Antiinflammatory Drugs (NSAID) and Short-term Mortality in the Elderly

Til Stürmer[1,2], **Sebastian Schneeweiss**[1], **M. Alan Brookhart**[1], **Kenneth J Rothman**[1], **Jerry Avorn**[1], and **Robert J Glynn**[1,2]

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

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

Abstract

Little is known about optimal application and behavior of exposure propensity scores (EPS) in small studies. Based on a cohort of 103,133 elderly Medicaid beneficiaries, the effect of nonsteroidal anti-inflammatory drug (NSAID) use on 1-year all-cause mortality was assessed based on the assumption that there is no protective effect, and the preponderance of any observed effect would be confounded. To study the comparative behavior of EPS, disease risk scores (DRS), and 'traditional' disease models, we randomly re-sampled 1,000 subcohorts of 10,000, 1,000 and 500 people. The number of variables was limited in disease models, but not EPS and DRS. Estimated EPS were used to adjust for confounding by matching, inverse probability of treatment weighting (IPTW), stratification, and modeling. The crude rate ratio (RR) of death for NSAID users was 0.68. 'Traditional' adjustment resulted in a RR of 0.80 (95% confidence interval:0.77–0.84). The RR closest to 1 was achieved by IPTW (0.85;0.82–0.88). With decreasing study size, estimates remained further from the null, which was most pronounced for IPTW (N=500: RR=0.72;0.26–1.68). In this setting, analytic strategies using EPS or DRS were not generally superior to 'traditional'. Various ways to use EPS and DRS behaved differently with smaller study size.

Keywords

epidemiologic methods; research design; confounding factors (epidemiology); bias (epidemiology); cohort studies; nonsteroidal anti-inflammatory drugs

Abbreviations

AUC, area under the receiver operating characteristic curve; CI, confidence interval; EPS, exposure propensity score; DRS, disease risk score; IPTW, inverse probability of treatment weighting; NSAID, nonsteroidal antiinflammatory drug; OR, odds ratio; RR, relative risk

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

Propensity score methods (1) are increasingly used to control for confounding in non-experimental medical research (2). Propensity scores combine a large number of possible confounders into a single variable (the score). This concept of multivariate confounder scores can not only be used to account for different propensities of exposure, but also to account for different disease risks, as already noted by Miettinen in 1976 (3). To separate clearly these different scores, we will use the terms exposure propensity score (EPS) and disease risk score (DRS).

So far we know that EPS and DRS give non-nominal p-values under the null in situations with very strong, likely unrealistic exposure – confounder associations (4). Omitting an important confounder from analysis leads to similar magnitude and direction of bias when using EPS compared with outcome models (5). EPS have a reduced efficiency compared with outcome models (6). Also, EPS have been shown to perform better than outcome models if less than 8 events are observed per covariate that we want or need to control for (7). Little is known, however, about the effect of different ways of using EPS and DRS with respect to control for confounding and statistical efficiency, especially in small studies.

We illustrate the different ways both methods can be used to control for confounding in a large cohort study based on claims data, evaluating the association between nonsteroidal anti-inflammatory drug (NSAID) use and 1-year mortality in an elderly population. We chose the specific empirical example of the previously observed inverse association between NSAID and all-cause mortality (8,9) since there is no known biological reason to expect that NSAID use would cause a reduction in the risk of short-term mortality. Even if such an association existed, the observed magnitude of about 26% risk reduction (8) is implausible. Instead, the apparent association is likely a spurious one, due to patient selection leading to confounding bias: physicians are more likely to treat symptomatic pain with narcotic agents rather than NSAID in patients who are close to death (8). We also randomly re-sampled smaller sub-cohorts to assess the effect of study size on the performance of these methods.

MATERIALS AND METHODS

Exposure propensity scores

Exposure propensity scores (EPS) are defined as each subject's probability of exposure to a specific treatment given his or her observed covariates (1). The EPS function is usually estimated using a multivariable logistic regression model for the entire study population, but could be estimated with a variety of multivariable scoring functions. In a logistic model, the EPS range from 0 to 1 and reflect the estimated probability, based on the subject's characteristics, that the subject will receive the treatment of interest. Any two subjects with the same EPS can have different covariate values, but the distributions of covariates for all treated subjects should be similar to untreated subjects with the same EPS (1). Therefore, within each EPS stratum, some patients will have received the treatment of interest while others will not, although they have similar estimated probabilities of receiving treatment given their observed covariates. By estimating the EPS and estimating the exposure-disease association within homogeneous levels of the EPS, in theory and under the assumption of no unmeasured confounding, one can theoretically achieve 'virtual randomization', in which comparable patients are separated into treated and untreated groups (1).

Disease risk score

Disease risk scores (DRS) reduce the number of covariates by summarizing the predictive information for disease risk of all potential confounders in a multivariable model, conditional on non-exposure (3,4). The DRS can then be used as a summary confounder and controlled using stratification or a multivariable outcome model. Just as the purpose of randomization is

to create groups for comparison that have the same underlying risk for disease, apart from the effect of exposure, the DRS achieves this comparability based on the multivariate distribution of identified confounders.

Study population

The study population was assembled for an analysis of pain medication use in the elderly, and consists of all New Jersey residents 65 years or older who filled prescriptions (as defined below) within Medicaid or the Pharmaceutical Assistance to the Aged and Disabled (PAAD) program and were hospitalized any time between January 1, 1995 and December 31, 1997. Patients discharged after December 31, 1997 as well as those residing in a nursing home before hospitalization were excluded. For individuals with more than one such hospitalization, one was selected at random to permit valid estimation of the 1-year risk for all-cause mortality in this population at the time of a hospital admission. Eligible individuals were those who filled at least one prescription within 120 days before hospitalization and at least one prescription more than 365 days before hospitalization, since covariates were assessed during that time period. The index date was the date of hospitalization.

The exposure of interest, NSAID use, was defined as having filled at least one prescription for any NSAID during the 120 days before hospital admission. For all subjects, the following covariates were then extracted: age (in years), sex, race (Caucasian, African-American, other), and 17 diagnoses based on inpatient and outpatient visits that are part of the Charlson comorbidity index (10) within 365 days before the index date (AIDS, congestive heart failure, chronic obstructive lung disease (COPD), dementia, hematologic disease, cancer, metastatic cancer, myocardial infarction, diabetes (with and without complications), liver disease (mild and severe), peripheral vascular disease, peptic ulcer, renal disease, arthritis (rheumatoid arthritis or osteoarthritis), and stroke). Further covariates were indicators for prescriptions of distinct generic entities filled within 120 days before the index date, including narcotics, other analgesics, ACE inhibitors, beta blockers, calcium-channel blockers, thiazide diuretics, other antihypertensives, lipid lowering drugs, antiarrhythmics, coumadin, digitalis, rheologic drugs, oral antidiabetics, insulin, anti-diarrheals, H2 blockers, other anti-ulcer drugs, anticonvulsants, beta-agonists, xanthines, steroids, other bronchodilators, loop diuretics, potassium, anxiolytics, antidepressants, phenobarbital, other antipsychotics, sedatives, stimulants, penicillins, cephalosporins, macrolide antibiotics, quinolones, sulfonamides, folic acid, influenza vaccines, glaucoma drugs, topical antibiotics, topical sulfonamides, topical enzymes. We computed the number of hospitalizations (3 categories), number of physician visits (3 categories), and screening examinations, including cholesterol, ECG, mammography, Pap-smear, prostate specific antigen during that 1-year time interval. All 71 covariates were available for inclusion in the analyses. In the absence of a record for a specific diagnosis, procedure, or prescription patients were coded as free of these characteristics. As a result of this coding rule, there were no subjects for whom exposure, confounder, or outcome information was missing (with the exception of unknown race which was classified as other than white or black).

We assessed time until death or study end at 365 days of follow-up (whichever came first), starting from the date of hospital admission, based on exact linkage to Medicare claims data (11). The study was approved by the Center for Medicare and Medicaid Services and the Institutional Review Board of the Brigham and Women's Hospital.

Analytic strategies

EPS stratification, modeling, and matching—We estimated the EPS of NSAID use (yes/no) during the last four months before hospitalization using logistic regression and a forward variable selection with an alpha-value of 0.3. The value of 0.3 was less stringent than

the value of 0.2 used in the 'traditional' disease models to allow more variables to be entered into the EPS models compared with the 'traditional' models. The estimated EPS was used in several ways. We first adjusted the multivariate Cox proportional hazards outcome model including the EPS as categories (quintiles), as linear splines (12) or as a linear predictor (continuous). Second, we matched every exposed participant to one unexposed participant on the EPS (1:1 matching) and used a stratified Cox proportional hazards model in the matched sample to control for confounding. We used two different matching algorithms: greedy matching, using calipers of the estimated EPS with increasing width to find matches (13), and fixed 1-digit matching, using a fixed width of 0.1 (i.e. matching on EPS +/− 0.05).

Inverse probability of treatment weighting—Inverse probability of treatment weighting (IPTW) uses the estimated EPS to assign individual weights to all observations resulting in an altered composition of the study population (14). The altered study population is then analyzed using a Cox proportional hazards model with NSAID use as the only covariate (15). We used 'stabilized' weights that take the marginal prevalence of the exposure into account to maximize efficiency and to obtain a re-weighted study population of equal size (14,15). The weights for exposed participants are obtained by dividing P, the marginal prevalence of exposure, by the individual EPS, and those in unexposed by dividing (1−P) by (1−EPS).

'Traditional' multivariable disease model—In addition to an unadjusted estimate of the association between NSAID use and short-term mortality, we also used a 'traditional' multivariable Cox proportional hazards model to adjust for confounding. Variables were included by forward selection using an alpha-value of 0.2, a value that has been shown to perform well with respect to control for confounding (16). To avoid overfitting, the number of variables was restricted to have at least 8 outcomes per variable in the model (7).

Disease risk score—We then estimated the DRS of 1-year mortality from all causes using a Cox proportional hazards model and a forward variable selection with an alpha-value of 0.3, including all 71 covariates described above at the beginning of the selection algorithm as well as the primary exposure, NSAID use. The value of 0.3 was again chosen to allow more variables to be entered into the DRS model than into the disease model. For the same reason, the overall number of variables was not restricted. The regression coefficients from this model were then multiplied by the individual covariate values of the variables entered into the model, except for NSAID use, which was set to 0 (non-use) for all participants (3,4). The sum of these products gave the subject-specific DRS, which was then used to control for confounding in separate Cox proportional hazards models of the study outcome. We included the DRS as categories (5 or 10) or as a continuous linear predictor together with the primary exposure, NSAID use.

Combination of methods—Finally, we combined some of these methods by simultaneously adjusting for EPS and DRS and by adding a selection of risk indicators to the EPS (again using forward variable selection). Although this ad hoc approach is not equivalent to 'doubly robust' estimation (17) it might nevertheless offer advantages if either the EPS model or the 'traditional' disease model are misspecified.

Random re-sampling of sub-cohorts

From the total cohort of 103,133 elderly, we created 1,000 randomly sampled subcohorts of 10,000, 1,000, or 500 persons, with replacement, and applied the analytic strategies described above within each re-sampled sub-cohort (3,000 cohorts overall). Using this approach, we obtained the empirical distribution of the compositions of the cohorts and selected model characteristics as well as parameter estimates for each of the 13 analytic strategies.

RESULTS

In table 1 we describe the study population of 103,133 hospitalized elderly. The mean age was 79 years and 75 percent were women. The most prevalent comorbidity was congestive heart failure (33%), followed by diabetes (30%), and cancer (17%). During the year preceding the hospitalization, the tertiles for the number of physician visits were 6 and 12, and 16% were hospitalized at least twice. During the 4 months before hospitalization, 18,326 elderly had at least 1 prescription of an NSAID (18%).

In table 2 we present the number of NSAID users, the number of deaths, as well as various characteristics of the models used in the different analytic strategies for the full cohort, and the distribution of these values for the re-sampled sub-cohorts. During the 1-year follow-up period, over 20% hospitalized patients died, either during hospitalization or afterwards (21,928 in the full cohort), a similar proportion than the overall proportion of NSAID use (18%).

Effect of analytic strategy and study size on model specification

In all scenarios (the full cohort, and the sub-cohorts of size 10,000, 1,000, and 500), more variables were independent predictors of the outcome than of the exposure. As a result, the DRS involved more covariates than the corresponding EPS (see table 2). In the full cohort, forward variable selection resulted in almost the same number of covariates being included in the EPS, the 'traditional' disease model, and in the model combining the EPS and risk indicators. Owing to our limits for the maximum number of covariates in the disease models, the number of covariates included in these models was smaller for smaller study sizes. In the N=500 sub-cohort, for example, only 12 covariates were included in the outcome models, compared with 26 in the EPS and 30 in the DRS models.

Despite the decreasing number of covariates used to estimate the EPS with decreasing size of the sub-cohorts, the median area under the receiver operating characteristic curve (AUC), estimating the ability of the EPS model to discriminate exposure status (18), increased from 0.68 (N=10,000) to 0.79 (N=500). The AUC can range from 0.5 (chance prediction) to 1.0 (perfect prediction).

In the full cohort, 99.6% of all NSAID users could be matched to non-users when we used either greedy matching or fixed width caliper 1-digit matching. Despite this large proportion, both matching strategies resulted in a loss of over 70% of all events (71.2 and 70.9%, respectively), because a large proportion of unexposed were not included. With decreasing size of the sub-cohorts, the proportion of successfully matched subjects decreased. Taking the decreasing absolute number of events with decreasing size of the sub-cohorts into account, the decline in number of events on which final analyses were based was even more pronounced. These results were essentially the same for both matching techniques.

Effect of analytic strategy and study size on NSAID effect estimates

Table 3 describes the association between NSAID use and 1-year mortality from Cox proportional hazards models using various approaches to control for confounding. Without any control for confounding, NSAID use appeared to be associated with an over 30% mortality risk reduction. With decreasing size of the sub-cohorts, this estimate remained stable, while the empirical 95% confidence interval got wider.

The unadjusted NSAID association was 0.68, a 32% mortality risk reduction, an effect that should be nearly all due to uncontrolled confounding. Every increase in the effect estimate from the unadjusted RR towards an RR of 1.0 can therefore be interpreted as an improved adjustment for confounding. Controlling for age and sex had only a minor effect on these estimates. Using up to 71 covariates, all the analytic strategies resulted in estimates ranging

from 0.85 for the EPS greedy matched and the IPTW analyses in the full cohort to 0.72 for the latter analysis in the sub-cohort of N=500.

Using the EPS in various ways to adjust for confounding in the full cohort resulted in estimates for the NSAID-mortality association ranging from 0.81 (quintiles) to 0.83 (continuous). The same point estimate was obtained when using fixed 1-digit matching, whereas the point estimate using greedy matching was slightly closer to the null (0.85). Owing to the loss of information, both matched estimates were slightly less precise. Using IPTW based on the estimated EPS also resulted in a point estimate of 0.85. All outcome models, including the 'traditional' model, all DRS models and the model including the EPS in combination with risk indicators, resulted in essentially identical estimates between 0.80 and 0.81.

Results from sub-cohorts with N=10,000 were nearly identical to those from the full cohort (with the exception of slightly wider, now empirical, confidence intervals). With decreasing size of the sub-cohorts, however, estimates from all analytic strategies moved further away from the null and closer to the crude, indicating increasing residual confounding. Despite the fact that the number of variables used in the corresponding models decreased much more sharply with decreasing study size in the outcome models than in the EPS and DRS, this decrease was not reflected in any substantial differences between these strategies. Specifically, the use of DRS did not translate into any gain compared with the 'traditional' outcome model, neither with respect to point estimate, nor precision (results stratifying on 10 rather than 5 categories of the DRS were virtually identical and are therefore not presented). Taking the absence of major differences into account, greedy matching resulted in the estimate closest to the null (0.80) and IPTW resulted in an estimate (0.72) very close to the age and sex adjusted estimate (0.71) in the smallest sub-cohort. Both analyses had wide confidence intervals compared with the other analytic strategies.

DISCUSSION

We observed no major difference between different analytic techniques and applications of these techniques in this particular setting. Specifically, neither EPS nor DRS was superior to traditional multivariable modeling in small studies with a limited number of outcomes, as was hypothesized earlier (19).

Since the use of EPS (but not DRS) comes at the price of losing potentially useful information about predictors of the outcome (since the covariates are not included in the disease model), and we furthermore know much less about variables selection and model building strategies for EPS compared to 'traditional' disease models (16), it seems desirable to use EPS only if a reduction in bias or an improvement in efficiency could be achieved.

Cook and Goldman compared the performance of tests of significance under the null hypothesis (i.e. assuming no difference between treatments) for EPS, DRS, and for 'traditional' multivariable outcome models using simulations (4). EPS appeared to produce nominal results in most circumstances, but not in situations with very strong treatment - confounder associations. This result was even more pronounced for DRS. Such a constellation was not present in our realistic example.

In some practical situations the choice of analytic method will be limited. Because 10 events per covariate are usually considered to be a minimum requirement for stable estimates in multivariable models (18,20), EPS analyses combining multiple covariates into a single score are especially desirable if the outcome is rare (19,21). A recent simulation study comparing EPS with multivariable outcome models concluded that EPS performed better in situations with less than 8 outcomes per covariate (7). We therefore used this proportion to limit the maximum number of variables available for selection in all our outcome models. Since

precision of estimates is likely to be of minor importance for EPS and DRS, we used a value of alpha = 0.3 in these models to allow more variables to be entered than in the disease model, for which a value of 0.2 has been shown to perform well for control of confounding (16). Neither the restriction of the absolute number of variables nor the more stringent p-value requirement handicapped the performance of 'traditional' disease models compared with EPS or DRS.

We chose the example of NSAID and all-cause mortality since NSAID are unlikely to affect mortality substantially in an elderly population (8). The Physicians' Health Study trial reported an equal number of deaths in both the aspirin as well as the placebo arm after 5 years of (low dose) treatment (22), although the overall number of deaths was too small to rule out a meaningful difference. Even if NSAID were protective for some cancers, including colorectal cancer (23,24), chronic use in the elderly is associated with several adverse outcomes, including increased risk of gastrointestinal hemorrhage (25–28), impaired kidney function (29–31), hypertension (32,33), and perhaps even cardiovascular disease, stemming from their possible antagonism of the preventive effect of aspirin (34) (since cohort enrollment ended in 1997, NSAID use does not include COX-2 inhibitors). Therefore, either no association with mortality or if anything a slightly increased risk of mortality seems biologically more plausible than a reduced risk of death.

Glynn et al. have argued that in an elderly population selected drug classes, including lipid lowering drugs, NSAID, or anti-glaucoma drugs, are more likely to be prescribed to healthier subjects (8). Drugs with a preventive component are less likely to be prescribed if death seems near based on the assessment of the prescribing physician (9). Thus, even if we do not know the precise relation between NSAID and mortality, our conclusions are valid for a wide range of possible effects including a reduction of up to15% in risk. More pronounced risk reductions in mortality from all causes seem biologically implausible. The generally similar estimates resulting from the application of the various analytic strategies might be an indication that we were able to control adequately for observed confounding, although we do not know what the best estimate of the NSAID-mortality association given the observed covariates would be.

Our results are limited to one specific setting with essentially the same prevalence of exposure and cumulative incidence of disease (around 20% each). This restriction might explain why we did not see differences between the EPS and the DRS. It would not explain, however, why we did not observe any of these methods that combine multiple variables into a single score to perform better than 'traditional' disease models. Generally speaking, the data structure at hand is likely to influence the choice of the preferred method. EPS are likely to perform better than 'traditional' outcome models or DRS with respect to control for confounding when the exposure is prevalent and the disease is rare since it may be possible to build a richer model of the exposure than of the disease and vice versa (21). We suppose that DRS might have an advantage over 'traditional' outcome models with respect to bias and precision if the disease is rare, because 1) they allow the truncation of the risk score distribution so that only the range of scores that is common to both exposed and unexposed is included in the analysis; and 2) the final disease model can be fit with only two variables (i.e. the exposure of interest and the risk score).

The set of variables available in this claims database might not be broad enough to predict exposure or outcome sufficiently. This limitation would only invalidate our comparisons, however, if a small subset of variables included in all models were responsible for all the confounding with additional variables showing no confounding above and beyond these 'core' confounders. It is nevertheless intriguing that differences between the methods are minor compared with the remaining residual confounding, assuming that there is no protective effect of NSAID on short-term all-cause mortality. Incorporating additional information on factors

strongly associated with the prescription of NSAID and short-term mortality not available in claims data (e.g. measures of over- and underweight or activities of daily living, 35, 36) seems more promising than using different strategies to analyze the available data in our specific example.

The parameters estimated using individual matching on the EPS and IPTW are not exactly the same as the ones from the other analytic strategies (37,38). The population-averaged interpretation of these estimates might explain why they are closest to the null in the full cohort, but not why the IPTW estimate seems furthest away from the null in the smallest studies. The latter might be due to influential weights attributed to observations with 'wrong' exposure status, i.e. exposed observations with a very low estimated propensity for exposure or vice versa.

We conclude that in the setting of claims data of an elderly population, various ways to apply exposure propensity scores and disease risk scores to control for confounding were not generally superior to 'traditional' multivariable outcome modeling. Differences in effect estimates between analytic strategies became more pronounced with smaller study size.

References

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
2. Stürmer T, Schneeweiss S, Avorn J, et al. Determinants of use and application of propensity score (PS) methods in pharmacoepidemiology. Pharmacoepidemiol Drug Saf 2003;12(Suppl 1):S121.
3. Miettinen OS. Stratification by a multivariate confounder score. Am J Epidemiol 1976;104:609–20. [PubMed: 998608]
4. Cook EF, Goldman L. Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. J Clin Epidemiol 1989;42 :317–324. [PubMed: 2723692]
5. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics 1993;49:1231–1236.
6. Drake C, Fisher L. Prognostic models and the propensity score. Int J Epidemiol 1995;24:183–187. [PubMed: 7797341]
7. Cepeda MS, Boston R, Farrar JT, et al. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol 2003;158:280–7. [PubMed: 12882951]
8. Glynn RJ, Knight EL, Levin R, et al. Paradoxical relations of drug treatment with mortality in older persons. Epidemiology 2001;12:682–9. [PubMed: 11679797]
9. Redelmeier DA, Tan SH, Booth GL. The treatment of unrelated disorders in patients with chronic medical diseases. New Engl J Med 1998;338:1516–20. [PubMed: 9593791]
10. Schneeweiss S, Seeger JD, Maclure M, et al. Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. Am J Epidemiol 2001;154:854–64. [PubMed: 11682368]
11. Yuan Z, Cooper GS, Einstadter D, et al. The association between hospital type and mortality and length of stay. Med Care 2000;38:231–45. [PubMed: 10659696]
12. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology 1995;6:356–65. [PubMed: 7548341]
13. Parsons LS. Reducing bias in a propensity score matched-pair sample using greedy matching techniques. 2001; http://www2.sas.com/proceedings/sugi26/p214–26.pdf
14. Robins JM. Marginal structural models. Proceedings of the American Statistical Association 1997. Section on Bayesian Statistical Science, pp. 1–10; http://www.biostat.harvard.edu/%7Erobins/msm-web.pdf
15. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550–60. [PubMed: 10955408]

16. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. Am J Epidemiol 1993;138:923–36. [PubMed: 8256780]
17. Robins JM, Rotnitzky A. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". Statistica Sinica 2001;11:920–936.
18. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models. Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87. [PubMed: 8668867]
19. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. Ann Intern Med 2002;137:693–5. [PubMed: 12379071]
20. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373–9. [PubMed: 8970487]
21. Cepeda MS. The use of propensity scores in pharmacoepidemiologic research. Pharmacoepidemiol Drug Saf 2000;9:103–4.
22. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. N Engl J Med 1989;321:129–35. [PubMed: 2664509]
23. Chan TA. Nonsteroidal anti-inflammatory drugs, apoptosis, and colon cancer chemoprevention. Lancet Oncol 2002;3:166–74. [PubMed: 11902503]
24. Peleg II, Wilcox CM. The role of eicosanoids, cyclooxigenases, and nonsteroidal anti-inflammatory drugs in colorectal tumorigenesis and chemoprevention. J Clin Gastroenterol 2002;34:117–25. [PubMed: 11782603]
25. Garcia Rodriguez LA, Hernandez-Diaz S. Relative risk of upper gastrointestinal complications among users of acetaminophen and nonsteroidal anti-inflammatory drugs. Epidemiology 2001;12:570–6. [PubMed: 11505178]
26. Hernandez-Diaz S, Garcia Rodriguez LA. Epidemiologic assessment of the safety of conventional nonsteroidal anti-inflammatory drugs. Am J Med 2001;110 (Suppl 3A):20S–7. [PubMed: 11173046]
27. Ofman JJ, MacLean CH, Straus WL, et al. A metaanalysis of severe upper gastrointestinal complications of nonsteroidal antiinflammatory drugs. J Rheumatol 2002;29:804–12. [PubMed: 11950025]
28. Solomon DH, Glynn RJ, Bohn R, et al. The hidden cost of nonselective nonsteroidal anti-inflammatory drugs in older patients. J Rheumatol 2003;30:792–8. [PubMed: 12672201]
29. Gurwitz JH, Avorn J, Ross-Degnan D, et al. Nonsteroidal anti-inflammatory drug-associated azotemia in the very old. J Am Med Assoc 1990;264:471–5.
30. Field TS, Gurwitz JH, Glynn RJ, et al. The renal effects of nonsteroidal anti-inflammatory drugs in old people: findings from the Established Populations for Epidemiologic Studies of the Elderly. J Am Geriatr Soc 1999;47:507–11. [PubMed: 10323640]
31. Stürmer T, Erb A, Keller F, et al. Determinants of impaired renal function with use of nonsteroidal anti-inflammatory drugs: the importance of half-life and other medications. Am J Med 2001;111:521–7. [PubMed: 11705427]
32. Gurwitz JH, Avorn J, Bohn RL, et al. Initiation of antihypertensive treatment during nonsteroidal anti-inflammatory drug therapy. JAMA 1994;272:781–6. [PubMed: 8078142]
33. Dedier J, Stampfer MJ, Hankinson SE, et al. Nonnarcotic analgesic use and the risk of hypertension in US women. Hypertension 2002;40:604–8. [PubMed: 12411450]
34. Kurth T, Glynn RG, Walker AM, et al. Inhibition of clinical benefits of aspirin on first myocardial infarction by nonsteroidal antiinflammatory drugs. Circulation 2003;108:1191–5. [PubMed: 12939216]
35. Stürmer T, Spiegelman D, Schneeweiss S, Avorn J, Glynn RJ. Correcting effect estimates for unmeasured confounding in cohort studies with validation data using propensity score calibration. Am J Epidemiol (in press)
36. Schneeweiss S, Glynn RJ, Tsai EH, Avorn J, Solomon DH. Assessment of bias by unmeasured confounders in pharmacoepidemiologic claims data studies using external data. Epidemiology 2004 (in press).
37. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? Control Clin Trials 1998;19:249–56. [PubMed: 9620808]

38. Johnston SC, Henneman T, McCulloch CE, et al. Modeling treatment effects on binary outcomes with grouped-treatment variables and individual covariates. Am J Epidemiol 2002;156:753–60. [PubMed: 12370164]

Table 1

Description of the study population of 103,133 elderly

Age (years), mean (SD)	78.8	(7.6)
Female gender, N (%)	76,782	(75)
Race, N (%)		
White	82,039	(80)
Black	13,720	(13)
Other	7374	(7)
Diagnoses based on claims data, N (%)		
Myocardial infarction	10,319	(10)
Congestive heart failure	34,466	(33)
Diabetes	31,164	(30)
Cancer	17,364	(17)
Arthritis (RA or OA)	4,846	(5)
Health care system use, N (%)		
Number of physician visits		
0 – 5	38,924	(38)
6 – 11	33,450	(32)
12+	30,759	(30)
Number of hospitalizations		
0	64,703	(63)
1	22,293	(22)
2+	16,137	(16)
Medications		
NSAID	18,326	(18)
Thiazides	6,377	(6)
Steroids	7,848	(8)
Anticoagulants	7,613	(7)

Table 2

Cohort compositions and selected model characteristics according to analytic strategy and study size

Cohort size	Full cohort	Re-sampled sub-cohorts					
	103,133	10,000		1,000		500	
		M	(2.5th – 97.5th)*	M	(2.5th – 97.5th)*	M	(2.5th – 97.5th)*
Number of NSAID users	18,296	1772	(1701–1849)	176	(153 – 198)	88	(72 – 105)
Number of deaths	21,928	2127	(2053–2201)	212	(193 – 239)	106	(90 – 125)
Number of variables in models:							
Exposure propensity score (EPS)	55	42	(35 – 49)	28	(20 – 37)	26	(18 – 36)
Disease risk score (DRS)	65	45	(39 – 52)	31	(24 – 39)	30	(22 – 40)
'Traditional' outcome model	63	40	(34 – 46)	24	(18 – 28)	12	(10 – 14)
EPS & risk indicators	65	41	(34 – 47)	24	(18 – 28)	12	(10 – 14)
Area under ROC curve (c-statistic) EPS	0.67	0.68	(0.67 – 0.69)	0.74	(0.70 – 0.78)	0.79	(0.73 – 0.85)
Success of matching on EPS							
Greedy matching							
Percent of exposed matched to unexposed	99.6	98.2	(97.1 – 99.0)	86.9	(79.4 – 93.7)	73.2	(59.9 – 85.6)
Percent of outcomes used in analyses	28.8	28.7	(26.4 – 30.9)	25.5	(18.7 – 32.9)	21.4	(12.8 – 31.0)
Fixed 1-digit matching							
Percent of exposed matched to unexposed	99.6	98.2	(97.1 – 99.0)	86.8	(79.5 – 93.6)	73.1	(59.9 – 85.5)
Percent of outcomes used in analyses	29.1	29.0	(26.8 – 31.4)	26.1	(19.5 – 33.4)	21.7	(12.5 – 31.7)

*Median (2.5th and 97.5th percentiles) of values from 1,000 sub-cohorts re-sampled at random from full cohort of 103,133 with replacement

Table 3

Association between NSAID use and 1-year mortality in a population-based cohort of hospitalized elderly according to analytic strategy and study size

Cohort size	Full cohort		Re-sampled sub-cohorts					
	103,133		10,000		1,000		500	
	RR*	(95% CI)*	RR†	(95% CI)†	RR†	(95% CI)†	RR†	(95% CI)†
Unadjusted	0.68	(0.66 – 0.71)	0.68	(0.60 – 0.77)	0.68	(0.43 – 0.98)	0.67	(0.35 – 1.11)
Age and sex adjusted	0.72	(0.69 – 0.75)	0.72	(0.63 – 0.82)	0.71	(0.45 – 1.05)	0.71	(0.36 – 1.23)
'Traditional' outcome model	0.80	(0.77 – 0.84)	0.80	(0.71 – 0.91)	0.77	(0.47 – 1.18)	0.75	(0.35 – 1.37)
Exposure propensity score (EPS) adjusted								
Quintiles	0.81	(0.78 – 0.84)	0.81	(0.72 – 0.91)	0.78	(0.50 – 1.15)	0.78	(0.39 – 1.31)
Linear splines	0.83	(0.79 – 0.86)	0.82	(0.73 – 0.93)	0.79	(0.51 – 1.17)	0.78	(0.38 – 1.35)
Continuous	0.83	(0.80 – 0.86)	0.83	(0.73 – 0.93)	0.80	(0.52 – 1.17)	0.79	(0.39 – 1.33)
Exposure propensity score matched								
Greedy	0.85	(0.80 – 0.89)	0.82	(0.70 – 0.97)	0.81	(0.46 – 1.41)	0.80	(0.29 – 1.80)
Fixed 1-digit	0.83	(0.79 – 0.87)	0.80	(0.69 – 0.95)	0.79	(0.44 – 1.30)	0.75	(0.27 – 1.83)
Inverse probability of exposure weighted	0.85	(0.82 – 0.88)	0.84	(0.73 – 0.96)	0.79	(0.44 – 1.36)	0.72	(0.26 – 1.68)
Disease risk score (DRS) adjusted								
Quintiles	0.81	(0.77 – 0.84)	0.80	(0.71 – 0.90)	0.77	(0.48 – 1.18)	0.75	(0.36 – 1.46)
Continuous	0.80	(0.77 – 0.84)	0.80	(0.71 – 0.91)	0.77	(0.48 – 1.21)	0.76	(0.32 – 1.57)
Combined strategies ("doubly robust")								
EPS & DRS	0.81	(0.78 – 0.84)	0.81	(0.71 – 0.92)	0.79	(0.47 – 1.21)	0.77	(0.31 – 1.55)
EPS & risk indicators	0.81	(0.78 – 0.84)	0.81	(0.71 – 0.92)	0.78	(0.48 – 1.22)	0.78	(0.32 – 1.56)

* relative risks and their 95% confidence intervals from Cox proportional hazards model (age and sex adjusted: Mantel-Haenszel estimate)

† Median (2.5^{th} and 97.5^{th} percentiles) of relative risk estimates from 1,000 cohorts re-sampled at random from full cohort with replacement