



---

## PRACTICE OF EPIDEMIOLOGY

---

### Analytic Strategies to Adjust Confounding using Exposure Propensity Scores and Disease Risk Scores: Nonsteroidal Antiinflammatory Drugs and Short-term Mortality in the Elderly

Til Stürmer<sup>1,2</sup>, Sebastian Schneeweiss<sup>1</sup>, M. Alan Brookhart<sup>1</sup>, Kenneth J. Rothman<sup>1</sup>, Jerry Avorn<sup>1</sup>, and Robert J. Glynn<sup>1,2</sup>

<sup>1</sup> Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA.

<sup>2</sup> Division of Preventive Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA.

Received for publication July 29, 2004; accepted for publication December 14, 2004.

---

Little is known about optimal application and behavior of exposure propensity scores (EPS) in small studies. In a cohort of 103,133 elderly Medicaid beneficiaries in New Jersey, the effect of nonsteroidal antiinflammatory drug use on 1-year all-cause mortality was assessed (1995–1997) based on the assumption that there is no protective effect and that the preponderance of any observed effect would be confounded. To study the comparative behavior of EPS, disease risk scores, and “conventional” disease models, the authors randomly resampled 1,000 subcohorts of 10,000, 1,000, and 500 persons. The number of variables was limited in disease models, but not EPS and disease risk scores. Estimated EPS were used to adjust for confounding by matching, inverse probability of treatment weighting, stratification, and modeling. The crude rate ratio of death was 0.68 for users of nonsteroidal antiinflammatory drugs. “Conventional” adjustment resulted in a rate ratio of 0.80 (95% confidence interval: 0.77, 0.84). The rate ratio closest to 1 (0.85) was achieved by inverse probability of treatment weighting (95% confidence interval: 0.82, 0.88). With decreasing study size, estimates remained further from the null value, which was most pronounced for inverse probability of treatment weighting ( $n = 500$ : rate ratio = 0.72, 95% confidence interval: 0.26, 1.68). In this setting, analytic strategies using EPS or disease risk scores were not generally superior to “conventional” models. Various ways to use EPS and disease risk scores behaved differently with smaller study size.

anti-inflammatory agents, non-steroidal; bias (epidemiology); cohort studies; confounding factors (epidemiology); epidemiologic methods; research design

---

Abbreviations: DRS, disease risk score; EPS, exposure propensity score; IPTW, inverse probability of treatment weighting; NSAID, nonsteroidal antiinflammatory drug.

---

Propensity score methods (1) are increasingly used to control for confounding in nonexperimental 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 be used to account for not only different propensities of exposure but also different

disease risks, as noted by Miettinen in 1976 (3). To clearly separate these different scores, we use the terms exposure propensity score (EPS) and disease risk score (DRS).

So far, we know that EPSs and DRSs give nonnominal  $p$  values under the null hypothesis in situations in which exposure-confounder associations are very strong and likely

---

Correspondence to Dr. Til Stürmer, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120 (e-mail: til.sturmer@post.harvard.edu).

unrealistic (4). Omitting an important confounder from analysis leads to a similar magnitude and direction of bias when using EPSs compared with outcome models (5). EPSs have a reduced efficiency compared with outcome models (6). EPSs have been shown to perform better than outcome models if fewer than eight 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 EPSs and DRSs with respect to control for confounding and statistical efficiency, especially in small studies.

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

## MATERIALS AND METHODS

### Exposure propensity scores

EPSs 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 by using a multivariable logistic regression model for the entire study population, but it could be estimated with a variety of multivariable scoring functions. In a logistic model, the EPSs 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 those for 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 achieve "virtual randomization," in which comparable patients are separated into treated and untreated groups (1).

### Disease risk scores

DRSs reduce the number of covariates by summarizing the predictive information for disease risk of all potential

confounders in a multivariable model, conditional on nonexposure (3, 4). The DRS can then be used as a summary confounder and can be controlled for by 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 of 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 to analyze pain medication use by the elderly, and it consists of all New Jersey residents aged 65 years or older who filled prescriptions (as defined below) through Medicaid or the Pharmaceutical Assistance to the Aged and Disabled 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 persons with more than one such hospitalization, one hospitalization was selected at random to permit valid estimation of the 1-year risk of all-cause mortality in this population at the time of a hospital admission. Eligible persons 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 (acquired immunodeficiency syndrome, congestive heart failure, chronic obstructive pulmonary disease, 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 those for narcotics, other analgesics, angiotensin-converting enzyme inhibitors, beta blockers, calcium channel blockers, thiazide diuretics, other antihypertensives, lipid-lowering drugs, antiarrhythmics, coumadin, digitalis, rheologic drugs, oral antidiabetics, insulin, antidiarrheals, histamine<sub>2</sub> receptor blockers, other antiulcer 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, and topical enzymes.

We computed number of hospitalizations (three categories), number of physician visits (three categories), and

screening examinations, including cholesterol, electrocardiogram, mammography, Papanicolaou smear, and 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 for NSAID use (yes/no) during the last 4 months before hospitalization by 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 "conventional" disease models to allow more variables to be entered into the EPS models compared with the "conventional" models. The estimated EPS was used in several ways. We first adjusted the multivariate Cox proportional hazards outcome model including the EPSs 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: 1) greedy matching, using calipers of the estimated EPS with increasing width to find matches (13); and 2) fixed one-digit matching, using a fixed width of 0.1 (i.e., matching on  $EPS \pm 0.05$ ).

**Inverse probability of treatment weighting (IPTW).** This strategy 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 by 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 participants by dividing  $(1 - P)$  by  $(1 - EPS)$ .

**"Conventional" multivariable disease model.** In addition to an unadjusted estimate of the association between NSAID use and short-term mortality, we also used a "conventional" multivariable Cox proportional hazards model to adjust for confounding. Variables were included by forward selection by 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 allow at least eight outcomes per variable in the model (7).

**Disease risk score.** We then estimated the DRS for 1-year mortality from all causes by 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 (nonuse) 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 DRSs 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 "conventional" disease model is misspecified.

### Random resampling of subcohorts

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 resampled subcohort (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

Table 1 describes 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 percent), followed by diabetes (30 percent) and cancer (17 percent). During the year preceding the hospitalization, the tertiles for number of physician visits were 6 and 12, and 16 percent of the elderly were hospitalized at least twice. During the 4 months before hospitalization, 18,326 elderly were given at least one prescription of an NSAID (18 percent).

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 resampled subcohorts. During the 1-year follow-up period, more than 20 percent hospitalized patients died either during hospitalization or afterwards (21,928 in the full

**TABLE 1. Description of the study population of 103,133 elderly, New Jersey, 1995–1997**

	No.	%
Age in years (mean (SD*))	78.8 (7.6)	
Female gender	76,782	75
Race		
White	82,039	80
Black	13,720	13
Other	7,374	7
Diagnosis based on claims data		
Myocardial infarction	10,319	10
Congestive heart failure	34,466	33
Diabetes	31,164	30
Cancer	17,364	17
Arthritis†	4,846	5
Health care system use		
Physician visits (no.)		
0–5	38,924	38
6–11	33,450	32
≥12	30,759	30
Hospitalizations (no.)		
0	64,703	63
1	22,293	22
≥2	16,137	16
Medications		
NSAIDs*	18,326	18
Thiazides	6,377	6
Steroids	7,848	8
Anticoagulants	7,613	7

\* SD, standard deviation; NSAID, nonsteroidal antiinflammatory drug.

† Rheumatoid arthritis or osteoarthritis.

cohort), a proportion similar to the overall proportion of NSAID use (18 percent).

### Effect of analytic strategy and study size on model specification

In all scenarios (the full cohort and the subcohorts 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 (table 2). In the full cohort, forward variable selection resulted in almost the same number of covariates being included in the EPS, the “conventional” disease model, and 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  $n = 500$  subcohort, 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 subcohorts, the median area under the receiver operating characteristic curve, which estimates the ability of the EPS model to discriminate exposure status (18), increased from 0.68 ( $n = 10,000$ ) to 0.79 ( $n = 500$ ). This area can range from 0.5 (chance prediction) to 1.0 (perfect prediction).

In the full cohort, 99.6 percent of all NSAID users could be matched to nonusers when we used either greedy matching or fixed-width-caliper, one-digit matching. Despite this large proportion, both matching strategies resulted in a loss of over 70 percent of all events (71.2 percent and 70.9 percent, respectively) because a large proportion of those unexposed were not included. With decreasing size of the subcohorts, the proportion of successfully matched exposed subjects decreased. When we considered the decreasing absolute number of events with decreasing size of the subcohorts, the decline in the 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 derived from Cox proportional hazards models using various approaches to control for confounding. Without any control for confounding, NSAID use appeared to be associated with more than a 30 percent reduction in mortality risk. With decreasing size of the subcohorts, this estimate remained stable while the empirical 95 percent confidence interval became wider.

The unadjusted NSAID-mortality association was 0.68, a 32 percent reduction in mortality risk—an effect that should be nearly all due to uncontrolled confounding. Every increase in the effect estimate from the unadjusted rate ratio toward a rate ratio 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. When as many as 71 covariates were used, all of 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 subcohort 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 fixed one-digit matching was used, whereas the point estimate using greedy matching was slightly closer to the null value (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 “conventional” model, all DRS models, and the model including the EPS in combination with risk indicators—resulted in essentially identical estimates of between 0.80 and 0.81.

**TABLE 2. Cohort compositions and selected model characteristics according to analytic strategy and study size, New Jersey, 1995–1997**

	Full cohort (no.) ( <i>n</i> = 103,133)	Resampled subcohorts (cohort size)					
		10,000		1,000		500	
		M*	2.5th, 97.5th*	M*	2.5th, 97.5th*	M*	2.5th, 97.5th*
No. of NSAID† users	18,296	1,772	1,701, 1,849	176	153, 198	88	72, 105
No. of deaths	21,928	2,127	2,053, 2,201	212	193, 239	106	90, 125
No. of variables in the models							
EPS†	55	42	35, 49	28	20, 37	26	18, 36
DRS†	65	45	39, 52	31	24, 39	30	22, 40
“Conventional” outcome model	63	40	34, 46	24	18, 28	12	10, 14
EPS and risk indicators	65	41	34, 47	24	18, 28	12	10, 14
Area under the ROC† curve ( <i>c</i> 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							
% 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
% 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 one-digit matching							
% 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
% 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 (M), and the 2.5th and 97.5th percentiles, of values from 1,000 subcohorts resampled at random from the full cohort of 103,133, with replacement.

† NSAID, nonsteroidal antiinflammatory drug; EPS, exposure propensity score; DRS, disease risk score; ROC, receiver operating characteristic.

Results from subcohorts in which  $n = 10,000$  were nearly identical to those from the full cohort (with the exception of slightly wider, now empirical confidence intervals). With decreasing size of the subcohorts, however, estimates from all analytic strategies moved further away from the null and closer to the crude value, 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, use of DRSs did not translate into any gain compared with the “conventional” outcome model with respect to either point estimate or precision (results stratifying on 10 rather than five categories of the DRS were virtually identical and are therefore not presented here). Taking the absence of major differences into account, greedy matching resulted in the estimate closest to the null value (0.80), and IPTW resulted in an estimate (0.72) very close to the age- and sex-adjusted estimate (0.71) in the smallest subcohort. 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 the EPS nor the DRS was superior to “conventional” multivariable modeling in small studies with a limited number of outcomes, as was hypothesized earlier (19).

Use of the EPS (but not the 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 know much less about variable selection and model-building strategies for EPS compared with “conventional” disease models (16). Therefore, it seems desirable to use the EPS only if bias could be reduced or efficiency improved.

Cook and Goldman (4) compared the performance of tests of significance under the null hypothesis (i.e., assuming no difference between treatments) for EPS, DRS, and “conventional” multivariable outcome models by using simulations. EPSs appeared to produce nominal results in most circumstances, but not in situations with very strong treatment-confounder associations. This result was even

**TABLE 3. Association between nonsteroidal antiinflammatory drug use and 1-year mortality in a population-based cohort of hospitalized elderly according to analytic strategy and study size, New Jersey, 1995–1997**

	Full cohort ( <i>n</i> = 103,133)		Resampled subcohorts (cohort size)					
			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
“Conventional” 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
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
EPS 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 one-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 treatment 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
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 and 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 and 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

\* RR, rate ratio; CI, confidence interval; EPS, exposure propensity score; DRS, disease risk score.

† From Cox proportional hazards models (age and sex adjusted: Mantel-Haenszel estimates).

‡ Median, and the 2.5th and 97.5th percentiles, of rate ratio estimates from 1,000 cohorts resampled at random from the full cohort of 103,133, with replacement.

more pronounced for DRSs. 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 is usually considered 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 the EPS performed better in situations with fewer than eight outcomes per covariate (7). We therefore used this proportion to limit the maximum number of variables available for selection in all of our outcome models. Since precision of estimates is likely to be of minor importance for the 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 “conventional” disease models compared with the EPS or DRS.

We chose the example of NSAIDs and all-cause mortality since NSAIDs 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

and the placebo arms 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 NSAIDs were protective for some cancers, including colorectal (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) (because cohort enrollment ended in 1997, NSAID use does not include cyclooxygenase-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. (8) have argued that in an elderly population, selected drug classes, including lipid-lowering drugs, NSAIDs, or antiglaucoma drugs, are more likely to be prescribed to healthier subjects. 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 NSAIDs and mortality, our conclusions are valid for a wide range of possible effects, including a reduction of up to 15 percent in risk. More pronounced risk reductions in mortality from all causes seem biologically implausible.

The generally similar estimates resulting from applying the various analytic strategies might indicate that we were able to control adequately for observed confounding, although we do not know what the best estimate of the NSAID-mortality association would be, given the observed covariates.

Our results are limited to one specific setting with essentially the same prevalence of exposure and cumulative incidence of disease (about 20 percent each). This restriction might explain why we did not find differences between the EPS and the DRS. However, it would not explain why we did not observe any of these methods that combine multiple variables into a single score to perform better than “conventional” disease models. Generally speaking, the data structure at hand is likely to influence the choice of the preferred method. EPSs are likely to perform better than “conventional” outcome models or DRSs 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 DRSs might have an advantage over “conventional” outcome models with respect to bias and precision if the disease is rare, because 1) they allow truncation of the risk score distribution so that only the range of scores common to both exposed and unexposed subjects 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 sufficiently predict exposure or outcome. However, this limitation would invalidate our comparisons only if a small subset of variables included in all models were responsible for all of 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 NSAIDs on short-term, all-cause mortality. Incorporating additional information on factors strongly associated with the prescribing of NSAIDs and on 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 by 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 were closest to the null value in the full cohort but not why the IPTW estimate seemed furthest away from the null value in the smallest studies. The latter finding might be due to influential weights attributed to observations with “wrong” exposure status, that is, exposed observations with a very low estimated propensity for exposure, or vice versa.

We conclude that in the setting of claims data on an elderly population, various ways to apply EPSs and DRSs to

control for confounding were not generally superior to “conventional” multivariable outcome modeling. Differences in effect estimates between analytic strategies became more pronounced with smaller study size.

## REFERENCES

- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
- 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.
- Miettinen OS. Stratification by a multivariate confounder score. *Am J Epidemiol* 1976;104:609–20.
- 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;42:317–24.
- Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 1993;49:1231–6.
- Drake C, Fisher L. Prognostic models and the propensity score. *Int J Epidemiol* 1995;24:183–7.
- 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.
- Glynn RJ, Knight EL, Levin R, et al. Paradoxical relations of drug treatment with mortality in older persons. *Epidemiology* 2001;12:682–9.
- Redelmeier DA, Tan SH, Booth GL. The treatment of unrelated disorders in patients with chronic medical diseases. *N Engl J Med* 1998;338:1516–20.
- 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.
- 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.
- Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology* 1995; 6:356–65.
- 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>).
- Robins JM. Marginal structural models. In: 1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science, 1998:1–10. (<http://www.biostat.harvard.edu/%7Erobins/msm-web.pdf>).
- Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–60.
- Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol* 1993;138:923–36.
- Robins JM, Rotnitzky A. Comment on the Bickel and Kwon article, “Inference for semiparametric models: some questions and an answer.” *Statistica Sinica* 2001;11:920–36.
- 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.

19. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann Intern Med* 2002;137:693–5.
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.
21. Cepeda MS. The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiol Drug Saf* 2000;9:103–4.
22. Final report on the aspirin component of the ongoing Physicians' Health Study. Steering Committee of the Physicians' Health Study Research Group. *N Engl J Med* 1989;321:129–35.
23. Chan TA. Nonsteroidal anti-inflammatory drugs, apoptosis, and colon cancer chemoprevention. *Lancet Oncol* 2002;3:166–74.
24. Peleg II, Wilcox CM. The role of eicosanoids, cyclooxygenases, and nonsteroidal anti-inflammatory drugs in colorectal tumorigenesis and chemoprevention. *J Clin Gastroenterol* 2002;34:117–25.
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.
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.
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.
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.
29. Gurwitz JH, Avorn J, Ross-Degnan D, et al. Nonsteroidal anti-inflammatory drug-associated azotemia in the very old. *JAMA* 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.
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.
32. Gurwitz JH, Avorn J, Bohn RL, et al. Initiation of antihypertensive treatment during nonsteroidal anti-inflammatory drug therapy. *JAMA* 1994;272:781–6.
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.
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.
35. Stürmer T, Schneeweiss S, Avorn J, et al. 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, 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.
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.
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.