© *Basic & Clinical Pharmacology & Toxicology* 2006, **98**, 253–259.
Printed in Denmark . All rights reserved

Pharmacology & Toxicology

ISSN 1742-7835

MiniReview

Indications for Propensity Scores and Review of their Use in Pharmacoepidemiology

Robert J. Glynn, Sebastian Schneeweiss and Til Stürmer

Divisions of Pharmacoepidemiology and Pharmacoeconomics and of Preventive Medicine, Brigham and Women's Hospital, Harvard Medical School, and Departments of Biostatistics and Epidemiology, Harvard School of Public Health, Boston, MA, U.S.A.

(Received July 31, 2005; Accepted December 9, 2005)

Abstract: Use of propensity scores to identify and control for confounding in observational studies that relate medications to outcomes has increased substantially in recent years. However, it remains unclear whether, and if so when, use of propensity scores provides estimates of drug effects that are less biased than those obtained from conventional multivariate models. In the great majority of published studies that have used both approaches, estimated effects from propensity score and regression methods have been similar. Simulation studies further suggest comparable performance of the two approaches in many settings. We discuss five reasons that favour use of propensity scores: the value of focus on indications for drug use; optimal matching strategies from alternative designs; improved control of confounding with scarce outcomes; ability to identify interactions between propensity of treatment and drug effects on outcomes; and correction for unobserved confounders via propensity score calibration. We describe alternative approaches to estimate and implement propensity scores and the limitations of the C-statistic for evaluation. Use of propensity scores will not correct biases from unmeasured confounders, but can aid in understanding determinants of drug use and lead to improved estimates of drug effects in some settings.

Use of propensity scores in pharmacoepidemiologic studies has increased substantially over the past few years, yet evidence is lacking that this approach will systematically give better estimates of drug effects than those obtained from conventional regression approaches. If one compares the distributions of variables included in a propensity score between users of a drug and non-users matched on the propensity score, the balance of these distributions between groups will frequently be better than if drug allocation were randomized (Joffe & Rosembaum 1999). However, randomization tends to balance the unmeasured confounders, whereas matching on the propensity score often will not.

Thus, propensity score methods and conventional multivariate methods (Drake 1993) have similar inability to control unmeasured confounding. In this context, this article considers whether increased use of propensity scores is warranted. We begin with definitions, a review of the properties of the propensity score, and a description of its increasing use. We summarize available empirical comparisons and simulation studies of drug effects estimated by the propensity score versus conventional regression methods. We mention specific circumstances when use of propensity scores can improve estimates in pharmacoepidemiologic studies. We comment on alternative ways to implement propensity scores and to evaluate their performance. Finally, we summarize their limitations, point to a few areas for additional research and give recommendations for their use.

Address for correspondence: Robert J. Glynn, Division of Preventive Medicine, Brigham and Women's Hospital, 900 Commonwealth Avenue, Boston, MA, U.S.A. (fax +1 617 734 1437, e-mail rglynn@rics.bwh.harvard.ed).

Development, definitions and properties

Miettinen (1976) saw the value of summation of the evidence on confounding in a single score insofar as it allows one to display the relationship of exposure with outcome within strata of the summary score in a way that might reveal relationships that could be obscured in a multivariate analysis. He envisioned this score developed from either of two approaches: from a function that relates the potential confounders to the outcome among the unexposed, also called a disease risk score; or from a function that relates the potential confounders to exposure among the non-diseased, termed the exposure score. The disease risk score is only occasionally used to control for confounding in pharmacoepidemiologic studies (Ray *et al*. 2002), perhaps because a high correlation between the risk score and drug use can lead to overestimation of the statistical significance of the drug effect in some applications (Pike *et al*. 1979). Refinement of the exposure score into the propensity score has received widespread attention, although use of the propensity score as a continuous variable in a regression model

may have problems similar to the disease risk score if the correlation between the propensity score and actual drug use is very high.

In a series of papers, Rosenbaum & Rubin (1983, 1984 & 1985) refined the exposure score to remove its perspective on the non-diseased and demonstrated the balancing properties of what they termed the propensity score, as well as the performance of its alternative implementations. If Z is an indicator of the exposure of interest, for example Z=1 if a subject initiates use of a specific medication, and Z=0 for non-use of this drug, and X is a vector of potential determinants of drug use, possibly including both discrete and continuous variables, then the propensity score is the conditional probability of receiving treatment given the covariates; that is, PS(X)=Pr(Z=1|X). The function PS(X) is most commonly estimated by logistic regression, although other approaches are possible, including discriminant function analysis, classification and regression trees, or neural networks.

The propensity score has an important balancing property that underlies its value for observational analysis (Rosenbaum 2005). If large enough groups of exposed and unexposed subjects are found with the same value of PS(X), then these two groups will have the same distributions of all components of X. This allows direct estimation of unconfounded risk ratios and risk differences in cohort studies. In fact, stratification or matching on the propensity score can yield a better balance of measured covariates between exposed and unexposed subjects than would be observed under randomized treatment assignment (Joffe & Rosenbaum 1999). The critical limitation is that the propensity score does not share with randomization the ability to balance unmeasured confounders.

Use of propensity scores and comparisons with alternatives

Recent overviews have described the use of propensity scores in medical research and compared estimates of relationships between exposures and outcomes obtained from propensity score methods to those obtained from multivariate models (Shah *et al*. 2005; Stürmer *et al*. 2006). A systematic literature search (Stürmer *et al*. 2006) found an exponential increase in use of propensity scores over the past several years (fig. 1). From a baseline with between 6 and 9 published papers using these methods between 1998 and 2000, annual numbers of publications using propensity score methods increased to 39, 51 and 71 in 2001, 2002 and 2003, respectively. Among 177 published studies that used propensity score methods to evaluate the relationship of a dichotomous exposure with an outcome, medications were the most common treatment studied (34% of studies), followed by surgical interventions (28%), interventional catheterization (7%), and other medical procedures and lifestyle factors.

The reason for the sharp increase in use of propensity scores over the past few years is unclear. Possibly, frequently cited presentations to clinical audiences and researchers

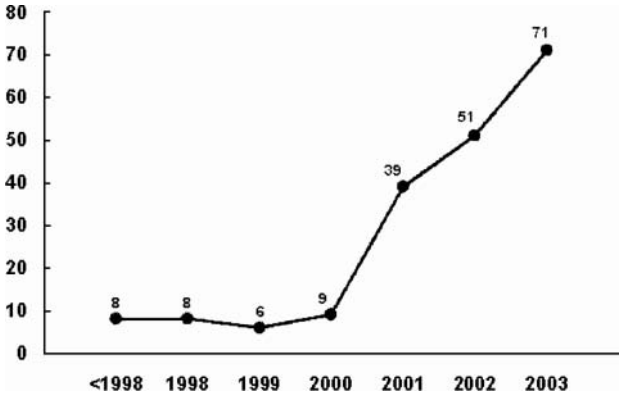

Fig. 1. Frequency of publications using propensity score methods by years.

(Rubin 1997; Joffe & Rosenbaum 1999) and tutorials (D'Agostino 1998) have influenced use.

Published studies have increasingly used both propensity score methods and regression models to evaluate the relationship between an exposure and an outcome, and reviews have compared estimates in these settings (Shah *et al*. 2005; Stürmer *et al*. 2006). A limitation of comparisons between estimates from conventional multivariate models and those based on control for the propensity score is that the approaches used to model confounding variables and the methods of construction and modeling of the propensity score vary widely across studies and are sometimes not fully described. Nonetheless, comparisons of estimated effects of drugs from multivariate models versus propensity score analysis can shed light on the performance of these approaches in real applications. Among 78 exposure-outcome associations in 43 studies evaluated both by propensity scores and regression models, statistical significance differed between the two methods in only 8 (10%) cases (Shah *et al*. 2005). The propensity score methods tended to give estimates slightly closer to the null. Another comparison of 69 studies that reported results from both propensity score and regression model approaches found only 9 (13%) to have all propensity score estimates differing by more than 20% from regression model estimates (Stürmer *et al*. 2006). Thus, there is little evidence for substantially different answers between propensity score and regression model estimates in actual usage.

Simulation studies offer the ability to compare analytic approaches in a setting where true relationships are known. Cook & Goldman (1989) compared estimates based on propensity scores, regression models and disease risk scores and found generally comparable performance of the three methods. They noted exaggerated levels of statistical significance in analyses based on propensity scores and disease risk scores in settings with a high correlation between exposure and confounders. Generally, propensity score methods displayed greater robustness to such high correlations than disease risk scores.

Cepeda *et al*. (2003) focused their simulation studies on

the setting with small numbers of events relative to the number of potential confounders. This is particularly relevant to pharmacoepidemiology where one often studies rare outcomes that occur in patients with multiple risk factors and many possible indications and contra-indications for drug use. They found that with fewer than eight events per confounder, analysis based on propensity scores yielded estimates that were less biased, more robust, and more precise than a regression approach based on logistic regression. By contrast, propensity score methods had poorer coverage than regression methods with larger numbers of events per confounder. These results are entirely consistent with the known poor performance of regression models with small numbers of events per variable (Peduzzi *et al*. 1996), and indicate an important situation where propensity score methods are clearly preferred (Braitman *et al*. 2002).

Another important topic evaluated in simulation studies is the impact of omitted covariates on the performance of estimates based on the propensity score. Often, available databases with detailed information on drug use either lack information on an important covariate or can only measure it crudely. Drake (1993) showed that omitted covariates yield comparable bias in estimates based on propensity scores relative to those based on regression models. She further demonstrated that failure to specify the response model correctly induces greater bias than incorrect specification of the propensity score and that the propensity score does not yield balance in the distributions of omitted covariates between treated and untreated subjects.

Five reasons to use propensity scores in pharmacoepidemiology

Theoretical advantages.

While analyses based on propensity scores often give similar estimates to those from regression models, and the balance in observed covariates can give the false sense that unobserved covariates are also balanced, propensity scores offer important theoretical advantages in pharmacoepidemiology. Confounding by indication is often the main challenge to validity in pharmacoepidemiology and the propensity score focuses directly on the indications for use and non-use of the drug under study. Patients with contraindications to use of a drug (or those with absolute indications) may have no comparable exposed subjects (or unexposed subjects) for valid estimation of relative or absolute differences in outcomes. These subjects are not usually recognized with conventional response modeling and might be influential due to effect measure modification or model misspecification. Graphical comparison of propensity scores in exposed versus unexposed subjects can identify these areas of non-overlap that are otherwise difficult to describe in a multivariate setting with many factors influencing treatment decisions (fig. 2 for an illustration).

The propensity score has direct scientific interest in studies that focus on determinants of drug initiation or persistence with therapy. Consideration of the propensity score can broaden one's perspective to include barriers to treatment. For example, frailty and comorbidity that are difficult to measure in large databases can lead to decreased use of preventive drug therapies. Shown in table 1 are several markers of frailty and comorbidity that are related to decreased propensity to use lipid-lowering drugs among older residents of New Jersey. Recognition of the importance of such factors and their inclusion in propensity scores can lead to improved control for confounding, relative to an analysis that does not control for these factors (Glynn *et al*. 2006). Further, understanding of the role such factors can play in drug use is of fundamental interest in pharmacoepidemiology and the propensity score naturally focuses on this issue.

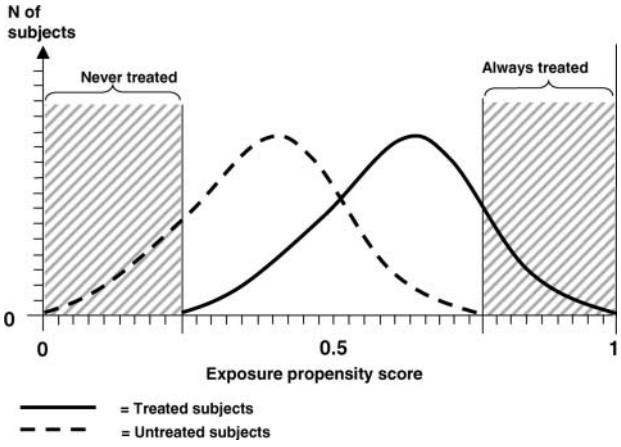

Fig. 2. The non-overlap of the exposure propensity score distribution among treated and untreated study subjects. In this example subjects with very low propensity score are never treated while subjects with very high propensity score are all treated.

Value of propensity scores for matching or trimming the population.

Matching or stratification on the propensity score offers several advantages relative to inclusion of an estimated linear propensity score in a conventional multivariate model. First, a matched analysis will eliminate those exposed subjects (e.g. those with absolute indications for therapy) with no comparable controls as well as those unexposed subjects with measurable contra-indications. Second, matched or stratified analyses do not make strong assumptions of linearity in the relationship of propensity with the outcome. Third, and perhaps most importantly, a matched data set allows for a simple, transparent analysis.

The balancing property of the propensity score has implications for optimal matching strategies in both cohort and cross-sectional studies in pharmacoepidemiology. Matching on the propensity score will outperform other matching strategies with many covariates in the sense that optimal balance of covariates will be achieved between exposed and unexposed groups (Gu & Rosenbaum 1993). The balance achieved in prospective studies will mimic that of randomization but, of course, will hold only for variables that are measured and included in the propensity score.

Table 1.

Correlates of lower propensity to use lipid-lowering drugs. Data from enrollees in New Jersey drug benefit programs age 65 years or older.

	Propensity quintile				
	1	2	3	4	5
N	22,492	22,493	22,493	22,493	22,492
Nursing home resident, %	82	46	24	12	5
Cardiac arrhythmia, %	27	22	19	16	14
Other neurologic disorders, %	18	10	7	5	3
Fluid, electrolyte disorder, %	32	16	11	9	8
Congestive heart failure, %	38	29	24	22	21
Dementia, %	31	10	3	1	0.5
COPD*, %	26	23	21	18	14

*Chronic obstructive pulmonary disease.

A limitation of matching is that many unexposed subjects not matched to exposed subjects, and possibly some unmatched exposed subjects, are excluded from analysis. This can lead to a loss of information and a decrease in the precision of the estimated association between the drug and the outcome. As an alternative, one can trim the population for analysis through exclusion of those subjects in the two tails of the propensity score distribution where overlap between those who use and do not use the drug of interest may be limited. This can be viewed as a principled approach to eliminate extreme observations that may be unduly influential and problematic in a multivariate analysis because of minimal covariate overlap between exposed and unexposed subjects. The reduction of the population for analysis is appropriate if the excluded subjects are those who are not candidates for drug therapy, or possibly if the other tail of the distribution consists entirely of people with an absolute need for the drug. Although trimming has these theoretical advantages, optimal trimming strategies (e.g. exclusion of the extreme 1% or 2% of the propensity score distributions) are unknown.

Improved estimation with few outcomes.

As previously noted, one common setting in pharmacoepidemiology where use of the propensity score can provide clearly improved estimates of drug effects occurs when one has relatively few outcomes compared with the number of potentially important covariates. In this setting, reliable estimation of many parameters in multivariate models is not possible because maximum likelihood estimation requires many outcomes per included parameter in a model (Harrell *et al.* 1996). Use of the propensity score provides an effective way to reduce the dimensionality of the covariates before modeling. The rule of eight proposed by Cepeda *et al.* (2003) (fewer than eight outcomes per included covariate) gives a helpful guideline on when use of the propensity score should effectively improve estimation.

Propensity score by treatment interactions.

Consideration of the propensity score focuses on the real possibility that the effectiveness of a drug may vary according to the strength of the indication for its use. Among patients with weak indication for use, or among those with contraindications for use, a drug may provide no benefit or even be harmful, while in patients with clear indications for use, the drug may provide substantial benefits. These clinically relevant concerns are frequently overlooked in analyses of pharmacoepidemiologic studies, but the propensity score provides a natural perspective to elucidate them.

The example of Kurth *et al.* (2006) illustrates the relevance of this perspective for pharmacoepidemiology. They studied the effect of treatment with tissue plasminogen activator (t-PA) on in-hospital mortality among 6,269 ischemic stroke patients in Westphalia. Their population included some treated patients with low propensity to receive treatment and small numbers of untreated patients with a high propensity to receive t-PA (fig. 3). Stratified analysis by levels of the propensity score revealed heterogeneity in efficacy perhaps due to side effects of treatment. Treated patients with low propensity to receive t-PA had substantially elevated death rates relative to untreated patients. However, among those with propensity to receive t-PA above 5%, the relative odds of death in treated versus untreated patients was approximately 1. It is unclear how this anticipated interaction would be identified outside the framework of the propensity score, if it arises from a combination of factors.

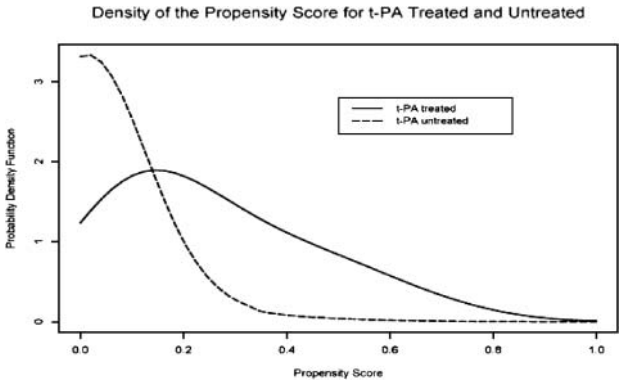

Fig. 3. Distribution of the propensity score.

Propensity score calibration to correct for measurement error.

In almost all pharmacoepidemiologic studies, some covariates are either not measured or measured with error. Neither standard applications of propensity scores nor use of regression models may adequately adjust for such unmeasured or mis-measured covariates. However, it may be possible to obtain a more reliable estimate of the propensity score in a sub-study with more detailed covariate information and then use this gold-standard propensity score to correct the main-study effect of the drug on the outcome (Stürmer *et al*. 2005a). One can view this approach as an application of regression calibration to correct for the measurement error in the main study propensity score that is available for all study subjects (Carroll *et al*. 1995). Use of propensity score calibration allows one to account for multiple unobserved confounders that may have available information only for a subgroup of study subjects.

To illustrate the method, consider a study of the relationship of use of non-steroidal anti-inflammatory drugs (NSAIDs) with 1-year mortality in a large cohort of older people (Stürmer *et al*. 2005a). The main study follows 103,133 residents of New Jersey age 65 or older for 1 year. As is common in data base studies, one has information on drug use, mortality and many determinants that allow for estimation of the propensity to NSAID use. However, other potentially important determinants of NSAID use, including cigarette smoking, non-prescription aspirin use, body mass index and education, may be available in a smaller, separate study such as the Medicare Current Beneficiary Survey (MCBS). The available data elements from these sources are illustrated in fig. 4. One can estimate both the error prone and the gold standard propensity score in the validation sample that also contains information on NSAID use but is too small for reliable evaluation or lacks information on the outcome of 1-year mortality.

Analyses based only on the main study data found a significant 20% reduction in mortality among NSAID users in a multivariate regression model (relative risk (RR) 0.80; 95% confidence interval (CI): 0.77–0.83) that was virtually unchanged upon control for the error-prone propensity score available in the main study (RR 0.81; 95% CI: 0.78–0.84). A similar protective effect of NSAIDs on mortality was observed in a prior observational study and could not be explained by available measures of confounding variables (Glynn *et al*. 2001). Application of propensity score calibration, based on the relationship of the gold-standard propensity score with the error prone propensity score and actual NSAID use in the validation study, resulted in a more plausible RR of 1.06 (95% CI: 1.00–1.12).

Propensity score calibration illustrates the magnitude of the bias that can arise from uncontrolled confounding. Propensity score calibration relies on the often unverifiable assumption inherent in corrections based on regression calibration that the error-prone propensity score is independent of the outcome given the gold-standard propensity score. If this assumption does not hold, propensity score calibration can increase bias in some scenarios (Stürmer *et al*. 2005a). The approach may perform better with internal validation studies where detailed information on confounders is available for a sample of the subjects included in the main study.

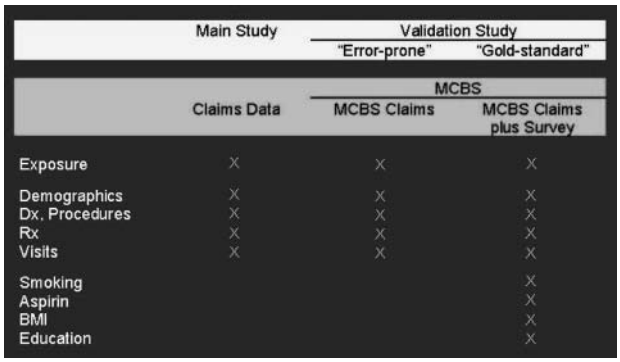

Fig. 4. Propensity score calibration.

Practical considerations for estimation and evaluation of the propensity score

The great majority of applications of propensity scores have used logistic regression to estimate the score. Other approaches such as classification and regression trees have been used (Cook & Goldman 1988), and neural networks can also be considered. However, logistic regression may be more accessible to readers, and it is not clear that alternative approaches will yield scores that give better adjustment for confounding.

Construction of the propensity score should consider barriers as well as indications for treatment. In building the propensity score, use of non-parsimonious models with consideration of interaction terms is recommended (D'Agostino 1998). Rubin (1997) recommended inclusion of variables that are strongly related to outcome, regardless of their apparent effect on the exposure. In simulations of small to moderate sized studies, Brookhart *et al*. (2006) found that inclusion of such variables increases the precision of the estimated exposure effect. However, these authors also found that inclusion of variables strongly related to exposure but unrelated or only weakly related to the outcome can substantially increase the mean squared error of the estimated treatment effect. Thus, maximal prediction of treatment status may not be optimal in developing a propensity score.

These results have implications for the common practice of reporting the area under the ROC curve (or C-statistic) as a measure of the adequacy of a propensity score. A very high C-statistic can indicate non-overlap in the distribution of propensity scores between treated and untreated subjects, and suggests an inability to make comparisons between treated and untreated subjects. This is related to the documented poor performance of analyses that use a linear propensity score in the presence of very good discrimination

of treated and untreated subjects by covariates (Cook & Goldman 1989), and the limited ability to obtain reliable estimates of highly correlated variables in regression models (if the propensity score and actual treatment have very high correlation). With a high C-statistic it is particularly important to consider analytic approaches such as matching or stratification to reduce the influence of subjects with extreme propensity score values. Additionally, a high C-statistic cannot be taken as evidence that the propensity score included every important confounder (Weitzen *et al*. 2005).

A relevant evaluation of the usefulness of a propensity score in a specific setting should compare the balance of covariates between exposed and unexposed subjects within strata of the score (Rubin 2004). Lack of balance can indicate the need to add higher order or non-linear terms to the propensity score. Alternatively, imbalances can identify subjects in the tails of the distribution of the score with contraindications or absolute indications for treatment. These subjects can then be excluded from analysis.

Alternative implementations of propensity score analysis

Once a propensity score is constructed, several alternative analytic strategies are available for its implementation. Common implementations include control for the propensity score in a regression model, matched or stratified analysis, inverse probability weighting and combinations of these approaches. A matched analysis based on a well-formulated propensity score has the advantage of deleting from analysis those subjects with contra-indications (or absolute indications) for treatment who have no available treated (or untreated) comparison subject. An analysis that uses inverse propensity score weights has population-based interpretations (Robins *et al*. 2000), but can be very sensitive to the estimated weights (Stürmer *et al*. 2005b). If the propensity score is included in a multivariate model together with actual treatment, options include use of the continuous linear propensity score, indicators of quintiles of the score, or allowance for non-linearity through use of splines. Use of a continuous, linear score makes a strong assumption about the relationship between propensity and disease risk, and estimated treatment effects can be biased if this assumption does not hold (Rosenbaum & Rubin 1983). As a possible mixed strategy, one can include the propensity score together with all potential confounding variables and treatment status in a common multivariate model with the hope of improving confounder control if either the relationship of the propensity score or the confounders with the outcome is correctly specified. However, evidence of improved control for confounders and less biased estimates of treatment effects through this approach is unavailable.

Stürmer *et al*. (2005b) used resampling strategies to compare performance of alternative implementations on estimated treatment effects in studies of varying sizes. Effect estimates based on inverse-probability weighting performed well in larger samples. However, in small samples this approach was sensitive to patients with extreme propensity scores. This was also noted in the example of Kurth *et al*. (2006) where patients treated with t-PA with low propensity scores had a large impact on inverse probability weighted estimates. More work is needed on optimal weights to discount influential outliers in application of inverse-probability weighting to estimation with propensity scores.

Within the range of circumstances considered by Stürmer *et al*. (2005b), use of alternative propensity score approaches demonstrated no superiority in terms of reduced bias or increased precision relative to conventional multivariate models. Further, the hybrid strategy with both propensity scores as well as available confounders in the same model did not give clearly better estimates than multivariate models without the propensity score. Overall, the alternative implementations of propensity score methods gave estimates similar to each other and to conventional multivariate models in this setting.

Conclusions and future directions

The propensity score has the important balancing property that treated and untreated subjects with the same propensity score will typically have comparable distributions of measured covariates that will often be more similar than the distributions of these covariates between groups of persons with randomly assigned treatment. Unlike the setting of randomized treatments, one cannot expect the balance in distributions of covariates included in the propensity score to extend to other covariates not included in the propensity score. Thus, use of a propensity score does not resolve the traditional concern in pharmacoepidemiology that patients who receive a drug differ in disease severity or have other prognostic differences with untreated patients. Further, for many of the study sizes and designs common to pharmacoepidemiology, there is no evidence that an analysis utilizing propensity scores will substantially decrease bias from confounding, relative to conventional estimation in a multivariate model.

Much of the work on propensity scores assumes a dichotomous treatment evaluated at a single point in time. Often more than one treatment option is available and while modeling of multi-category choices, for example through polytomous logistic regression, is straightforward, experience in this area is limited. More challenging are conceptualizations of time-varying propensity as patients make decisions to initiate, continue or terminate treatments over time. Variables related to the initiation of therapy may differ from those associated with persistence. Analysis of these processes will need to account for intermediate variables that may be influenced by prior treatments and the prior disease course that can also influence disease outcomes.

Although use of propensity scores is not guaranteed to reduce bias due to confounding, its use in pharmacoepidemiology can still be recommended for several reasons. Most fundamentally, the propensity score focuses on the multifaceted determinants of drug use, and understanding these determinants has intrinsic interest in pharmacoepidemiolo-

gy. Comparison of the distributions of the propensity score between exposed and unexposed subjects can identify those with absolute indications or contra-indications to therapy for whom no comparison may be available. Stratification on the propensity score may be important if the effect of the therapy may reasonably vary according to the strength of the indication for its use. In some common settings such as studies with many covariates and few outcomes the propensity score offers a straightforward approach to reduce the dimensionality of the array of confounders and improve their control. Finally, if covariates either unavailable or mismeasured in the main study are measured with greater validity in a substudy, propensity score calibration offers one potentially useful approach to adjust for the potential bias in estimates based solely on the main study.

Acknowledgements

Supported by grants AG18833 and AG23178 from the National Institute on Aging.

References

Braitman, L. E. & P. R. Rosenbaum: Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann. Intern. Med.* 2002, **137**, 693–695.

Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn & T. Stürmer: Variable selection in propensity score models. *Amer. J. Epidemiol*. 2006, in press.

Carroll, R. J., D. Ruppert & L. A. Stefanski: *Measurement error in nonlinear models*. Chapman and Hall, London 1995.

Cepeda, M. S., R. Boston, J. T. Farrar & B. L. Strom: Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Amer. J. Epidemiol.* 2003, **158**, 280–287.

Cook, E. F. & L. Goldman: 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.

Cook, E. F. & L. Goldman: Asymmetric stratification. An outline for an efficient method for controlling confounding in cohort studies. *Amer. J. Epidemiol.* 1988, **127**, 626–639.

D'Agostino, R. B. Jr.: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* 1998, **17**, 2265–2281.

Drake, C.: Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 1993, **49**, 1231–1236.

Glynn, R. J., E. L. Knight, R. Levin & J. Avorn: Paradoxical relations of drug treatment with mortality in older persons. *Epidemiology* 2001, **12**, 682–689.

Glynn, R. J., S. Schneeweiss, P. S. Wang, R. Levin & J. Avorn: Selective prescribing led to over-estimation of the benefits of lipid-lowering drugs. *J. Clin. Epidemiol.* 2006, in press.

Gu, X. S. & P. R. Rosenbaum: Comparison of multivariate matching methods: structures, distances and algorithms. *J. Computat. Graph. Stat.* 1993, **2,** 405–420.

Harrell, F. E. Jr., K. L. Lee & D. B. Mark: Multivariable prognostic models. Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med*. 1996, **15**, 361–387.

Joffe, M. M. & P. R. Rosenbaum: Invited commentary: propensity scores. *Amer. J. Epidemiol*. 1999, **150**, 327–333.

Kurth, T., A. M. Walker, R. J. Glynn, K. A. Chan, J. M. Gaziano, J. M. Robins & K. Berger: Results of multivariable logistic regression, propensity matching, propensity adjustment and propensity-based weighting under conditions of non-uniform effect. *Amer. J. Epidemiol.* 2006, in press.

Miettinen, O. S.: Stratification by a multivariate confounder score. *Amer. J. Epidemiol*. 1976, **104**, 609–620.

Peduzzi, P., J. Concato, E. Kemper, T. R. Holford & A. R. Feinstein: A simulation study of the number of events per variab le in logistic regression. *J. Clin. Epidemiol*. 1996, **49**, 1373–1379.

Pike, M. C., J. Anderson & N. Day: Some insights into Miettinen's multivariate confounder score approach to case-control study analysis. *Epidemiol. Comm. Health* 1979, **33**, 104–106.

Ray, W. A., C. M. Stein, K. Hall, J. R. Daugherty & M. R. Griffin: Non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease: an observational cohort study. *Lancet* 2002, **359**, 118–123.

Robins, J. M., M. A. Hernan & B. Brumback: Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000, **11**, 550–560.

Rosenbaum, P. R. & D. B. Rubin: The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983, **70**, 41–55.

Rosenbaum, P. R. & D. B. Rubin: Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* 1984, **79**, 516–524.

Rosenbaum, P. R. & D. B. Rubin: Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* 1985, **39**, 33–38.

Rosenbaum, P. R.: Propensity score. In: *Encyclopedia of biostatistics*, Second edition. Eds.: P. Armitage & T. Colton. John Wiley and Sons, Chichester, 2005, pp. 4267–4272.

Rubin, D. B.: Estimating causal effects from large data sets using the propensity score. *Ann. Intern. Med.* 1997, **127**, 757–763.

Rubin, D. B.: On principles for modeling propensity scores in medical research. *Pharmacoepidemiol. Drug Safety* 2004, **13**, 855–857.

Shah, B. R., A. Laupacis, J. E. Hux & P. C. Austin: Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J. Clin. Epidemiol.* 2005, **58**, 550–559.

Stürmer, T., M. Joshi, R. J. Glynn, J. Avorn, K. Rothman & S. Schneeweiss: A review of the application of propensity score methods yielded increased use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J. Clin. Epidemiol.* 2006, in press.

Stürmer, T., S. Schneeweiss, J. Avorn & R. J. Glynn: Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Amer. J. Epidemiol.* 2005a, **162**, 279–289.

Stürmer, T., S. Schneeweiss, M. A. Brookhart, K. J. Rothman, J. Avorn & R. J. Glynn: Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal anti-inflammatory drugs and short-term mortality in the elderly. *Amer. J. Epidemiol.* 2005b, **161**, 891–898.

Weitzen, S., K. L. Lapane, A. Y. Toledano, A. L. Hume & V. Mor: Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of omitted confounders. *Pharmacoepidemiol. Drug Safety* 2005, **14**, 227–238.