Propensity Score Methods for Confounding Control in Nonexperimental Research

M. Alan Brookhart, PhD; Richard Wyss, MS; J. Bradley Layton, PhD; Til Stürmer, MD

Background

Nonexperimental studies are increasingly used to investigate the safety and effectiveness of medical products as they are used in routine care. One of the primary challenges of such studies is confounding, systematic differences in prognosis between patients exposed to an intervention of interest and the selected comparator group. In the presence of uncontrolled confounding, any observed difference in outcome risk between the groups cannot be attributed solely to a causal effect of the exposure on the outcome.

Confounding in studies of medical products can arise from a variety of different sociomedical processes.[1] The most common form of confounding arises from good medical practice, physicians prescribing medications and performing procedures on patients who are most likely to benefit from them. This leads to a bias known as confounding by indication, which can cause medical interventions to appear to cause events that they prevent.[2,3] Conversely, patients who are perceived by a physician to be near the end of life may be less likely to receive preventive medications, leading to confounding by frailty or comorbidity.[4–6] Additional sources of confounding bias can result from patients' health-related behaviors. For example, patients who initiate a preventive medication may be more likely than other patients to engage in other healthy, prevention-oriented behaviors leading to bias known as the healthy user/adherer effect.[7–9]

Many statistical approaches can be used to remove the confounding effects of such factors if they are captured in the data. The most common statistical approaches for confounding control are based on multivariable regression models of the outcome. To yield unbiased estimates of treatment effects, these approaches require that the researcher correctly models the effect of the treatment and covariates on the outcome. However, correct specification of an outcome model can be challenging, particularly in studies involving many confounders, rare outcomes, or strong treatment effect heterogeneity that must be correctly modeled.[10] Propensity score (PS) models are an approach for estimating treatment effects that do not rely on modeling the outcome.[11] Rather, PS methods rely on a model of the treatment given the confounders. In many cases, these models may be easier to specify.

In this article, we outline the steps involved in the implementation of a PS analysis, including a description of the various ways that a PS can be used. We focus on the practical application of these methods in the area of nonexperimental studies of medical interventions. Using a large insurance claims database, we illustrate the discussed concepts using a substantive example involving the comparison of angiotensin-converting enzyme inhibitors (ACEis) and angiotensin receptor blockers (ARBs) on the risk of angioedema, a well-established adverse effect of ACEi initiation.[12–15]

The Propensity Score

A PS is defined as the conditional probability of treatment or exposure given all confounders.[11] Rosenbaum and Rubin[11] formalized PS methods and showed that all confounding can be controlled through the use of the PS. They demonstrated that among patients with the same PS, treatment is unrelated to confounders. Therefore, the treated and untreated tend to have the same distribution of measured confounders, something that we would also achieve using randomization. Robins et al[16] extended the application of PSs through the development of inverse probability of treatment weighted (IPTW) estimation, and other weighting approaches have been proposed.[17]

Given a PS, treatment effects are usually estimated by matching, weighting, stratification, or adjustment for the PS in a multivariable regression model. In the presence of treatment effect heterogeneity, these different approaches may result in estimates of different treatment effects (contrasts).[18] In PS matching, each treated subject is matched to ≥1 control subjects depending on the matching algorithm used. The effect estimate obtained from PS matching is generalizable only to populations similar to the matched patients. In many applications, one is matching a small group of treated patients to a larger population of untreated patients. When all treated patients can be matched, this results in an estimate of the average treatment effect in the treated (ATT). In some cases, we cannot find untreated matches for each treated patient, and therefore the estimate is not completely generalizable to the entire treated population. PS methods based on matching can also result in a substantial reduction in sample size.

From the Department of Epidemiology, Gillings School of Global Public Health (M.A.B., R.W., J.B.L., T.S.) and Cecil G. Sheps Center for Health Services Research (M.A.B., J.B.L.), University of North Carolina at Chapel Hill.

The online-only Data Supplement is available at http://circoutcomes.ahajournals.org/lookup/suppl/doi:10.1161/CIRCOUTCOMES.113.000359/-/DC1.

Correspondence to M. Alan Brookhart, PhD, Department of Epidemiology, UNC Gillings School of Global Public Health, UNC-Chapel Hill, McGavran-Greenberg, CB #7435, Chapel Hill, NC 27599-7435. E-mail abrookhart@unc.edu

(*Circ Cardiovasc Qual Outcomes*. 2013;6:604-611.)

***Circ Cardiovasc Qual Outcomes* is available at http://circoutcomes.ahajournals.org** **DOI: 10.1161/CIRCOUTCOMES.113.000359**

See Guo and Fraser[19] for an overview of different approaches to matching.

PSs can also be used to generate weights that can be used to control confounding. This approach, unlike matching, does not result in a reduction of the original sample size. The purpose of PS weighting is to reweight the individuals within the original treated and control samples to create a so-called pseudopopulation in which there is no longer an association between the confounders and treatment. Two types of weighting are commonly used: IPTW and standardized mortality ratio weighting.[17]

IPTW is defined as the inverse of the estimated PS for treated patients and the inverse of one minus the estimated PS for control patients. Patients who receive an unexpected treatment are weighted up to account for the many patients like them who did receive treatment. Patients who receive a typical treatment are weighted down because they are essentially over-represented in the data. These weights create a pseudopopulation where the weighted treatment and control groups are representative of the patient characteristics in the overall population. Therefore, IPTW results in estimates that are generalizable to the entire population from which the observed sample was taken. The treatment effect obtained after applying IPTW is referred to as the population average treatment effect (ATE).[18,20]

Precision of estimated effects from an IPTW analysis can be improved by stabilizing the weights. This is done by multiplying the previously defined weights by the marginal probability of receiving treatment for those treated and the marginal probability of not receiving treatment for those not treated.[21] Stabilized weights do not increase or decrease bias but can increase precision in the estimated treatment effects by reducing the variance of the weights.[16,21,22]

For standardized mortality ratio weighted (SMRW) estimation, treated patients are given a weight of 1 whereas weights for control patients are defined as the ratio of the estimated PS to 1 minus the estimated PS.[18,20] SMRW reweights the control patients to be representative of the treated population. SMRW results in an estimate of ATT.[20] Unlike PS matching, which also often estimates the ATT, no treated individuals are excluded from this analysis. Both IPTW and SMRW require the use of a robust variance estimator, similar to the variance estimator used in the generalized estimating equation methodology. This approach results in confidence intervals (CIs) that are conservative, that is, have a slightly greater than nominal coverage.

Researchers also commonly estimate treatment effects by conducting an analysis stratified across the PS. Often the strata are taken to be quintiles or deciles of the PS. Treatment effects are estimated within these strata, and a summary effect is generated by taking the weighted average of the stratum-specific estimates using weights that are proportional to the number of outcomes in each stratum, optimizing statistical efficiency of estimation. Assuming uniform treatment effects, this approach results in an estimate of the ATE in the population. In the presence of heterogeneous treatment effects, however, this approach may no longer result in an estimate of the ATE.

Finally, it is possible to combine PS and regression methods in various ways. For example, researchers often include the PS in a regression model along with other covariates or use regression adjustment in cohorts that have already been matched on the PS. These approaches are not unreasonable and may help to remove some residual confounding. However, they may not result in estimates of a parameter of interest, such as ATE or ATT (although these could be obtained from the fitted model using marginalization). One principled approach to combining regression and PS weighting is through the use of the augmented IPTW estimator.[23,24] This estimator depends on both a PS and multivariable outcome model and results in an estimate of the ATE. The estimator is often more efficient than an IPTW estimator. The augmented IPTW estimator also possesses an appealing doubly robust property, meaning that to be consistent, it needs only 1 of the 2 models to be correctly specified.[25]

Estimating the PS

In practice, the true probability of treatment is unknown and therefore must be estimated from the available data. The ability of estimated PSs to control for measured confounding is contingent on both correctly selecting variables for the PS model and specifying the functional form of the relation between selected covariates and treatment.[11,26,27] Logistic regression is the most widely used method to estimate PSs.[28] Models that are more flexible than logistic regression are increasingly used for PS estimation and have been found to perform well in specific settings.[29–31]

Once a model has been chosen, the analyst must select which variables need to be included in the model. Ideally, this process is guided by subject-matter knowledge. In practice, however, treatment assignment is usually determined by a complex interaction of patient, physician, and healthcare system factors that are often incompletely understood.

Various model building and variable selection strategies have been proposed to help researchers select variables for inclusion in a PS. Variable selection strategies range from simply selecting covariates a priori based on expert and substantive knowledge to approaches that are more empirical or data-driven where large PS models are constructed that control for large numbers of covariates.[32–34]

The choice of variables that one includes in a PS model can influence both the validity and efficiency of effect estimates. Simulation studies have demonstrated that the best predictive models of exposure do not necessarily result in optimal PS models.[26] For example, the inclusion of variables in a PS model that affect the outcome but not treatment are beneficial because they decrease the variance of the estimated treatment effect.[26] Conversely, including variables that affect only treatment can be harmful.[26] These variables can increase the variability of effect estimates and, in the presence of unmeasured confounding, increase bias.[1,35,36] Studies of variable selection for PS suggest that optimal PS models, in terms of bias and precision, include all variables that affect the outcome of interest regardless of whether they are important determinants of treatment.[26] Ideally, this should be determined from subject-matter knowledge, for example, as coded in causal graphs.[37]

In the common setting of rare outcomes but common treatments, it is possible to build much larger models of treatment

than of the outcome. This allows one to control for many more covariates in a PS analysis. This has led researchers to use algorithms that result in large PS models, including the so-called high-dimensional propensity score algorithm.[33,34] These methods have performed well in several empirical examples; however, theoretically, in some situations they could result in a more biased estimate than more parsimonious PS models.

Regardless of the particular approach adopted, 1 helpful feature of PS methods is the ability to explicitly evaluate the performance of an estimated PS model by assessing the balance of covariates after matching or weighting on the estimated PSs. Residual imbalances in the covariates indicate a possible problem with PS model specification. This process allows the researcher to evaluate and modify the PS model before attempting to estimate the treatment effect. Various strategies for PS variable and model selection based on the evaluation of covariate balance have been proposed.[29]

Once the PS model is fit and the estimated PSs are generated, it is common to plot the frequency distribution (or estimated density function) of the PS within each treatment group. These plots allow the researcher to identify regions of the PS distributions with little or no overlap where treatment effects cannot be reliably estimated. It is reasonable to consider removing patients from these regions from the analysis, an approach referred to as PS trimming. Matching will automatically remove most of the patients from this region because they cannot be matched, and SMRW methods will reduce their influence by giving them small weights.

IPTW methods can be particularly sensitive to the influence of patients who receive unexpected treatments. Because IPTW estimates the ATE in the population, it must upweight individuals in the population who are given an apparently unusual treatment.[38] If treatment effects differ in these patients or an unmeasured reason for the unexpected treatment affects the risk for the outcome,[39] IPTW estimates can be unlike estimates from other approaches. This has been observed in several empirical examples.[20,40]

In situations where a small number of patients influence the results of the analysis because of their large weights, one should carefully investigate potential causes of this issue and rule out problems such as data errors. It may be reasonable to consider removal of these patients from the analysis through PS trimming. However, this changes the target population for inference, and the benefits of estimating causal effects on a well-defined population are lost. If the cause of the large weights is unmeasured confounding, however, trimming may decrease bias.[39] In this case, the disadvantage of losing the causal interpretation would be moot because generalizability is no longer relevant in the setting of a biased estimator.

It is worth noting that both PS methods and multivariable outcome models can identify treatment effects in patients in the nonoverlapping regions of the PS by model extrapolation, that is, assuming that the effect of treatment in the patients who are always treated or never treated is similar to the treatment effect in other patients. This assumption is often unknowingly made and can lead to misleading results. Therefore, an advantage of PS methods is the ability to identify patients whose treatment effects cannot be reliably estimated.

Exploring Effect Modification by the PS

Differences between the various approaches to using the PS occur when there is substantial treatment effect heterogeneity.[18] If the treatment effect varies with the PS, different PS methods will give different results.[22,40] It can be informative to report estimated effects by strata of the PS distribution. For example, in a study of thrombolytic therapy for ischemic stroke, the IPTW approach suggested that thrombolytic therapy was associated with a large risk of in-hospital mortality.[20] However, based on PS matching, the authors found little evidence of a substantially increased mortality attributable to the treatment.

By examining treatment effects across PS strata, the authors discovered that treatment in patients with a low probability of receiving treatment was associated with a greatly increased risk of mortality. This finding suggested the possibility that these patients may have possessed an unmeasured contraindication for treatment. If so, the treatment effect estimate generated by IPTW generalized to many patients who should not have received thrombolytic therapy and therefore may not be of great clinical interest. Note that the ATT estimate based on PS matching or SMRW reduces the potential for estimating the treatment effect in patients with contraindications. Examining treatment effects across strata of the PS is an effective way of identifying treatment effect heterogeneity. However, further analysis would be required to identify the true source of the heterogeneity.

PS Methods for Multicategorical Treatments

In the setting of a treatment that has multiple levels, the PS becomes a vector, that is, the predicted probability of each treatment category. These can be estimated using a model for a categorical outcome, such as multinomial logistic regression. IPTW methods can be used directly in this setting. As in the case of a dichotomous treatment, each patient receives a weight equal to the inverse of the probability that they would receive their actual treatment. Stratification and matching on a multivariate PS are possible but not preferred in this setting because the PS is no longer single dimensional.

Example: Angioedema Risk Among New Users of ACEis Versus ARBs

For our illustrative example, we identified a cohort of new users of ACEis or ARBs in a large, US employer–based insurance claims database—the MarketScan Commercial Claims and Encounters and Medicare Supplementary and Coordination of Benefit (Truven Healthcare, Inc.). The database contains patient billing information for in- and outpatient procedures and diagnoses, pharmacy medication dispensing, and enrollment information for enrolled employees, spouses, dependents, and retirees.

New ACEi or ARB use was defined as a pharmacy dispensing of an ACEi or ARB to individuals who had been free of antihypertensive use (β-blockers, calcium channel blockers, α-blockers, thiazide diuretics, ACEi or ARB medications) for 6 months. To restrict to ACEi or ARB monotherapy, we also excluded individuals who initiated another antihypertensive within 1 day of the index ACEi or ARB prescription. Patients

were followed up for 1 year after initiation. The outcome was an occurrence of angioedema, defined as an *International Classification of Diseases, Ninth Revision, Clinical Modification* code of 995.1 associated with an in- or outpatient encounter. Patients with angioedema occurring during the washout period were excluded. This setting may be particularly well suited to PS methods because angioedema is a relatively rare outcome, and the database provides a large collection of candidate covariates that one may wish to include in a model as potential confounders.[10] Furthermore, the risk of angioedema is known to vary across race and may be heterogeneous across other subgroups.[41] The PS approach allows us to estimate a population treatment effect without the need to explicitly specify these interactions.

Estimating the Probability of ACEi Versus ARB Use

For the example considered in this article, we identified covariates a priori based on the literature, substantive knowledge, and the availability of covariates within the data. Covariates were defined from diagnoses and procedures that were performed during the 6-month baseline period. Considered covariates included markers of cardiovascular risk and cardiovascular disease management, recent acute events, other cardiovascular medication use and coadministration (diuretics, statins and other anticholesterol drugs, and anticoagulants), and patient characteristics. A list of these covariates along with a description of their distributions stratified by exposure status is shown in Table 1.

Using logistic regression, we estimated the PSs by modeling the main effects of the covariates listed in Table 1. To assess the comparability of the covariate distributions between the ACEi and ARB groups, we plotted the distributions of the estimated PSs stratified by exposure status (Figure 1). Treated (ACEi initiators) and control (ARB initiators) patients with similar PS values will, on average, have similar covariate distributions. Therefore, the overlapping region of the PS distributions identifies the subset of the observed population where the patient populations are comparable.

PS Implementation and Causal Effect Estimation

In each of the example analyses in this article, we compared rates of angioedema in ACEi versus ARB users with Cox proportional hazard models in which the outcome was censored by plan disenrollment or administratively by 1 year after the index date. We present hazard ratios (HRs) and 95% CIs.

We matched ACEi initiators to ARB initiators 1-to-1, without replacement, using a varying width caliper matching algorithm (5-to-1 digit matching). Because there were many more ACEi users and substantial overlap in the PS distributions between ACEi and ARB users, almost all of the ARB users could be matched, but many ACEi users were discarded. From the matched set of observations, the treatment effect was then estimated using an unadjusted Cox proportional hazards model.

We next estimated the treatment effect using PS weighting, including IPTW and SMR weighted approaches. For IPTW, stabilized weights were used to reduce variance of the estimated treatment effect. The estimated weights were incorporated into a Cox regression model that only included the treatment variable.

Finally, we conducted an analysis that was stratified across deciles of the PS. We plotted the strata-specific estimates in an effort to identify the existence of systematic trends in the strength and direction of the estimated effects as a function of the estimated PS (Figure 2). A stratified summary estimate was calculated using unadjusted Cox regression with an indicator variable for each PS strata.

For comparison, we also estimated the unadjusted treatment effect using the crude outcome model and the multivariable-adjusted treatment effect by controlling for all covariates explicitly in an outcome model. These 2 outcome models do not implement any PS analysis and are, therefore, not of primary interest but serve for comparison with the performance of the PS methods. All analyses were conducted in SAS. Example code that can be used to conduct these analyses is provided in the Appendix in the online-only Data Supplement.

Results

We identified 947 004 patients initiating ACEi and 289 167 patients initiating ARBs. Table 1 shows the relative similarities between ACEi and ARB groups before PS adjustment. Comparability of a large proportion of the observed sample is also demonstrated by a large amount of overlap in the estimated PS distributions (Figure 1).

Despite the similarities in the characteristics between ACEi and ARB groups, adjusting for the covariates in Table 1 through PS matching and weighting improved balance of the observed characteristics. For example, before PS adjustment, diabetes mellitus was strongly associated with receiving an ACEi versus an ARB. Among patients with diabetes mellitus, 31% received an ACEi and 22% received an ARB. After PS weighting or matching, the proportion with diabetes mellitus is approximately equal between the 2 groups, with 22% of patients having diabetes mellitus among the ACEi and ARB groups after PS matching, 29% after IPTW, and 31% after SMRW (Table 1).

As discussed, different methods of PS implementation result in estimates that generalize to different populations. The characteristics of these different populations are evident in the difference in the means and frequencies of patient characteristics reported in Table 1. Using the example of diabetes mellitus, PS matching changes the proportion of patients with diabetes mellitus in the ACEi group to match that in the ARB group (Table 1). This pattern is expected because most ARB users could be matched to an ACEi user. SMRW adapted the proportion of individuals with diabetes mellitus in the ARB group to reflect the proportion in the unadjusted treatment (ACEi) population. IPTW resulted in a proportion that reflects the proportion of patients with diabetes mellitus in the overall population. Similar patterns are found for all covariates (Table 1).

Results of the estimated treatment effects are presented in Table 2. The unadjusted effect estimate resulted in an estimated HR of 1.77 (95% CI, 1.57–2.00). After adjustment for potential confounding factors, PS matching, IPTW, and SMRW all resulted in elevated HRs compared with the unadjusted estimate, as shown in Table 2 (matching: HR=1.91 [95% CI, 1.67–2.19]; IPTW: HR=1.87 [95% CI, 1.64–2.13];

Table 1. Covariate Distribution by Treatment Groups in the Overall Population, PS-Matched Population, and SMRW and IPTW Populations

	Overall		Propensity Score Matched		SMRW		IPTW	
	ARB (n=289 167)	ACEi (n=947 004)	ARB (n=288 401)	ACEi (n=288 401)	ARB (n=950 218*)	ACEi (n=947 004*)	ARB (n=289 919*)	ACEi (n=946 946*)
Patient characteristics								
Mean age, y (SD)	55.6 (13.3)	55.3 (13.9)	55.6 (13.3)	55.7 (13.3)	55.8 (25.3)	55.3 (13.9)	55.8 (13.8)	55.4 (13.8)
Male, %	48.0	52.5	48.0	47.7	52.2	52.5	51.2	51.5
Medicare, %	22.3	22.5	22.3	22.4	24.0	22.5	23.6	22.5
CVD management, %								
Angiography	0.1	0.3	0.1	0.2	0.4	0.3	0.3	0.3
Cardiac stress test	7.2	6.1	7.2	7.2	6.1	6.1	6.4	6.4
Echocardiograph	9.6	9.4	9.6	9.7	9.6	9.4	9.6	9.5
Mean lipid tests (SD)	0.59 (0.90)	0.61 (0.93)	0.59 (0.90)	0.59 (0.90)	0.61 (0.93)	0.61 (0.93)	0.59 (0.91)	0.60 (0.93)
Angioplasty	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1
Coronary stent placement	0.3	0.7	0.3	0.3	0.7	0.7	0.6	0.6
CABG	0.1	0.2	0.1	0.1	0.3	0.2	0.2	0.1
Comorbidities and acute events, %								
MI	0.2	0.6	0.2	0.2	0.8	0.6	0.7	0.5
MI in past 3 wk	0.1	0.4	0.1	0.1	0.6	0.4	0.5	0.3
Former MI	0.3	0.4	0.3	0.3	0.4	0.4	0.4	0.3
Unstable angina	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
Unstable angina in past 3 wk	0.2	0.6	0.2	0.3	0.7	0.6	0.6	0.5
Ischemic heart disease	6.2	6.7	6.2	6.1	7.2	6.7	7.0	6.6
Stroke	3.6	3.9	3.6	3.5	4.1	3.9	4.0	3.9
Diabetes mellitus	22.3	31.2	22.2	22.2	31.3	31.2	29.2	29.1
CKD	2.2	1.6	2.2	2.1	1.8	1.6	1.9	1.8
ESRD	0.7	0.5	0.7	0.6	0.5	0.5	0.6	0.5
Hypertension	55.2	44.3	55.2	54.7	44.4	44.3	46.9	46.9
Hyperlipidemia	27.6	26.7	27.6	27.7	26.3	26.7	26.6	26.9
Atrial fibrillation	1.6	1.8	1.6	1.6	2.0	1.8	1.9	1.8
Heart failure	2.0	2.8	2.0	1.9	3.1	2.8	2.9	2.6
Prevalent medication use, %								
Statins	25.7	27.3	25.7	25.9	25.8	27.3	25.8	26.9
Antiplatelets	3.3	3.1	3.2	3.2	3.2	3.1	3.2	3.2
Potassium-sparing diuretics	0.9	0.8	0.9	0.9	0.8	0.8	0.8	0.8
Loop diuretics	5.2	4.9	5.2	5.1	4.9	4.9	5.0	4.9
Niacin	1.3	1.2	1.3	1.3	1.2	1.2	1.2	1.3
Fibrates	3.4	3.5	3.3	3.3	3.3	3.5	3.3	3.4
Ezetimibe	4.2	3.5	4.2	4.2	3.4	3.5	3.6	3.7
Anticoagulants	2.6	2.7	2.5	2.5	2.7	2.7	2.7	2.7
Concurrent medication initiation, %								
Statins	5.8	11.4	5.8	5.8	11.5	11.4	10.2	10.1
Antiplatelets	0.5	1.3	0.5	0.6	1.5	1.3	1.3	1.1
Potassium-sparing diuretics	0.2	0.3	0.2	0.2	0.4	0.3	0.3	0.3
Loop diuretics	0.8	1.6	0.8	0.9	1.9	1.6	1.7	1.5
Niacin	0.3	0.4	0.3	0.3	0.5	0.4	0.4	0.4
Fibrates	0.7	1.2	0.7	0.7	1.3	1.2	1.1	1.1
Ezetimibe	1.1	1.1	1.1	1.1	1.3	1.1	1.3	1.1
Anticoagulants	0.2	0.6	0.2	0.3	0.7	0.6	0.6	0.5

ACEi indicates angiotensin-converting enzyme inhibitor; ARB, angiotensin receptor blocker; CABG, coronary artery bypass graft; CKD, chronic kidney disease; CVD, cardiovascular disease; ESRD, end-stage renal disease; IPTW, inverse probability of treatment weighted; MI, myocardial infarction; PS, propensity score; and SMRW, standardized mortality ratio weighting.

*Synthetic n values derived from weights.

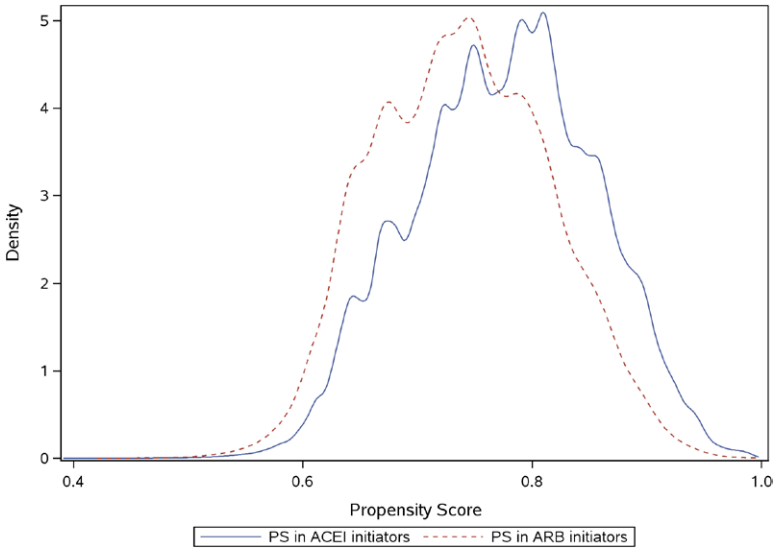

Figure 1. Estimated density of the propensity scores (PSs) among new users of angiotensin-converting enzyme inhibitors (ACEis) and angiotensin receptor blockers (ARBs).

SMRW: HR=1.86 [95% CI, 1.62–2.14]). The similarity among the estimates obtained from the different methods of PS implementation suggests that there is an absence of strong treatment effect heterogeneity. In other words, the risk of angioedema associated with ACEi use seems to be constant over various subgroups.

To further explore the impact of potential effect heterogeneity, we calculated treatment effect estimates within each of the 10 PS strata. The estimates are graphically depicted in Figure 2. The estimated HRs seem to fluctuate randomly around their mean, suggesting that no strong systematic treatment effect heterogeneity exists across values of the PS (Figure 2).

Conclusions

PS methods have become widely used tools for confounding control in nonexperimental studies of medical products and procedures. Regardless of the approach that one adopts to control for confounding in nonexperimental research, it is important that the researcher understands the underlying assumptions inherent in the chosen statistical method, as well as the interpretation of the results and the population(s) to which they are generalizable. We have described the assumptions necessary for valid PS analysis, the different treatment effects obtained by the various PS methods, and the populations to which these treatment effects are generalizable.

The validity of PS and multivariable outcome models require the strong assumption that all confounders are accurately measured and the exposure or outcome model is properly specified. However, PS methods provide several advantages over multivariable outcome models. First, they allow the researcher to identify patients who are never treated or untreated. These patients provide no information about treatment effects without making model assumptions that, if incorrect, could bias estimates of treatment effectiveness. Second, PS models require that analysts correctly model the effect of covariates on treatment, rather than the effect of covariates and treatment on the outcome. It may be difficult to correctly specify multivariable outcomes, particularly when treatment effects are heterogeneous across patient subgroups. In the setting of strong treatment effect heterogeneity, PS

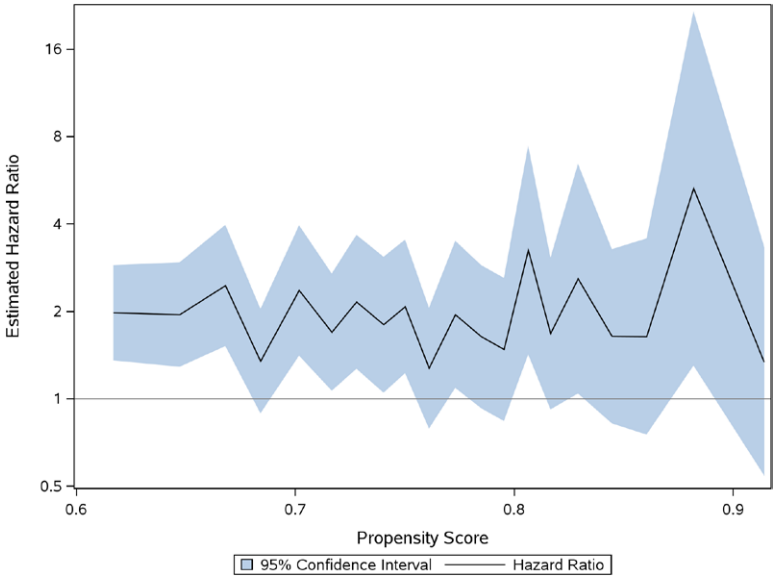

Figure 2. Estimated treatment effects and 95% confidence interval within deciles of the estimated propensity scores.

Table 2. Estimated Treatment Effects Comparing New Users of ACEi to New Users of ARBs on Risk of Angioedema After PS Adjustment

Model	Treatment	n	Events (%)	HR	95% CI
	ARBs	289 167	310 (0.1)	…	…
Crude (unadjusted)	ACEi	947 004	1713 (0.2)	1.77	1.57–2.00
Multivariable adjusted				1.87	1.65–2.11
	ARBs	288 401	309 (0.1)	…	…
PS matched	ACEi	288 401	601 (0.1)	1.91	1.67–2.19
	ARBs	950 218	938 (0.1)	…	…
SMRW	ACEi	947 004	1713 (0.2)	1.86	1.62–2.14
	ARBs	289 919	292 (0.1)	…	…
IPTW	ACEi	946 946	1751 (0.2)	1.87	1.64–2.13
	ARBs	289 167	310 (0.1)	…	…
Summary stratified	ACEi	947 004	1713 (0.2)	1.87	1.66–2.12

ACEi indicates angiotensin-converting enzyme inhibitor; ARB, angiotensin receptor blocker; CI, confidence interval; HR, hazard ratio; IPTW, inverse probability of treatment weighted; PS, propensity score; and SMRW, standardized mortality ratio weighting.

methods allow the researcher to estimate average effects of treatments in different populations without needing to explicitly specify the interactions in the model. Finally, in the usual setting of a common exposure and a rare outcome, researchers can construct much larger models of the PS. This is advantageous in studies using healthcare databases that provide a large number of weak confounders.[34]

As we have described, in the presence of treatment effect heterogeneity, different approaches to using the PS result in estimates of different treatment effects (contrasts). When deciding which particular PS approach to use, one should consider which treatment effect is of greatest interest and also whether the parameter can be reasonably estimated with the available data. For example, when comparing treated with untreated, it may be difficult to estimate ATE because there may be many untreated patients in the population who do not have an indication for treatment and therefore who would be rarely treated. In such situations, the ATT may be both more clinically relevant and also more reliably estimated. In studies comparing 2 candidate treatments (ie, comparative effectiveness research), ATE may be both easily estimated and the most useful to both clinicians and policymakers.

Despite the usefulness of PS methods in nonexperimental research, it should be noted that PS methods alone do not correct for errors introduced in the design or measurement of variables. For example, bias can be introduced by immortal person time (ie, a period of time in which exposed patients cannot experience the event because of the exposure definition),[42] selection bias,[43] control of causal intermediates,[39] and measurement error of the exposure or outcome. Many design issues can be addressed through the use of incident user designs and active comparators.[44] However, even with careful design and appropriate statistical adjustment, it is unlikely that all biases within healthcare database research can be completely addressed. Given the complexity of the underlying medical, sociological, and behavioral processes that determine exposure to medical products and interventions as well as the limitations of typical healthcare databases, there will often exist substantial uncertainty about how one should specify PS models to control confounding.[1]

Because of the inherent challenges in nonexperimental research, we suggest that researchers explore and report the sensitivity of results to changes in the epidemiological design and specifications of the statistical models. If the results are robust to such changes, the study more strongly supports the possibility that the estimates are indeed reflecting true causal relations.

Disclosures

Drs Brookhart and Stürmer are supported by National Institute on Aging (R01-AG023178 and R01-AG042845). The other authors report no conflicts.

References

1. Brookhart MA, Stürmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Med Care*. 2010;48(6 Suppl):S114–S120.
2. Walker AM. Confounding by indication. *Epidemiology*. 1996;7:335–336.
3. Kramer MS, Wilkins R, Goulet L, Séguin L, Lydon J, Kahn SR, McNamara H, Dassa C, Dahhou M, Masse A, Miner L, Asselin G, Gauthier H, Ghanem A, Benjamin A, Platt RW; Montreal Prematurity Study Group. Investigating socio-economic disparities in preterm birth: evidence for selective study participation and selection bias. *Paediatr Perinat Epidemiol*. 2009;23:301–309.
4. Glynn RJ, Schneeweiss S, Wang PS, Levin R, Avorn J. Selective prescribing led to overestimation of the benefits of lipid-lowering drugs. *J Clin Epidemiol*. 2006;59:819–828.
5. Glynn RJ, Knight EL, Levin R, Avorn J. Paradoxical relations of drug treatment with mortality in older persons. *Epidemiology*. 2001;12:682–689.
6. Winkelmayer WC, Levin R, Setoguchi S. Associations of kidney function with cardiovascular medication use after myocardial infarction. *Clin J Am Soc Nephrol*. 2008;3:1415–1422.
7. Simpson SH, Eurich DT, Majumdar SR, Padwal RS, Tsuyuki RT, Varney J, Johnson JA. A meta-analysis of the association between adherence to drug therapy and mortality. *BMJ*. 2006;333:15.
8. Brookhart MA, Patrick AR, Dormuth C, Avorn J, Shrank W, Cadarette SM, Solomon DH. Adherence to lipid-lowering therapy and the use of preventive health services: an investigation of the healthy user effect. *Am J Epidemiol*. 2007;166:348–354.
9. Dormuth CR, Patrick AR, Shrank WH, Wright JM, Glynn RJ, Sutherland J, Brookhart MA. Statin adherence and risk of accidents: a cautionary tale. *Circulation*. 2009;119:2051–2057.
10. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*. 2003;158:280–287.
11. Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41–55.

12. Leenen FH, Nwachuku CE, Black HR, Cushman WC, Davis BR, Simpson LM, Alderman MH, Atlas SA, Basile JN, Cuyjet AB, Dart R, Felicetta JV, Grimm RH, Haywood LJ, Jafri SZ, Proschan MA, Thadani U, Whelton PK, Wright JT; Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial Collaborative Research Group. Clinical events in high-risk hypertensive patients randomly assigned to calcium channel blocker versus angiotensin-converting enzyme inhibitor in the antihypertensive and lipid-lowering treatment to prevent heart attack trial. *Hypertension*. 2006;48:374–384.
13. Makani H, Messerli FH, Romero J, Wever-Pinzon O, Korniyenko A, Berrios RS, Bangalore S. Meta-analysis of randomized trials of angioedema as an adverse event of renin-angiotensin system inhibitors. *Am J Cardiol*. 2012;110:383–391.
14. Piller LB, Ford CE, Davis BR, Nwachuku C, Black HR, Oparil S, Retta TM, Probstfield JL; ALLHAT Collaborative Research Group. Incidence and predictors of angioedema in elderly hypertensive patients at high risk for cardiovascular disease: a report from the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *J Clin Hypertens (Greenwich)*. 2006;8:649–656; quiz 657.
15. Stojiljkovic L. Renin-angiotensin system inhibitors and angioedema: anesthetic implications. *Curr Opin Anaesthesiol*. 2012;25:356–362.
16. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550–560.
17. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*. 2003;14:680–686.
18. Stürmer T, Rothman KJ, Glynn RJ. Insights into different results from different causal contrasts in the presence of effect-measure modification. *Pharmacoepidemiol Drug Saf*. 2006;15:698–709.
19. Guo S, Fraser MW. *Propensity Score Analysis: Statistical Methods and Applications*. Thousand Oaks, CA: SAGE Publications; 2010.
20. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol*. 2006;163:262–270.
21. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168:656–664.
22. Cole SR, Hernán MA, Robins JM, Anastos K, Chmiel J, Detels R, Ervin C, Feldman J, Greenblatt R, Kingsley L, Lai S, Young M, Cohen M, Muñoz A. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *Am J Epidemiol*. 2003;158:687–694.
23. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23:2937–2960.
24. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *Am J Epidemiol*. 2011;173:761–767.
25. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61:962–973.
26. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163:1149–1156.
27. Patrick AR, Schneeweiss S, Brookhart MA, Glynn RJ, Rothman KJ, Avorn J, Stürmer T. The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmacoepidemiol Drug Saf*. 2011;20:551–559.
28. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*. 2006;59:437–447.
29. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods*. 2004;9:403–425.
30. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf*. 2008;17:546–555.
31. Ellis AR, Dusetzina SB, Hansen RA, Gaynes BN, Farley JF, Stürmer T. Confounding control in a nonexperimental study of STAR*D data: logistic regression balanced covariates better than boosted CART. *Ann Epidemiol*. 2013;23:204–209.
32. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997;127(8 pt 2):757–763.
33. Johannes CB, Koro CE, Quinn SG, Cutone JA, Seeger JD. The risk of coronary heart disease in type 2 diabetic patients exposed to thiazolidinediones compared to metformin and sulfonylurea therapy. *Pharmacoepidemiol Drug Saf*. 2007;16:504–512.
34. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20:512–522.
35. Bhattacharya J, Vogt WB. Do instrumental variables belong in propensity scores? *NBER Work Pap Ser*. 2007. http://www.nber.org/papers/t0343. Accessed September 5, 2013.
36. Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM, Glynn RJ. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol*. 2011;174:1213–1222.
37. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10:37–48.
38. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution–a simulation study. *Am J Epidemiol*. 2010;172:843–854.
39. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*. 2009;20:488–495.
40. Lunt M, Solomon D, Rothman K, Glynn R, Hyrich K, Symmons DP, Stürmer T; British Society for Rheumatology Biologics Register; British Society for Rheumatology Biologics Register Control Centre Consortium. Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. *Am J Epidemiol*. 2009;169:909–917.
41. Brown NJ, Ray WA, Snowden M, Griffin MR. Black Americans have an increased rate of angiotensin converting enzyme inhibitor-associated angioedema. *Clin Pharmacol Ther*. 1996;60:8–13.
42. Suissa S. Immortal time bias in pharmaco-epidemiology. *Am J Epidemiol*. 2008;167:492–499.
43. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615–625.
44. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol*. 2003;158:915–920.

KEY WORDS: epidemiologic methods ■ propensity score

Circulation
Cardiovascular Quality and Outcomes

Propensity Score Methods for Confounding Control in Nonexperimental Research

M. Alan Brookhart, Richard Wyss, J. Bradley Layton and Til Stürmer

Circ Cardiovasc Qual Outcomes. 2013;6:604-611; originally published online September 10, 2013;
doi: 10.1161/CIRCOUTCOMES.113.000359
Circulation: Cardiovascular Quality and Outcomes is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 1941-7705. Online ISSN: 1941-7713

SUPPLEMENTAL MATERIAL:

Example SAS code

```
/*************************************************************************/
/* This code demonstrates estimating a propensity score, calculating weights,   */
/* evaluating the distribution of the propensity score by treatment group, and  */
/* evaluating treatment effect heterogeneity over the distribution of the       */
/* propensity score.                                                            */
/*                                                                              */
/* This program written in SAS 9.2 TM, February 2013                            */
/*                                                                              */
/* Prepared by Bradley Layton, PhD at the University of North Carolina at       */
/* Chapel Hill                                                                  */
/*************************************************************************/

/*************************************************************************/
/* Variable Definitions:                                                        */
/*                                                                              */
/* x = binary [0,1] treatment variable                                          */
/* y = binary [0,1] outcome variable                                            */
/* c1 - c5 = binary [0,1] covariates                                            */
/* y_dur = time until censoring for event y                                     */
/*************************************************************************/



/*****************************************/
/* Estimating a propensity score           */
/*****************************************/


/* Modeling treatment = 1 given covariates and outputting data with the propensity score
   into a new dataset, 'PS_DATA'
   PS = estimated propensity score          */

PROC LOGISTIC DATA=raw_data DESCENDING;
    MODEL x = c1 c2 c3 c4 c5;
    OUTPUT OUT=ps_data PROB=ps;
    TITLE "Estimation of the propensity score from measured covariates";
RUN;



/*****************************************/
/* Evaluating the PS distribution          */
/*****************************************/


/* Creating PS treatment groups for plotting */

DATA ps_data;
    SET ps_data;

    IF x = 1 THEN treated_ps = ps;
        ELSE treated_ps = .;

    IF x = 0 THEN untreated_ps = ps;
        ELSE untreated_ps = .;
```

```
RUN;


/* Plot the overlap of the PS distributions by treatment group

   Turn on ODS output to get high quality graphics saved as an image file
   PLOTS=ALL gives you multiple plots. If you only want the overlay plot,
   use PLOTS=DENSITYOVERLAY */

ODS GRAPHICS ON;

PROC KDE DATA=ps_data;
    UNIVAR untreated_ps treated_ps / PLOTS=densityoverlay;
    TITLE "Propensity score distributions by treatment group";
RUN;

ODS GRAPHICS OFF;



/*****************************************/
/* Calculating PS weights                */
/*****************************************/


/* Calculating the marginal probability of treatment for the stabilized IPTW */

PROC MEANS DATA=ps_data(keep=ps) NOPRINT;
    VAR ps;
    OUTPUT OUT=ps_mean MEAN=marg_prob;
RUN;

DATA _NULL_;
    SET ps_mean;
    CALL SYMPUT("marg_prob",marg_prob);
RUN;


/* Calculating weights from the propensity score */

DATA ps_data;
    SET ps_data;

    *Calculating IPTW;

    IF x = 1 THEN iptw = 1/ps;
        ELSE IF x = 0 then iptw = 1/(1-ps);

    *Calculating stabilized IPTW;

    IF x = 1 THEN siptw = &marg_prob/ps;
        ELSE IF x = 0 THEN siptw = (1-&marg_prob)/(1-ps);

    *Calculating SMRW;

    IF x = 1 THEN smrw = 1;
        ELSE IF x = 0 THEN smrw = ps/(1-ps);

    LABEL   ps = "Propensity Score"
```

```
            iptw = "Inverse Probability of Treatment Weight"
            sisptw = "Stabilized Inverse Probability of Treatment Weight"
            smrw = "Standardized Mortality Ratio Weight";

RUN;



/*****************************************/
/* Evaluating the weights and preparing for */
/* trimming if necessary                    */
/*****************************************/


/* Performing univariate analysis on the weight variables by treatment status
   to check for extreme weights */

PROC UNIVARIATE DATA=ps_data;
    CLASS x;
    VAR iptw siptw smrw;
    TITLE "Evaluating weights by treatment group";
RUN;


/* Identifying percentiles at the upper and lower extremes of the untreated and treated
    PS distributions for trimming, if needed.  If other percentiles are needed, they can
    be created in the OUTPUT statement either by using a predefined SAS percentile,
    or by creating one in PCTLPTS=" */

PROC UNIVARIATE DATA=ps_data NOPRINT;
    CLASS x;
    VAR ps;
    OUTPUT OUT=ps_pctl MIN=min MAX=max P1=p1 P99=p99 PCTLPTS=0.5 99.5 PCTLPRE=p;
    title "Distribution of Propensity Score for Statin use, by statin use";
RUN;


/* Labeling the percentiles at the lower extremes of the treated in macro variables which can be
    called later.
    Defining the minimum, 0.5th percentiles, and 1st percentile of the treated */

DATA _NULL_;
    SET ps_pctl;
    WHERE x = 1;
    CALL SYMPUT("treated_min",min);
    CALL SYMPUT("treated_05",p0_5);
    CALL SYMPUT("treated_1",p1);
RUN;


/* Labeling the percentiles at the upper extremes of the untreated in macro variables
   which can be called later.
   Defining the maximum, 99th, and 99.5th percentile of the untreated. */

DATA _NULL_;
    SET ps_pctl;
    WHERE x = 0;
    CALL SYMPUT("untreated_max",max);
    CALL SYMPUT("untreated_99",p99);
    CALL SYMPUT("untreated_995",p99_5);
```

```
RUN;


/* When applying PS weights to analyses, these defined percentiles can be applied to trim areas
   of non-overlap and individuals treated contrary to prediction.
   To trim non-overlapping regions of the PS distribution, include the following statement
   in the modeling procedure: WHERE &treated_min <= ps <= &untreated_max;
   To trim those treated contrary to prediction, include the following
   statement: WHERE &treated_05 <= ps <= &untreated_995
   Trimming percentiles can be moved in progressively as far as desired */


/*****************************************/
/* Checking for treatment effect         */
/* heterogeneity                         */
/*****************************************/


/* Stratify the PS distribution into deciles
   If a different number of strata are desired, change to the desired number, N, in
    the GROUPS statement
   If changing the number of strata, also change the number in the '%DO i=0 %TO N'
   statement below to N-1  */

/* Create a new dataset with the strata indicator variables, 'ps_strata'
   Create a new variable for the strata rank, 'ps_decile' */

PROC RANK DATA=ps_data
    GROUPS = 10
    OUT = ps_data;
    VAR ps;
    RANKS ps_decile;
RUN;


/* Evaluate the outcomes, PS, and weights by treatment in each strata
   Check for heterogeneity of treatment effect across the strata
   Create a macro to perform the descriptive statistics and estimate
   effect measure in each decile */

%MACRO deciles;
%DO i= 0 %TO 9;

PROC FREQ DATA=ps_data;
    WHERE ps_decile = &i;
    TABLE y*x;
    TITLE "Outcome Y by Exposure X in PS deciles in decile &i";
run;

PROC MEANS DATA=ps_data MIN MEDIAN MAX MEAN;
    WHERE ps_decile = &i;
    CLASS X;
    VAR ps iptw siptw smrw;
    TITLE "Distribution of PS and weights by treatment in decile &i";
RUN;

/* y_dur is the time until censoring for event y */

PROC PHREG DATA=ps_data;
    WHERE ps_decile = &i;
```

```
    CLASS x / desc;
    MODEL y_dur*y(0) = x;
    HAZARDRATIO  x;
    TITLE "Hazard ratio of Exposure X on Outcome Y in decile &i";
RUN;

%END;
%MEND deciles;


/* Call the macro to estimate frequencies, PS distributions, and effect measures in each strata

%deciles;

/* The weights can be applied to various modelling PROCs in SAS, or PS matching
    can be performed using any number of matching algorithms */
```