Paper 2522-2018

Application of Propensity Score Models in Observational Studies

Nikki Carroll, Kaiser Permanente Colorado

ABSTRACT

Treatment effects from observational studies may be biased as patients are not randomly allocated to a treatment group. Propensity score methods are increasingly being used to address this bias. After propensity score adjustment, the distribution of baseline covariates will be balanced between treated and untreated patients. This paper reviews variable selection, balancing the propensity score, sensitivity analyses and presentation of results for 5 different propensity score methods: covariate adjustment, stratification, inverse probability of treatment weighted (IPTW), stabilized IPTW, and matching. Strengths and limitations of each method are illustrated by estimating the effect of anti-hypertension treatment on survival in advanced stage non-small cell lung cancer patients.

INTRODUCTION

Multiple analytic questions on a recent project brought to light that we needed to fully understand propensity score methods and how to appropriately use them. We set out to confirm we were using the method appropriately and to build a repository of references that we could use to defend our analysis to other investigators and/or manuscript reviewers. This paper is a summary of our learnings as well as how we applied these methods to our study. We will review 5 different propensity score methods: covariate adjustment, stratification, inverse probability of treatment weighted (IPTW)/ stabilized IPTW, and matching. We will then review variable selection, balancing the propensity score, potential sensitivity analyses that may be performed, and show how we applied these methods to our project.

STUDY DESCRIPTION

Preclinical studies have suggested that Angiotensin System Inhibitors (Angiotensin Receptor Blockers, and/or Angiotensin-Converting-Enzyme Inhibitors: ASIs) improve tumor perfusion and chemotherapy delivery. We looked at the effect of ASIs on the survival in patients receiving carboplatin/paclitaxel with or without bevacizumab chemotherapy for advanced non-small lung cancer (Menter, 2016). This retrospective study included patients diagnosed between 2005 and 2011 with Stage IIIB/IV non-small cell lung cancer who received the chemotherapy combination of carboplatin and paclitaxel with or without bevacizumab as part of their first course therapy. Four Kaiser Permanente regions participated in the study: Colorado, Northern California, Northwest, and Southern California. Table 1 shows a subset of demographic and clinical characteristics of the cohort. Patients on an ASI were older, had more cardiovascular disease (CVD), had more peripheral vascular disease (PVD), were more likely to be diabetic and were more likely to be taking other anti-hypertensive medications.

Table 1. Characteristics of patients receiving an ASI

Characteristic, N (%)	ASI	Other
Total N	273	1,192
Age at diagnosis (Mean/Std)	68 (8)	62 (10)
Non-White	90 (33)	381 (32)
Male	139 (51)	621 (52)
Tumor Grade Poor/Undiff/Unk	55 (71)	225 (21)
Cardiovascular Disease Diagnosis (CVD)	124 (45)	253 (21)
Peripheral Vascular Disease (PVD)	88 (32)	172 (14)
Diabetes	111 (41)	315 (26)
Smoking, Ever	177 (65)	748 (63)

Characteristic, N (%)	ASI	Other
Other Anti-HTN Medication Use	126 (46)	320 (27)

PROPENSITY SCORE METHODS

PROPENSITY SCORE REVIEW

Treatment effects from observational studies may be biased since the patients are not randomly allocated to a treated or untreated group. Propensity score methods are increasingly being used to address this bias. The propensity score is the probability of treatment assignment given a set of observed baseline characteristics. It allows you to mimic some of the characteristics of a randomized controlled trial by using this probability of treatment to balance differences in baseline covariates. After appropriately adjusting for the propensity score, the distribution of observed baseline covariates will be similar between treated and untreated patients.

The propensity score is usually created in logistic regression by modeling the likelihood of receiving treatment. Covariates include all characteristics that could affect the probability of treatment but not the outcome of interest. The probability for each person from the logistic regression model is considered the propensity to receive the treatment. This propensity score is then incorporated into a model to analyze the association with the outcome. There are various methods for incorporating the propensity score into the analysis and are discussed in detail below.

WHY PROPENSITY SCORES WORK

Estimates may be biased when characteristics between groups are imbalanced or when the treatment effects are not constant across the values of characteristics. Even after adjustment with conventional methods, residual confounding may still exist (Faires, 2010; D'Agostino, 2007). Some studies have shown that propensity score adjustment may be better alternatives to logistic regression to control for imbalance and increase comparability between groups (Faires, 2010; Cepeda, 2003; Groenwold, 2011).

BUILDING YOUR PROPENSITY SCORE

When building your propensity score, include variables that are related to treatment selection but not your outcome (Brookhart, 2006). Variables that reflect clinical or demographic factors used to determine which treatment a patient receives is a great place to start (McDonald, 2013). Common advice is to be over-inclusive to avoid leaving out a confounding variable, however, the optimal selection of a model is likely not one where ALL variables are included (Faires, 2010; Brookhart, 2006).

We built our propensity score using the following logistic regression model template:

```
proc logistic data=asi_data descending;
  model asi = [list of variables related to treatment of ASI];
  output out = propscore (keep=phat [list of variables you want to keep]);
run;
```

BALANCING YOUR PROPENSITY SCORE

Once your propensity score is estimated, it's important to make sure the measured covariates are balanced in order to reduce overt bias (Harder, 2010). There are several ways to assess the balance including:

- Graphic of the propensity score distribution. The distribution of the propensity score between the two groups should overlap. Nonoverlapping distributions suggest that one or more baseline covariates are strongly predictive of treatment selection and the analyst should consider re-doing variable selections and/or performing a stratified analysis (Curtis, 2007).
- Standardized differences of each covariate between treatment groups. Standardized differences are used to quantify the magnitude of the difference between baseline characteristics of two groups. They are calculated differently depending on the method by which you are incorporating

your propensity score into your outcome model (e.g., matched analysis vs stratified vs IPTW). One limitation to the use of standardized differences is the lack of consensus as to what value of a standardized difference denotes important residual imbalance between treated and untreated subjects. Some researchers have proposed that a standardized difference of 0.1 or more denotes meaningful imbalance existing in the baseline covariates (Faires, 2010; Austin, 2009; Normand, 2001).

- Stratify by deciles or quintiles. Baseline characteristics can be compared by stratifying the propensity score by deciles or quintiles. (Curtis, 2007; Austin, 2008). A side-by-side boxplot within the quintiles is a great graphic representation of this method (Austin, 2008).

WHEN SCORES DON'T BALANCE

If graphics and standardized difference calculations suggest your propensity score is not balanced, it may be necessary to re-estimate your propensity score. Suggestions in modifying your propensity score model include:

- Add more covariates
- Delete covariates
- Add interactions
- Substitute a non-linear term for a continuous one (e.g., cubic spline)
- Change the standardized differences threshold – choose one consistent with your model having been correctly specified

This may be an iterative process before you find your balanced model.

PROPENSITY SCORE ADJUSTMENT

Authors often will use multiple methods to report their findings. Consistency between these methods can help strengthen the findings and conclusions. Discrepancies in the results may indicate residual confounding or sensitivity to the study population, analysis approach, or both (McDonald, 2013). Methods to analyze your results with the propensity score include:

- Covariate adjustment
- Inverse Probability of Treatment Weighted / Stabilized IPTW
- Stratification
- Propensity score matching

COVARIATE ADJUSTMENT

This is the method most commonly seen in the literature and the method to which most readers can relate. The propensity score is simply included as an adjustment variable in in your model. You can also include other small important observed covariates that may have a strong relationship with your outcome or variables with noted residual imbalance after building your propensity score (Faires, 2010; D'Agostino, 1998). This method may be more sensitive to whether the propensity score has been accurately estimated (Austin, 2011).

For our study, the following variables were included in the final balanced model for this method: Age, Gender, Health Plan, COPD, CVD, PVD, Diabetes, Other HTN medication use, and Beta Blocker medication use. The distribution of the propensity score can be graphically represented through a histogram (Figure 1) using code similar to the following:

```
proc univariate data=asi_data noprint;
  class asi;
  histogram phat / normal (color=red) nrows=2;
run;
```

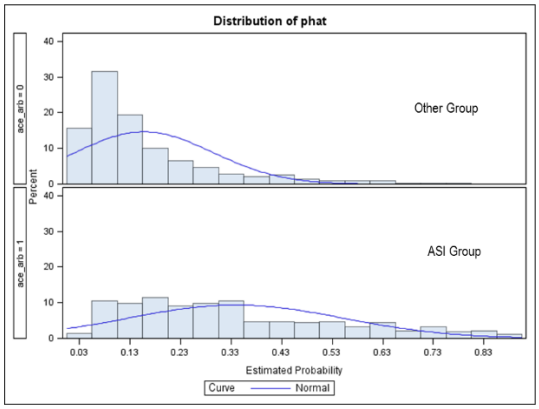

Figure 1. Distribution of propensity scores between Other and ASI groups

IPTW/STABILIZED IPTW

This method is used to estimate causal effects of treatments (Austin, 2011). One advantage of IPTW is that it requires fewer distributional assumptions about the underlying data, and it avoids the potential residual confounding that arises from stratification on a fixed number of strata (Curtis, 2007).

Stabilized weights address the situation when subjects with a very low probability of receiving the treatment creates weights that may be inaccurate or unstable (Austin, 2011). Stabilized weights might be the best option when analyzing using IPTW, however it can be tricky. Some weights are very large and thus influential, possibly resulting in a biased estimate of the treatment effect. Treated individuals with large weights should not be removed because those individuals are generally the best predictors of the outcome under comparison given that a large IPTW weight results from a small propensity score. To reduce the variability of the IPTW weights and give individuals with extreme weights less influence, Robins (2000) discussed a technique they referred to as stabilization. Stabilization is accomplished by multiplying the treatment and comparison weights by a constant equal to the expected value of being in the treatment or comparison groups, respectively. Because the IPTW weights in each group are multiplied by a constant, stabilization does not affect the point estimate of the treatment effect, but it does decrease the variance (Harder, 2010, Robins, 2000).

The following variables were included in the final balanced model for the IPTW method: Race/Ethnicity and Gender. There were other variables that we felt were important to include as additional adjustments in our model: Other HTN medication use, Age, Health Plan, and CVD.

STRATIFICATION

Cochran (1968) demonstrated that stratifying on the quintiles of a continuous confounding variable eliminated approximately 90% of the bias due to that variable. Rosenbaum and Rubin (1984) extended this result to stratification on the propensity score, stating that stratifying on the quintiles of the propensity score eliminates approximately 90% of the bias due to measured confounders when estimating a linear treatment effect (Austin, 2011, Cochran, 1968, Rosenbaum, 1984). When the propensity score has been correctly specified, the distribution of measured baseline covariates will be approximately similar between treated and untreated subjects within the same stratum (Austin, 2011). The most common approach is to divide subjects into 5 equal-sized groups using quintiles of the estimated propensity score (Austin, 2011). Increasing the number of strata used should result in improved bias reduction, although the marginal reduction in bias decreases as the number of strata increases (Austin, 2011). The effect of the treatment on outcomes can be estimated by comparing outcomes directly between treated and untreated subjects within strata. The stratum-specific estimates of treatment effect can then be pooled across stratum to estimate an overall treatment effect (Austin, 2011).

We chose to divide our subjects into 5 equal-sized groups using quintiles of the propensity score. Variables included in our final balanced model included: Age, Race/Ethnicity, Gender, COPD, CVD, PVD, Diabetes, Other Comorbid Diseases, Health Plan, Other HTN Medication use, and Beta Blocker

medication use. There are two methods we used to show balance between the groups for this method. Side-by-side boxplots (Figure 2) show the distribution of the propensity score within each quintile. Subjective observation shows that the propensity scores are fairly evenly distributed between the two groups within each quintile. The second method was to calculate standardized differences (Figure 3). Blue squares show the standardized different prior to propensity score adjustment and the orange circles show how the imbalance of the characteristics was abated after propensity score adjustment.

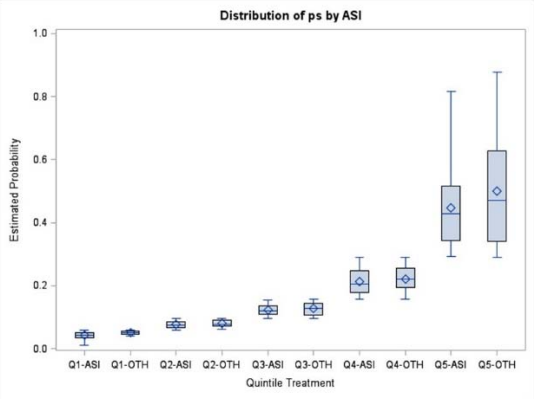

Figure 2. Distribution of propensity scores stratified by quintile and treatment group

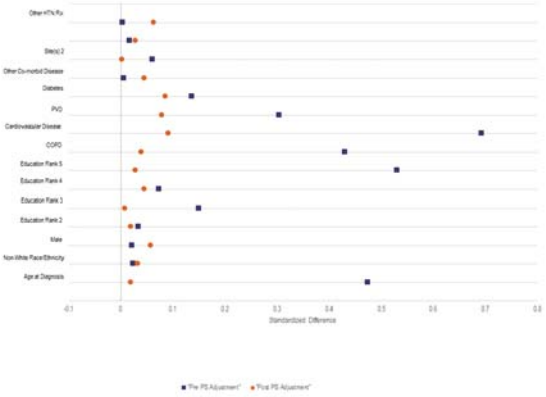

Figure 3. Standardized differences pre- and post-propensity score adjustment

PROPENSITY SCORE MATCHING

This method most closely mimics that of a randomized control trial. You can directly compare outcomes between treated and untreated subjects within the propensity score matched sample (Austin, 2011). Propensity score matching eliminates a greater proportion of the systematic differences in baseline characteristics between treated and untreated subjects than does stratification or covariate adjustment (Austin, 2011). Different methods of matching may introduce different types of bias. The use of nearest neighbor matching or optimal matching eliminates bias due to incomplete matching, because all treated subjects will be included in the matched sample (assuming that the number of untreated subjects is at least as large as the number of treated subjects). However, their use may result in the matching of more dissimilar subjects, and thus the estimated treatment effect may be contaminated by residual confounding. Caliper matching should result in the elimination of a greater degree of the systematic differences between treated and untreated subjects, but may introduce bias due to incomplete matching (Austin, 2009). To minimize the mean squared error of the estimated treatment effect, use the optimal caliper width of 0.2 times the standard deviation of the logit of the propensity score (Austin, 2011).

In some settings propensity score matching and IPTW removed systematic differences between treated and untreated subjects to a comparable degree; however, in some settings propensity score matching removed modestly more imbalance than did IPTW (Austin, 2011, Austin, 2009).

Variables in the final balanced model using propensity score matching included Age, Gender, CVD, Diabetes, and Other Anti-HTN medication use. We looked at the distribution of the propensity score using standardized differences (Figure 4) as well as the distribution of the propensity score between the groups in a histogram (Figure 5). Both figures show the propensity score corrected the imbalance between our groups. We matched 255 subjects which represented 93.4% of our ASI group and 21.4% of our Other group.

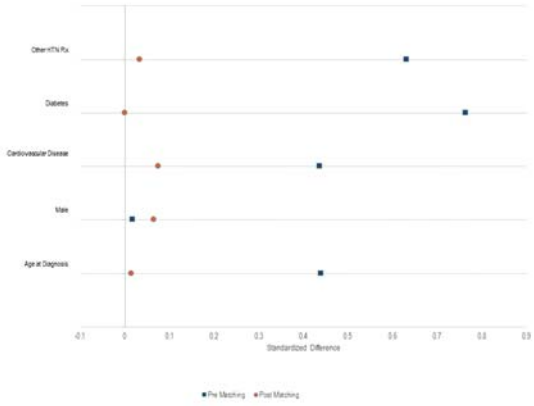

Figure 4. Standardized differences pre- and post-propensity score adjustment in matching model

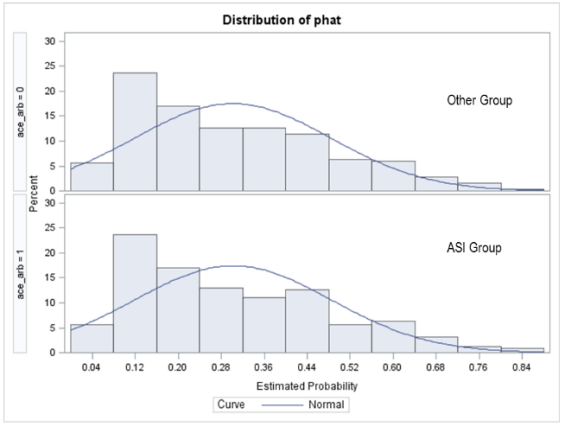

Figure 5. Distribution of propensity score for matching model

LIMITATIONS OF PROPENSITY SCORES

While these methods are powerful, there are several limitations: 1) they do not control for unobserved covariates (unless they are correlated with the observed covariates) (Faires, 2010); 2) they are only successful when there is substantial overlap between patient groups; 3) remaining unmeasured confounding may still be present; and 4) these methods cannot overcome initial selection bias.

POTENTIAL SENSITIVITY ANALYSES

Performing sensitivity analyses may inform how stable your methods are in adjusting for bias and may uncover hidden biases that may still remain. Some sensitivity analyses you may want to consider include:

- Examining the tails of the distribution of the propensity score and trim extreme weights to determine the amount of influence on your model (Curtis, 2007, Harder, 2010)
- Remove the subjects in non-overlapping regions of the distribution of the propensity score (Faires, 2010)
- If only performing a covariate adjustment, try doing a matching analysis. (Faires, 2010)
- Quantify the level of unmeasured confounding necessary to change the observed results (see details in Faires, 2010; Schneeweiss, 2006; Sturmer, 2005)

PRESENTING YOUR METHODS

Once you've built your propensity score and analyzed the data some suggestions to consider including in your write-up include:

- Showing the graphic representation of propensity score distribution overlap between your groups
- Listing all clinical variables used to generate your propensity score and how they were chosen
- What approach you used to balance the treatment groups
- Provide sufficient evidence that the two groups were balanced after propensity score adjustment
- Any sensitivity analyses you performed

Table 2 shows our final survival results for each method. Consistency across all methods confirms the stability of our analysis.

Table 2. Survival Results with all Propensity Score Methods

Model	Hazard Ratio	95% CI	p-value
Unadjusted	0.72	0.63-0.84	< 0.01
Covariate Adjustment	0.75	0.64-0.88	< 0.01
Stratification	0.73	0.62-0.86	< 0.01
IPTW Weighted	0.75	0.69-0.81	< 0.01
IPTW Stabilized	0.72	0.55-0.95	0.02
Matched Cohort	0.73	0.61-0.88	< 0.01

CONCLUSION

Propensity scores are one method to control for imbalances that may exist between treatment groups due to patients not being randomized. There are multiple propensity score methods that may be used and we've summarized these within this paper. In our study, we used propensity scores to adjust for imbalances that existed between our groups so the effect of treatment could be fully analyzed.

REFERENCES

Austin PC. Goodness of fit diagnostics for the propensity score model when estimating treatment effects using covariate adjusted with the propensity score. *Pharmacoepidemiology*. 2008:17:1202-1217

Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statist Med*. 2009:28:3083-3107

Austin PC. An introduction to Propensity Score Methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*. 2011:46:399-424

Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics.* 2011:10(2):150-161

Brookhart MA, Schneeweiss S, Rothman K, Glynn RJ, Avorn J, Sturmer T. Variable Selection for Propensity Score Models. *Am J Epidemiol*. 2006:163:1149-1156

Cepeda MSR, Boston JTF, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epi*. 2003:158(3):280-287

Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 1968:24, 295-313

Curtis LH, Hammill BG, Eisenstein EL, Kramer JM, Anstrom KJ. Using Inverse Probability-Weighted Estimators in Comparative Effectiveness Analyses with Observational Databases. *Medical Care*. 2007:45(10)2:S103-S107

D'Agostino Jr, RB, D'Agostino Sr, RB. Estimating treatment effects using observational data. *JAMA*. 2007:297(3):314-316

D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*. 1998:17(19):2265-2281

Faires DE, Leon AC, Haro JM, Obenchain RL, (Editors). Propensity score stratification and regression. Analysis of observational health care data using SAS. 1st. Edition. Cary, NC: SAS Institute, Inc.; 2010;23-46

Groenwold RHH, Vries F, Boer A, Pestman WR, Rutten FH, Hoes A, Klungel O. Balance measures for propensity score methods: a clinical example on beta-agonist use and the risk of myocardial infarction. *Pharmacoepidemiology and Drug Safety*. 2011:20:1130-1137

Harder VS, Stuart E, Anthony JC. Propensity Score Techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*. 2010:15(3):234-249

McDonald RJ, McDonald JS, Kallmes DF, Carter RE. Behind the numbers: Propensity Score Analysis – A primer for the diagnostic radiologist. *Radiology*. 2013:269(3):640-645

Menter AR, Carroll NM, Sakoda LC, Delate T, Hornbrook MC, Jain RK, Kushi LK, Quinn VP, Ritzwoller DP. Effect of Angiotensin System Inhibitors on Survival in Patients Receiving Chemotherapy for Advanced Non-Small Cell Lung Cancer. *Clin Lung Cancer.* 2016 Aug 20

Normand SLT, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, McNeil BJ. Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology.* 2001:54:387-398

Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in Epidemiology. *Epidemiology*. 2000:11, 550-560

Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 1983:45, 212-218

Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*. 1984:79, 516-524

Rubin D. Using propensity score to help design observational studies: application to the tobacco litigation. *Health Services & Outcomes Research Methodology.* 2001:2:169-188

Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology and Drug Safety*. 2006:15(5):291-303

Sturmer T Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology*. 2005:162:279-289

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Nikki Carroll
Kaiser Permanente Colorado
303-636-2455
nikki.m.carroll@kp.org