Propensity Score Matching for Multiple Treatment Comparisons in Observational Studies

Yuan Liu, Dana Nickleach, Joseph Lipscomb
Emory University, Atlanta, GA, USA
Corresponding author: Yuan Liu, email: yliu31@emory.edu

Abstracts

A major limitation of making inference about treatment effect based on observational data from a non-randomized study designs is the treatment selection bias, in which the baseline characteristics of the population under one treatment could dramatically differ from the other one. If not handled properly, such sources of heterogeneity will introduce confounding effects into a causal-effect relationship and result in bias in the estimation of treatment effect. The Propensity Score (PS) method is one of the approaches that have been widely used in practice to correct this selection bias through balancing the observed patients' characteristics among treatment groups. Until recently, the PS method has been applied exclusively for 2 treatment comparison settings (e.g. treatment vs. control) despite that it is frequently of interest to compare more than 2 treatments or interventions in medical and cancer research. PS covariate adjustment, inverse probability weighting (IPW) estimator, and PS matching are the three PS approaches commonly seen in two treatment comparisons, and among them, PS matching has been shown to have the greatest potential to eliminate the imbalance among covariates. However, not all of them are ready to be applied in the comparison of more than 2 treatments, especially for PS matching. To the best of our knowledge, we have not seen any such extension. In this study, we filled the gap and proposed an analytical approach to generalize PS matching for multiple (>= 2) treatments comparisons. This study was motivated by the desire to address comparisons of no adjuvant therapy, adjuvant chemotherapy alone and chemo-radiation therapy in resected pancreatic adenocarcinoma (rPAC) patients in a recent data analysis based on the National Cancer Data Base (NCDB). We present the proposed method and illustrate it in the above case study as well as compare it with other two PS approaches.

Key Words: propensity score, bias elimination, observational studies, matching

1. Introduction

A major limitation of making inference about treatment effect based on observational data from a non-randomized study design is the treatment selection bias, in which the baseline characteristics of the population treated by one intervention could systematically differ from that by the other intervention. Making direct estimate about treatment effect without taking such sources of heterogeneity into account introduces confounding effects into a causal-effect relationship and results in bias in the estimation. Complimentary to the conventional multivariable regression modeling, the Propensity Score (PS) method (Rosenbaum and Rubin 1983) is widely used in the observational studies to correct such selection bias. The propensity score is defined as a subject's *probability* of receiving a specific treatment assignment given the *observed* baseline covariates. A logistic regression model that predicts treatment assignment by observed baseline characteristics is commonly used to estimate the propensity score when two treatments are under investigation. The PS is also called balance score, and the beauty of it lies in fact that alignment of propensity scores across treatment groups balances baseline covariates accordingly, which enables us to assess treatment effect through fairly homogenous population groups. The process mimics what happens in randomized clinical trials (RCT),

in which the assignment of treatment is independent of the covariates. The result from RCT serves as the goal standard for estimating casual effectiveness. One major limitation of the PS method relative to RCT is that the unobserved confounders have no way to be adjusted. Many nice theoretical and practical reviews about this method are available elsewhere (Austin 2011; Agostino 1998; Joffe and Rosenbaum 1999; Pizer 2009).

Comparing treated vs. untreated populations is the most popular scenario that the PS method has been applied to in practice. The concept of generalized propensity scores generalizes PS to suite the situations where treatment could be a continuous dose or have multiple ordinal/categorical levels (Joffe and Rosenbaum 1999; Imbens 2000; Imai and van Dyk 2004). For the case of categorical treatments with k levels ($k > 2$), Imbens (2000) suggested using multinomial logistic regression to construct the predictive model of treatment assignment, by which each subject will have an estimated vector of PS, denoted as $\{P_1, P_2, \ldots, P_k\}$, which represents the probability of being assigned to each treatment given the covariates. With summation of one, only $k-1$ items are needed to carry out in the subsequent steps. The conventional PS estimated by a logistic regression model for two treatments can be viewed as a special case. Even though it is more suitable and desired by the scientific question of interest in medical studies, the utilization of generalized propensity score in practice is still limited.

PS covariate adjustment (PS-CA), inverse probability weighting (IPW) estimator, and PS matching are the three traditional PS approaches commonly seen in the two treatment setting, among which PS matching has been shown to have the greatest potential to eliminate the imbalance among covariates (Austin 2009a; Sekhon 2007). While it is straight forward to implement the PS-CA approach and IPW estimator for more than 2 treatments (Spreeuwenberg et al. 2010; Curtis et al. 2007), the matching approach of PS for more than 2 treatments has not been seen yet to the best of our knowledge. Motivated by a resent collaborative study that aimed to compare 3 treatments (no adjuvant therapy, adjuvant chemotherapy only and chemo-radiation therapy) in resected pancreatic adenocarcinoma (rPAC) patients, we extended the concept of nearest-neighbor-caliper-matching (Dehejia and Wahba 2002), already well-honed for the two-intervention case, and developed an algorithm that can be applied generally to any number of interventions under comparison by matching the generalized propensity scores.

This article is organized as follows. In the method section, we review the PS-CA approach and IPW estimator and propose the PS matching algorithm for the multiple treatment comparison. In the case study, we show the results by the three PS approaches along with the conventional multivariable regression model. The article is wrapped up by discussions.

2. Method

Suppose we have data $\{Y_i, X_i, T_i\}$, for $i = 1, 2, \ldots, N$, collected from a observational study for N subjects, in which Y_i is the outcome, X_i is the observed covariates, and T_i is the categorical treatment assignment with K ($K > 2$) levels. By multinomial logistic regression model, $\text{logit}(p_{ij}) = a_{0j} + a_{1j} * X_i$ and $j = 1,2, \ldots K$, for every subject we estimate the estimated probability of receiving each of treatment options given the observed covariates is denoted as $\{P_{i1}, P_{i2}, \ldots, P_{iK}\}$. The variables selected for the PS predicting model are those associated with the outcome (Brookhart et al. 2006; Austin, Grootendorst, and Anderson 2007). For the PS-CA approach, we follow the steps illustrated by Spreeuwenberg et al. (2010), and in the final step, $\{P_1, P_2, \ldots, P_{k-1}\}$ are added into the final model as covariates. The IPW estimator is a weighted analysis in which the weight

is the inverse probability of receiving the treatment actually received. If patient *i* received treatment *j*, then the weight is calculated as $w_i = 1/P_{ij}$ or $sw_i = Pr(T = j)/P_{ij}$, a stabilized weight when P_{ij} is extremely small and the number of subjects is extremely unbalanced among treatment groups (Cole and Hernán 2004; Sugihara 2010; Robins, Hernán, and Brumback 2000).

For the PS matching approach, we naturally extended the nearest-neighbor matching with caliper approach that has been widely used in the comparison of two treatments. The basic idea is to generalize the matching on a 1 dimensional line to a *K-1* dimensional space formed by PS vectors, and for a randomly generated origin, its nearest subjects from each treatment group within a certain radius (caliper) will form a matched group. After a matched group has been identified, it will be removed and the matching process will be continued until no more matches can be found. A detailed algorithm for $K = 3$ follows below:

(1) Uniformly generate M (e.g. M = 10000) origin points within the common support region formed by P1 and P2;
(2) For the i^{th} origin point , $i = 1,\ldots, M$, within the radius (caliper) from the origin, search for one nearest subject treated by treatment 1, one nearest treatment 2 subject, and one nearest treatment 3 subject to create a matched group. If success, remove the matched subjects from the region;
(3) Repeat step (2) for the $(i+1)^{th}$ origin point until reaching M.

The purpose of step (1) is to assure that each subject will have equal chance to be matched with others. However, the order of origin entering the matching process could result in different final matched groups. To allow such inherent sampling variation, we repeat the above algorithm B (e.g. B = 200) times, and construct a bootstrap confidence interval and p-value for the estimated treatment effect. The balance of covariates across treatment group after PS adjustment can be checked by pair wise standard differences (Austin 2009b).

3. Case Study

The study population consists of 7288 National Cancer Data Base (NCDB) incident cases of pancreatic adenocarcinoma (PAC) diagnosed in 1998-2002 who underwent surgical section of the primary PAC and had 5-year median follow up. Created in 1988, the NCDB, the largest disease-specific clinical registry in the U.S., contains detailed clinical, pathological, and demographic data on approximately 70% of all U.S. incident cancer cases. The purpose of this study was to investigate the differential impact on patient's overall survival (OS) by the three adjuvant therapies, which are no adjuvant therapy (NoAdjuvant), adjuvant chemotherapy only (ChemoOnly) and adjuvant chemo-radiotherapy (ChemoRad). The scientific finding has been accepted for publication in Annals of Surgical Oncology recently(Kooby et al. 2013).

To reduce the treatment selection biases to the greatest degree, we considered the propensity score matching method as proposed above, and the estimated treatment effect was compared to the ones from a conventional multivariable model, PS-CA approach and IPW estimator. The patients' demographic and disease characteristic variables that had an impact on OS were included in the PS predicting model by multinomial logistic regression. The Cox proportional hazard mode was mainly employed to assess the impact of the treatment on OS, and if matched sample was used, a stratified Cox model by the matched group was considered

4. Results

Among 7288 patients, 3094(45.5%) received no adjuvant therapy, 3596(49.3%) received ChemoRad, and 598(8.2%) had ChemoOnly, and there are systematic difference among these three groups, such as patients that received no adjuvant therapy are significantly older than the other two treatment groups; patients' tumor size is significantly smaller in the ChemoRad group; the cancer stage and grade are higher in the adjuvant therapy groups than the no adjuvant group, etc. (Due to the page limit, the results were not shown).

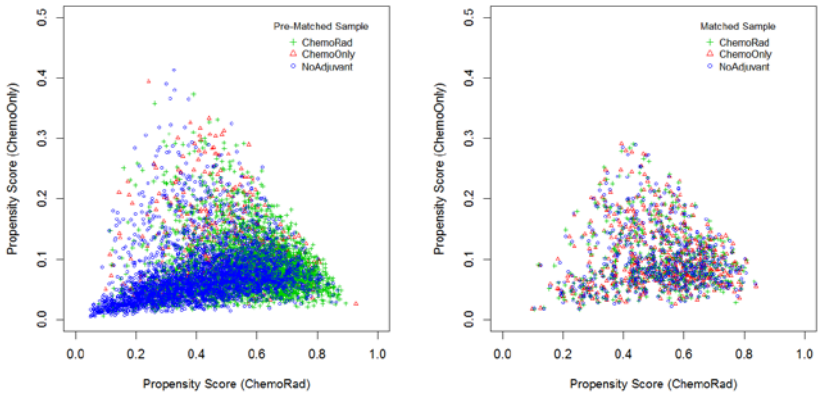

Figure 1 Distribution of (PS_ChemoRad, PS_ChemoOnly) in the pre-matched sample (7288 patients) and matched sample (1650 patients) by 1:1:1 matching using caliper 0.01.

In Figure 1, a PS 2-D space formed by PS_ChemoRad and PS_ChemoOnly is presented in the left panel. Note that by 1:1:1 matching, the maximum matched group size is 598 if all ChemoOnly patients can be matched. The right panel shows a matched sample by a caliper radius of 0.01. In the sensitivity analyses, a range of radius (caliper), from 0.005 to 0.03, was tested, and this range accounts for about 0.22% - 3.3% of the maximum distance of any two points in the left panel. As expected, a larger caliper links to a bigger number of matched groups (Table 1). However, the estimated hazard ratio and 95% confidence interval is quite stable by different calipers, and they all agree with the same conclusion that ChemoRad therapy improves the overall survival in resected PAC patients over those who received no adjuvant therapy or adjuvant chemotherapy alone.

Table 1 sensitivity analysis by different values of caliper.

Caliper	Median Number of Matched group	Hazard Ratio (95% CI) ‡	
		ChemoRad	ChemoOnly
0.005	359	0.71 (0.58, 0.83) **	1.02 (0.88, 1.20)
0.008	518	0.71 (0.61, 0.82) **	1.04 (0.91, 1.15)
0.01	557	0.70 (0.62, 0.78) **	1.03 (0.91, 1.15)
0.03	595	0.71 (0.63, 0.84) **	1.06 (0.94, 1.20)

** p-value < 0.001; * p-value < 0.05; ‡ NoAdjuvant group serves as reference level for hazard ratio, and confidence interval and p-value was constructed by 200 bootstrap samples.

In Table 2, we summarize the results by the other PS approaches as well as the conventional multivariable model. Without adjusting by any covariates or PS, the unadjusted model shows us no survival benefit by either adjuvant therapies, which is

obviously misleading as it ignores the systematic difference among the three groups in terms of other risk factors, while the remaining adjusted models agree with each other most of the time and confirm that ChemoRad therapy is associated with lower hazard of death. Some minor differences are among the adjusted models. If the baseline covariates are considered additionally in the IPW estimator or PS-CA approaches, the results don't differ from those by the conventional multivariable analysis, but by IPW or PS-CA alone, the benefit by ChemoRad relative to NoAdjuvant is diminished and ChemoOnly turns out to be significantly worse. By the PS matching approach, the hazard ratio for ChemoRad and ChemoOnly is the smallest, but conclusion would be the similar as those in the multivariable model.

Table 2 the comparisons of results from different PS approaches and conventional models.

Models	Hazard Ratio (95% CI)‡	
	ChemoRad	ChemoOnly
Unadjusted model	0.99 (0.94, 1.04)	1.40 (1.27, 1.54)**
Multivariable model	0.78 (0.74, 0.83)**	1.08 (0.98, 1.19)
IPW estimator	0.86 (0.81, 0.91)**	1.08 (1.01, 1.15)*
IPW estimator +	0.79 (0.74, 0.84)**	1.05 (0.99, 1.12)
PS-CA	0.84 (0.79, 0.89)**	1.12 (1.01, 1.23)*
PS-CA +	0.78 (0.74, 0.83)**	1.07 (0.97, 1.18)
PS matching (caliper = 0.01)	0.70 (0.62, 0.78)**	1.03 (0.91, 1.15)

+ Baseline covariates were included in the model; ** p-value < 0.001; * p-value < 0.05; ‡ NoAdjuvant group serves as reference level for hazard ratio.

5. Discussion

In this study, we generated an algorithm that extends the nearest-neighbor caliper PS matching approach to suite the scenario in which more than 2 treatments are of interest and showed its feasibility by a pancreatic cancer case study. The algorithm is not sensitive to the chosen size of caliper and is efficient when $K = 3$. The algorithm can even be applied to even a bigger number of K and different ratios of matching, it is computational intensive but not impracticable. In our future study, a simulation study will be carried out to compare the relative performance by the three PS approaches, and hence guide us how to implement them properly.

References

Agostino, Ralph B D. 1998. "TUTORIAL IN BIOSTATISTICS PROPENSITY SCORE METHODS FOR BIAS REDUCTION IN THE COMPARISON OF A TREATMENT TO A NON-RANDOMIZED CONTROL GROUP." *Statistics in Medicine* 2281 (19): 2265–2281.

Austin, Peter C. 2009a. "The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies." *Medical Decision Making : an International Journal of the Society for Medical Decision Making* 29 (6): 661–77.

———. 2009b. "Balance Diagnostics for Comparing the Distribution of Baseline Covariates Between Treatment Groups in Propensity-score Matched Samples." *Statistics in Medicine* 28 (25) (November 10): 3083–107.

———. 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research* 46 (3) (May): 399–424.

Austin, Peter C, Paul Grootendorst, and Geoffrey M Anderson. 2007. "A Comparison of the Ability of Different Propensity Score Models to Balance Measured Variables

Between Treated and Untreated Subjects: a Monte Carlo Study." *Statistics in Medicine* 26 (4) (February 20): 734–53.

Brookhart, M Alan, Sebastian Schneeweiss, Kenneth J Rothman, Robert J Glynn, Jerry Avorn, and Til Stürmer. 2006. "Variable Selection for Propensity Score Models." *American Journal of Epidemiology* 163 (12) (June 15): 1149–56.

Cole, Stephen R, and Miguel a Hernán. 2004. "Adjusted Survival Curves with Inverse Probability Weights." *Computer Methods and Programs in Biomedicine* 75 (1) (July): 45–9.

Curtis, LH, BG Hammill, and EL Eisenstein. 2007. "Using Inverse Probability-weighted Estimators in Comparative Effectiveness Analyses with Observational Databases." *Medical Care* 45 (10): 103–107.

Dehejia, RH, and S Wahba. 2002. "Propensity Score-matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84 (1): 151–161.

Feng, Ping, Xiao-Hua Zhou, Qing-Ming Zou, Ming-Yu Fan, and Xiao-Song Li. 2012. "Generalized Propensity Score for Estimating the Average Treatment Effect of Multiple Treatments." *Statistics in Medicine* 31 (7) (March 30): 681–97.

Imai, Kosuke, and David A van Dyk. 2004. "Causal Inference With General Treatment Regimes: Generlizing the Propensity Score." *Journal of the American Statistical Association* 99 (467) (September): 854–866.

Imbens, Guido W. 2000. "The Role of the Propensity Score in Estimating Dose-response Functions." *Biometrika* 87 (3) (September 1): 706–710.

Joffe, Marshall M, and Paul R Rosenbaum. 1999. "Invited Commentary: Propensity Scores." *Americqan Journal of Epidemiology* 150 (4): 327–333.

Kooby, David A, Theresa W Gillespie, Yuan Liu, Johnita Byrd-Sellers, Jerome Landry, John Bian, and Joseph Lipscomb. 2013. "Impact of Ajuvant Radiation Therapy on Survival Following Pancreastic Cancer Resection: An Appraisal of Data from the National Cancer Data Base." *Annals of Surgical Oncology*. In Press.

Kurth, Tobias, Alexander M Walker, Robert J Glynn, K Arnold Chan, J Michael Gaziano, Klaus Berger, and James M Robins. 2006. "Results of Multivariable Logistic Regression, Propensity Matching, Propensity Adjustment, and Propensity-based Weighting Under Conditions of Nonuniform Effect." *American Journal of Epidemiology* 163 (3) (February 1): 262–70.

Pizer, Steven D. 2009. "An Intuitive Review of Methods for Observational Studies of Comparative Effectiveness." *Health Services and Outcomes Research Methodology* 9 (1) (January 21): 54–68.

Robins, JM, MÁ Hernán, and B Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11 (5): 550–560.

Rosenbaum, PR, and DB Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Sekhon, Jasjeet S. 2007. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R." *Journal Of Statistical Software* VV (2): 1–51.

Spreeuwenberg, Marieke Dingena, Anna Bartak, Marcel a Croon, Jacques a Hagenaars, Jan J V Busschbach, Helene Andrea, Jos Twisk, and Theo Stijnen. 2010. "The Multiple Propensity Score as Control for Bias in the Comparison of More Than Two Treatment Arms: An Introduction from a Case Study in Mental Health." *Medical Care* 48 (2) (February): 166–74.

Sugihara, Masahiro. 2010. "Survival Analysis Using Inverse Probability of Treatment Weighted Methods Based on the Generalized Propensity Score." *Pharmaceutical Statistics* 9: 21–34.