January 2012

Comparing Propensity Score And Inverse Weighting Methods In A Longitudinal Time-To-Event Study

Kejia Zhu
Yale University, kejiazhu@gmail.com

YALE SCHOOL OF PUBLIC HEALTH

Comparing Propensity Score and Inverse Weighting Methods in a Longitudinal Time-to-event Study

by

Kejia Zhu

First Reader: Peter Peduzzi, PhD

Second Reader: Haiqun Lin, PhD

Division of Biostatistics, Yale School of Public Health, New Haven, CT

With Thanks to Ling Han, MD, MS

Program on Aging/Pepper Center Biostatistics Core, Department of Internal Medicine,

Yale School of Medicine, New Haven, CT

05/01/2012

Index

1 Abstract

We always strive to minimize the impact of bias in observational studies due to possible nonrandom treatment assignment. Propensity score and inverse weighting methods both attempt to achieve this goal. Inverse probability weighting is the method based on Horvitz and Thompson (1952) while propensity score is based on Rosenbaum and Rubin (1983). Because they are the most prevalent methods in longitudinal studies, these methods should be evaluated to find out which is better in reducing bias and producing accurate estimates. However, there are few studies comparing the two approaches. In a study of theory and simulated data, Ertefaie and Stephens (2010) demonstrated that, in simple cases, multivariate generalized propensity score (MGPS) routinely produced estimators with lower Mean-Square Error (MSE) when compared to inverse probability weighting (IPW). In the same paper, however, they were unable to show the same result in a longitudinal dataset. In this paper, I will perform similar comparisons in the treatment effect hazard ratio estimates as well as the efficiency of the estimates, specifically the variance of the two methods in an observational longitudinal public health study. I will only compare the direct effect of treatment, or the unconfounded and unmediated effect on expected response, since this is the only place where Propensity score and Inverse Weighting methods are comparable, and demonstrate that PS may not be the best method of analysis for reducing bias in longitudinal time-to-event studies, despite theoretical studies to the contrary. The results show that the treatment effect hazard ratio estimates with the two approaches are indistinguishable, although PS is consistently efficient while IPW varies based on whether stabilization occurs and on covariates.

2 Introduction

In a perfect world, the designers of experimental studies and controlled clinical trials would attempt to randomly assign subjects or patients to a condition or treatment group. However, in real-life studies, the “assignment” of a person to a treatment group or condition is usually not entirely random, unless specifically designed. Observational studies are, by design, not likely to be random in group or condition “assignment”. Prominent statisticians have spent years inventing and introducing methods to reduce the biases in experimental studies. And in the recent decade, competing methods have been slowly overtaken in popularity by weighting methods.

Even though inverse probability weighting (IPW) method was first introduced in the 1950s (Horvitz and Thompson, 1952), it was, for an extended period, widely considered inefficient relative to likelihood based methods (Clayton et al. 1998), and resulted in estimates sensitive to the precise form of the model for the probability of response (Little and Rubin, 1987), Robins and colleagues proposed improved IPW estimates that mitigated both problems in a series of papers in the 1990s, such as Robins et al. (1995), Robins and Rotnizky (1995), and Scharfstein et al (1999). IPW is a weight equal to the inverse of the probability of response by treatment group.

Another method of reducing bias is the propensity score (PS). PS, first proposed by Rosenbaum and Rubin (1983, 84), is defined as the conditional probability of receiving the treatment given pre-treatment covariates. IPW and PS are thus constructed in similar fashion. First, a model for treatment group or condition is fitted. The resulting conditional model for the outcome is then fitted either through weighting for IPW or matching for PS.

Despite the theoretical similarity of the two bias-reducing methods, there have been very few direct comparison studies between them. Ertefaie and Stephens (2010) found that PS outperformed IPW in mean squared error (MSE) in single and multiple interval simulation studies. However, no preference was found for either method using real data from Mother's Stress and Children's Morbidity Study, a small (N=167) longitudinal study. Tan (2007) also could not show superior efficiency of an IPW or IPW-like estimator over that of a regression estimator based on controlling for all pretreatment variables, essentially a PS estimator. Hirano et al. (2003) found the two estimators to be asymptotically equivalent.

In this paper, we compare PS and IPW in a large (n=5698) longitudinal time to event dataset from the Medicare Current Beneficiary Survey (MCBS). We will examine the performance indicators, such as the variance and MSE, of the estimators produced by the two methods for establishing the magnitude of the direct effect of treatment, without any of the confounding effects. The main focus is on direct effect of the treatment because it is the only situation where IPW and PS methods can be directly compared. The hazard ratio estimates of all other potential covariates for the effect of covariates are interesting to discuss but not directly comparable, because only IPW adjustment plays a role in estimating the effects of other covariates.

3 Methods

3.1 Propensity Score (PS)

As previously discussed, propensity score is a method of reducing bias in treatment effect estimation. At its most basic form, PS is defined for binary treatment as

$$\pi(x) = p(Tr = 1|x),$$

where $\pi(x)$ is the propensity score, Tr is treatment, and x is the covariate, by Rosenbaum and Rubin (1983) and is the most basic function of covariates that has the balancing property, which means treatment assignment is independent of covariates given the propensity score. This, of course, requires all confounding variables to be known as well as the existence of a real choice between treatment and control for each patient at the time of treatment selection, both critical criteria for what Rosenbaum and Rubin (1983) called the assumption of a strongly ignorable treatment assignment.

To produce the least biased propensity score model, it is important to not only include covariates that are correlated with treatment but also those correlated with outcome, as doing so would decrease the precision of the treatment effect estimate (Brookhart *et al*, 2006) Variables whose removal result in insignificant changes in estimated treatment effect and an increase in precision are seen as unlikely confounders and can be safely removed from propensity model (Hill and Kleinbaum, 2000). The average treatment effect can be computed from propensity score estimates using iterated expectation

$$\mu = \mathrm{E}[Y(1) - Y(0)] = \mathrm{E}_{\pi(X)}\{\mathrm{E}[Y(1)|\, \pi(X)] - \mathrm{E}[Y(0)|\, \pi(X)]\}$$

where $\mathrm{E}_{\pi(X)}$ is the expectation with respect to the distribution of $\pi(X)$ in the entire population (Ertefaie and Stephens, 2010)

The propensity scores produced can be used to find a conditional estimate of treatment effects given propensity score π, over the distribution of π. This can be best accomplished through matching between treatment and control patients, stratification, or using the PS directly as a covariate in the regression. Matching protects against misspecification of the propensity model but can significantly reduce sample size. Many existing user-generated programs and macros with numerical matching algorithms exist for SAS and other statistical analysis tools.

Stratification is similar to matching in effectiveness without the risk of losing subjects due to a lack of strong assumptions on time dependency of the effect of PS on the outcome. A quasi-standard of 5 strata exists for stratification based on the work of Cochran (1968). But Cochran also suggests that more than 5 strata should be used for larger datasets to further reduce imbalance. However, because stratification aims to produce treatment groups with similar probability of receiving treatment versus control, the individuals in the strata may be indistinguishable for further clinical decision making. (Curtis et al, 2007)

For our MCBS dataset, the estimated PS is used directly in the model as a covariate. It is an easy method to implement since the absolute standardized difference between the probability of outcome in the treated group and the probability of outcome in the untreated group can be determined (Austin 2008). But an incorrect assumption about the functional relationship between PS and outcome, such as assuming assignment to treatment group to be a prognostic factor by itself after controlling for other covariates, can negate the benefits of PS and lead to biased results (Rosenbaum and Rubin, 1983). Therefore, both the first step of propensity score model and the second step of establishing functional relationship between PS and the outcome need to be correct in order to correct bias.

In a dataset with time-varying covariates, a generalized propensity score method is more appropriate. If we set Y_{ij}, Tr_{ij}, and X_{ij}, as the response, treatment, and covariates of unit i at time j, respectively, then we can find π_{ij} as the propensity score. Moodie and Stephens (2010) then find, for every dose tr,

$$Y_{ij}\,(tr) \perp Tr_{ij} \,|\, \pi_{ij}\,(tr, X_{ij}),$$

and with it the $\mathrm{E}[Y_{ij}\,(tr) \,|\, \pi_{ij}\,(tr, X_{ij})]$ can be found as the unbiased estimate over the distribution of covariate X_{ij}.

Any proposed propensity score model can be adequate as long as balance is achieved, that is, the distribution of covariates X for different values of treatment Tr for each strata of PS π is approximately balanced. While any score that achieve balance will provide unbiased estimates of the treatment effect, the variance depends strictly on the definition of the PS (Ertefaie and Stephens 2010). In this case, variance is obtained using the standard model-based variance estimate. The results for the propensity score method were generated by Ling Han, MD, MS of the Yale Program on Aging/Pepper Center Biostatistics Core.

3.2 Inverse Probability Weighting (IPW)

The basic idea of IPW is conceptually easy to grasp and to program. A simple example is presented by missingdata.org.uk (2012)

Suppose we saw the following data,

Group	A			B			C		
Response	1	1	1	2	2	2	3	3	3

then the average response is 2. However if we observed

Group	A			B			C		
Response	1	?	?	2	2	2	?	3	3

then the average response is 13/6, which is biased. However, the probability of response is 1/3 in group A, 1 in group B and 2/3 in group C. We can therefore calculate a weighted average, where each observation is weighted by 1/{Probability of response}:

$$\frac{1\times\frac{3}{1}+(2+2+2)\times 1+(3+3)\times\frac{3}{2}}{\frac{3}{1}+1+1+1+\frac{3}{2}+\frac{3}{2}}$$

Thus, in this case, inverse probability weighting (IPW) has eliminated the bias by adjusting for measured data. More generally we will see it may still give biased parameter estimates, just less of them.

We can then expand the idea for treatment that is confounded. Unbiased estimates of the treatment parameter can even be obtained using a weighted analysis, assuming there are no unmeasured confounders. For a binary treatment (treated/untreated or treated/placebo), much like beta blocker use we have in our present data, for each subject *i* the weight $w_i = 1 + e^{-\eta_i}$ is assigned, where

$$\eta_i = \text{logit}\{p(Tr = tr_i \mid X = x_i)\} = \beta_0 + \beta_1 x_i$$

where tr_i is the observed binary treatment and x_i the observed confounder for subject *i*. This weight, essentially an ideological extension of the propensity score discussed earlier, can then be used in a weighted regression of Y on observed treatment T and components of X in order to find an average treatment effect.

While the estimated weight is theoretically asymptotically unbiased, in practice w_i is highly variable. This can be amended by stabilization, replacing the numerator of the weight with the marginal probability of receiving the treatment. The resulting stabilized weight, according to Robins (1997) is

$$sw_i = f\,(tr_i)\,/\,f\,(tr_i \mid x_i)$$

where (tr_i, x_i) are treatment and confounders as previously described, and f is the probability density function. The further expansion of this idea for multilevel and continuous treatment is discussed by Robins *et al* (2000), where the stabilized weight becomes

$$sw_i \propto f\,(tr_i)\,/\,f\,(tr_i \mid x_i),$$

provided that the initial logistic model calculating the weight is correctly specified. The Marginal structural models (MSM), a class of causal models for the estimation of the causal effect of a time-dependent exposure in the presence of time-dependent covariates that may be simultaneously confounder and intermediate variables from observational data, are consistently estimated by IPW estimators (Robins *et al*, 2000).

The IPW we already introduced, however, only accounts for confounders at baseline. To adjust for time-varying aspect of potential covariates, stabilized IPW for patient *i* at visit *j* are the product of inverse probability of exposure defined as

$$SW_{ij}^{X} = \prod_{k}^{j} \frac{f[X_{ik}|\bar{X}_{ik-1}, V_{i0}, \bar{C}_{ik-1} = \bar{0}]}{f[X_{ik}|\bar{X}_{ik-1}, \bar{L}_{ik-1}, \bar{C}_{ik-1} = \bar{0}]},$$

where $\bar{0}$ is a vector of zeroes, V_{i0} a vector of fixed baseline variables and $\bar{L}_{ik-1}$ a vector of time-varying variables. The denominator adjusts for bias while the numerator stabilizes. Therefore, unstabilized IPW for the same patients are just

$$NSW_{ij}^{X} = \prod_{k}^{j} \frac{1}{f[X_{ik}|\bar{X}_{ik-1}, \bar{L}_{ik-1}, \bar{C}_{ik-1} = \bar{0}]}.$$

The magnitude of nonpositivity bias increases with the number of time points and decreases with the use of appropriately stabilized weights (Cole and Hernán, 2008). In most cases, lack of stabilization results in larger variance estimates. Therefore, stabilization is generally preferred in order to achieve greater efficiency for no real cost.

IPW adjusted time-to-event analysis is a derivative of using such weights in controlling of confounding in analyzing survival data, as described by Robins et al (2000). A Cox regression model, weighted by estimated stabilized weights, accounts for cofounding by the covariate vector X because the “pseudo-population” created by weighting on the covariate X are unrelated to treatment Tr (Cole and Hernán, 2004). The use of the Cox model further removes the need for adjusting our IPW for the inverse probability of censoring. Variance of the hazard ratio estimate, obtained through a Cox model, is normally estimated through the robust variance estimator of Lin and Wei (1989) so that variance estimate is valid under null hypothesis and provides conservative confidence interval range. However, in this case we must utilize the sandwich estimator, similar to generalized estimating equations proposed by Zeger et al. (1988), accounting for the variability in estimating the weights. Based on Carpenter and Kenward (2006), if we write the estimating equation for β as $\Sigma^{n}_{i=1}\, \mu_i(\beta) = 0$, and the estimating equation for the logistic regression on the probability of observing X_{1i}, parameterized by α, as $\Sigma^{n}_{i=1}\, v_i(\alpha) = 0$. Then, let $w_i(\beta, \alpha) = (\mu^T, v^T)^T$. And finally, let

$$A = \sum_{i=1}^{n} \begin{pmatrix} \frac{\partial}{\partial \beta} \mu_i & 0 \\ 0 & \frac{\partial}{\partial \alpha} \nu_i \end{pmatrix}_{\hat{\beta}, \hat{\alpha}}$$

Then the sandwich estimator of the variance of treatment effect is the upper 4 x 4 block of

$$(A^{-1})\{\textstyle\sum_{i=1}^{n} w_i(\hat{\beta}, \hat{\alpha}) w_i^T(\hat{\beta}, \hat{\alpha})\}(A^{-1}).$$

SAS programming is used in both data transformation and analysis with this method.

4 Data and Analysis

4.1 MCBS

The dataset used in the analysis is a subset of 3752 patients between 2002 and 2006 in the large Medicare Current Beneficiary Survey (MCBS) that focuses on the effect of beta-blocker usage in patients with co-existing coronary artery disease and COPD or after myocardial infarction (MI) patients for combating adverse cardiovascular events. Even though randomized controlled trials have repeatedly shown beta-blockers to effectively protect the heart after MI according to the meta-analysis of Freemantle *et al.* (1999), this is an excellent opportunity to study the overall cardio-protective effectiveness of binary beta-blocker use in a strongly representative subset of the overall elderly population regardless of previous MIs. The subset includes patients from age 65 to 103, with the mean age being 78.39 years and median age 78, which is exactly what one would expect of the Medicare utilizing population.

Each individual had a variety of cardiovascular risk factors recorded, including sex, race, prior myocardial infarction, stent CABG, smoking status, etc. Based on enrollment date, each patient was observed for an entire year with no repeated measurement of the covariates and only seen when an adverse cardiac event occurs or at the end of the observational period. 1946 of the uncensored patients were followed for a further year with new measurements of the covariates under similar conditions. We use these data to determine treatment effect of beta-blocker use on cardiac health. Logistic regressions are used to fit the model for weights and propensity score over each interval over all relevant covariates, including the cardiovascular risk factors introduced earlier, as well as mobility, hypertension, diabetes, oral corticosteroid use, prior stroke, congenital heart defect, dementia, end stage renal disease, and a few others.

4.2 PS versus stabilized and unstabilized IPW

The resulting logistic models that include all the possible covariates, collectively shown as x_{ij}, for unit *i* and time *j* are

$$\text{logit}\{p(tr_{i1} = 1)\} = \alpha_0 + \alpha_1 x_{i1}$$

$$\text{logit}\{p(tr_{ij} = 1)\} = \beta_0 + \beta_1 tr_{i(t-1)} + \beta_2 c_{i(t-1)} + \beta_3 x_{i1}$$

for t =1 and t >1, where *c* is the response variable and *tr* is treatment.

We can then model our binary response, adverse cardiac symptoms, using cox regression with beta-blocker usage, all relevant time-varying covariates, and either PS as a covariate or weighting by IPW of treatment. The Cox proportional hazard model must then be generalized in order to fit the time-varying covariates. This is easily handled in theory for Cox regression models and in practice with modern statistical programming software such as SAS. Ties are resolved using the standard Breslow method for Cox regression and life model (Breslow, 1974). We compare the estimates and the variance of treatment effect of beta-blockers using PS, stabilized IPW, as well as unstabilized IPW. The results of the analysis are presented in Table 1 below.

Table 1: Parameters estimates based on IPW and PS for MCBS adverse cardiac events study.

		Hazard Ratio Estimate	Standard Error	95% Hazard Ratio Confidence Limits
Untabilized IPW	No Covariates	0.975	0.06411	(0.861, 1.105)
	All Covariates	0.985	0.08709	(0.830, 1.168)
Stabilized IPW	No Covariates	0.984	0.11878	(0.780, 1.242)
	All Covariates	1.029	0.11059	(0.828, 1.278)
PS		1.033	0.07487	(0.892, 1.197)

* Sandwich variance estimate used for all IPW method variances

^ Hazard Ratio Estimate compares beta-blocker users vs nonusers

We can see that the treatment effects are all not significant and very similar to each other. Because of the consistent overlap between estimated treatment effect confidence interval regardless of any of the methods listed, it is difficult to say that one is preferable to another in terms of estimation efficiency. In fact, even the general rule that stabilized IPW has lower variance was not true in our case. Adjusting for covariates did not make much of a difference in unstabilized IPW in terms of the hazard ratio estimate for beta-blocker users versus nonusers. But doing the same in stabilized IPW pushed the hazard ratio estimate in line with PS hazard ratio estimate. Unfortunately, the difference was still rather small and within the range of the confidence interval.

4.3 Bootstrap

To more accurately access the precision of the hazard ratio estimates of the treatment effect of beta-blocker use, we can use bootstrapping to find standard error. A bootstrap sample is an independent random sample of size *n* taken from dataset x with replacement. The bootstrap replication of statistic $\hat{\theta}$ = s(x) is $\hat{\theta}^*$ = s(x^*) and the bootstrap estimate of standard error is the observed standard deviation of repeated bootstrap replications. (Efron and Tibshirani, 1993) As the number of independent samples approaches infinity, $\hat{\theta}^*$ is approximately normally distributed and the bootstrap estimate of standard error approaches the estimate of the actual sample standard error. Precisely, the bootstrap estimate of standard error is calculated thusly,

$$\widehat{se}_k = \sqrt{\sum_{b=1}^{k} \frac{\left[\hat{\theta}^*(b) - \hat{\theta}^*(.)\right]^2}{k-1}}, \text{ where } \hat{\theta}^*(.) = \sum_{b=1}^{k} \frac{\hat{\theta}^*(b)}{k}.$$

For our data, we ran k=100 bootstrap sampling of the original dataset through the same models. Table 2 illustrates the comparison between model estimates and bootstrap estimates of hazard ratio for beta-blocker users versus nonusers and standard errors for all methods in Table 1.

Table 2: Comparison between model estimates and bootstrap estimates of hazard ratio for beta-blocker users versus nonusers and standard errors

		Hazard Ratio Estimate	Standard Error	Bootstrapped Hazard Ratio Estimate	Bootstrapped Standard Error
Unstabilized IPW	**No Covariates**	0.975	0.06411	0.976	0.06598
	All Covariates	0.985	0.08709	0.989	0.07608
Stabilized IPW	**No Covariates**	0.984	0.11878	0.983	0.07895
	All Covariates	1.029	0.11059	1.034	0.10037
PS		1.033	0.07487	1.037	0.07877

* Sandwich variance estimate used for all IPW method variances

^ Hazard Ratio Estimates are for beta-blocker users versus nonusers

We can see here that bootstrapped hazard ratio estimates and standard errors are very close to the estimates the model of the original dataset produced. One noticeable difference was in the standard error of the unadjusted treatment effect of beta-blocker use, where the standard error dropped from a high of 0.11878 to be more in line with other standard errors. More importantly, we can state that the standard error of the treatment effect adjusted for all important covariates with the stabilized IPW method is higher than that through other methods and therefore less efficient in this case.

4.4 Truncated IPW

We may further adjust the IPW models with truncation methods as suggested by Cole and Hernán (2008). They indicate that the choice of the model used to construct weights may impact the results of the marginal structural model. This choice is based on an informal bias-variance tradeoff between inclusion of a sufficient number of flexibly modeled confounders in the weight model and well-behaved weights. Truncation methods allow us to explore this trade-off and can give us a more refined comparison between the hazard ratio for beta-blocker users versus nonusers and variance estimates obtained through stabilized and unstabilized IPWs. The simplest way to explore such a tradeoff is through progressive truncation, developed by Kish (1992). In this method, weights are truncated by resetting the values of weights greater or less than percentile p(100-p) to the value of the weights at percentile p(100-p). The comparison of truncation-adjusted IPW between stabilized and unstabilized weights for our study is presented below in Table 3.

Table 3: Comparison of truncated IPW between stabilized and unstabilized weights for MCBS adverse cardiac events study.

	Truncated Percentile	Mean (SD)	Min	Max	Hazard Ratio Estimate	SE
Unstabilized IPW	0, 100	1.18 (0.68)	0.11	11.46	0.985	0.0871
	1, 99	1.16 (0.54)	0.67	4.04	0.967	0.0795
	5, 95	0.95 (0.35)	0.86	2.38	0.951	0.0779
	10, 90	1.07 (0.21)	0.91	1.64	0.956	0.0743
	25, 75	1.01 (0.03)	0.97	1.05	0.949	0.0735
	50, 50	1.02 (0.00)	1.02	1.02	0.948	0.0734
Stabilized IPW	0, 100	0.42 (0.31)	0.01	4.48	1.029	0.1106
	1, 99	0.41 (0.26)	0.05	1.58	0.999	0.0981
	5, 95	0.40 (0.21)	0.10	0.93	0.983	0.0899
	10, 90	0.38 (0.16)	0.14	0.64	0.978	0.0830
	25, 75	0.37 (0.11)	0.23	0.51	0.973	0.0762
	50, 50	0.39 (0.00)	0.39	0.39	0.948	0.0734

* Sandwich variance estimate used for all IPW method variances

^ Hazard Ratio Estimates are for beta-blocker users versus nonusers

Here the first row of each section corresponds to the marginal structure model adjusted for covariates for unstabilized and stabilized weight, respectively. Similarly, the last row of each section corresponds to the previously not shown baseline-adjusted model, which has a weight of 1 for every subject, is why the hazard ratio estimates are the same with either method. We can see that precision of the estimate increases as truncation increases. However, bias also increases since we are truncating more of the weights. Therefore, in this case and assuming the marginal structural model estimate is unbiased, it is unlikely for the small gain in precision to outweigh the increase in bias.

5 Discussion and Summary

Past studies, such as Chamberlain (1987), Hahn (1998), Hirano et al. (2003), and Ertefaie and Stephens (2010) have collectively demonstrate the theoretical superiority of propensity score method over inverse probability weighting method in MSE, variance, efficiency, and bias removal, assuming correct model specification in general longitudinal studies. But no simulation seems to exist for show the same for longitudinal time-to-event data. A real dataset like our study on the MCBS impact of beta-blocker use on adverse heart event data shows that, in a large empirical public health dataset with longitudinal time-to-event binary outcome, could not conclusively determine PS method as having better performance than IPW methods. We were able to show that PS results in a numeric variance estimate of the treatment effect hazard ratio estimate that is similar to that using the unstabilized IPW method, while maintaining a hazard ratio estimate that is more similar to that using the stabilized IPW.

Unfortunately, there were significant overlaps in the confidence interval estimates, which did not allow a conclusion on the statistical superiority of one method versus the other, especially without simulation to show the accuracy of the estimates. We would have also preferred larger resampling of bootstrapping for the precision analysis. However, hardware limitations restricted the bootstrapping capacity of SAS for such a large dataset and 25 to 200 replications is generally seem as sufficient for estimating a standard error through bootstrapping. The treatment in our study, beta-blocker use, did not have a significant treatment effect on adverse cardiac events. We cannot be sure that a study with a significant treatment effect would not result in clearer distinction between PS and IPW methods according to theoretical projections. Furthermore, the limited of time points limited the usefulness of the stabilized IPW method. Future analyses with multiple time points could be of interest.

Direct implementation of PS as a covariate in the model is a simpler process than the multiple modeling and calculation required for both truncated and stabilized IPW. However, the current development of PS limits its use to estimation of direct hazard ratio estimate for treated versus untreated, whereas the marginal structural method with IPW allows for the estimation of effects of other covariates as well, which tells researchers additional information about the dataset. We would have also liked to include the same comparison with the PS matching method. But computational limits eliminated that plan. Therefore the comparison between PS and IPW methods is not necessarily simply for theoretical efficiency and accuracy but also for whether the study is only interested in the treatment or other covariates as well. Furthermore, current developments in doubly robust IPW and potential developments in PS that allows for estimation of total effects could improve the current weakness in each method, respectively, and make for interesting future comparisons.

References

Austin PC (2008), “Primer on Statistical Interpretation on Methods Report Card on Propensity-Score Matching in the Cardiology Literature from 2004-2006: A Systemic Review”, *Circulation: Cardiovascular Quality and Outcomes*, 1:62-67.

Breslow N (1974), "Covariance Analysis of Censored Survival Data”, *Biometrics*, 30:89.

Brookhart MA, Schneeweiss s, Rothman KJ, Glynn RJ, Avorn J, Stürmer T (2006), “Variable Selection for Propensity Score Models”, *American Journal of Epidemiology*, 163:1149-1156.

Carpenter JR and Kenwood MG (2006), “A Comparison of Multiple Imputation and Doubly Robust Estimation for Analyses with Missing Data”, *Journal of the Royal Statistical Society*, 169:3:571-584.

Chamberlain G (1987), “Asymptotic efficiency in estimation with conditional moment restrictions”, *Journal of Econometrics*, 34:305-334.

Clayton DG, Spiegelhalter D, Dunn G and Pickles A (1998), “Analysis of Longitudinal Binary Data from Multi-Phase Sampleing (with discussion)”, *Journal of the Royal Statistics Society, Series B (statistical methodology)*, 71-87.

Cochran WG (1968), “The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies”, *Biometrics*, 24:295-313.

Cole SR and Hernán MA (2004), “Adjusted Survival Curves with Inverse Probability Weights”, *Computer Methods and Programs in Biomedicine*, 75:45-49.

Cole SR and Hernán MA (2008), “Constructing Inverse Probability Weights for Marginal Structural Models”, *American Journal of Epidemiology*, 168:6:656-664.

Curtis LH, Hammill BG, Eisenstein EL, Kramer JM, Anstrom KJ (2007), “Using Inverse Probability-Weighted Estimators in Comparative Effectiveness Analyses with Observational Databases”, *Medical Care*, 45(10 Supl 2):S103-107.

Ertefaie A and Stephens DA (2010), "Comparing Approaches to Causal Inference for Longitudinal Data: Inverse Probability Weighting versus Propensity Scores", *The International Journal of Biostatistics*, 6:2:14.

Efron B and Tibshirani RJ (1993), *An Introduction to Bootstrap*, London: Chapman & Hall.

Freemantle N, Cleland J, Young P, and Mason J (1999), “Beta Blockade after Myocardial Infarction: Systemic Review and Meta Regression Analysis”, *British Medical Journal*, 318(7200):1730-1737.

Hahn J (1998), “On the Role of the Propensity Score in Semiparametric Estimation of Average Treatment Effects”, *Econometrica*, 66:2:315-331.

Heinze G and Jüni P (2011), “An Overview of the Objectives of and the Approaches to Propensity Score Analyses”, *European Heart Journal*.

Hernán MA, Brumback B, and Robins JM (2000), “Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men”, *Epidemiology*, 11:561-570.

Hill HA and Kleinbaum DG (2000), “Bias in Observational Studies”, In: Gail M, Benichou J, editors, *Encyclopedia of Epidemiological Methods*, Chichester: Wiley, 94-100.

Hirano K, Imbens GW and Ridder G (2003), “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score”, *Econometric*, 71:4:1161-1189.

Horvitz DG and Thompson DJ (1952), "A Generalization of Sampling without Replacement from A Finite Universe", *Journal of the American Statistical Association*, 47:663–685.

Kish L (1992), “Weighting of Unequal P_j”, *Journal of Official Statistics*, 8:183-200.

Lin DY and Wei LJ (1989), “The Robust Inference for the Proportional Hazards Model”, *Journal of American Statistics Association*, 84:1074-1078.

Little RJA and Rubin, DB (1987), *Statistical Analysis with Missing Data*, J Wiley & Sons, New York.

Missingdata.org.uk (2012), “Idea Behind Inverse Probability Weighting”

Moodie EEM and Stephens DA (2010), “Estimation of Dose-Response Functions for Longitudinal Data”, *Statistical Methods in Medical Research*, 00:1-20.

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

Rosenbaum PR and Rubin DB (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score", *Journal of American Statistics Association,* 79:516-524.

Robins JM (1997), “Causal Inference from Complex Longitudinal Data", In M. Berkane (Ed.*), Latent Variable Modeling and Applications to Causality*, 69-117, New York: Springer-Verlag.

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

Robins JM and Rotnitzky A (1995), “Semiparametric Efficiency in Multivariate Regression Models”, *Journal of the American Statistical Association,* 90:122-129.

Robins JM, Rotnitzky A, and Zhao LP (1995), “Estimation of Regression Coefficients when Some Regressors are Not Always Observed”, *Journal of the American Statistical Association,* 89:846-866.

Rubin DB (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Scharfstein DO, Rotnitzky A, and Robins JM (1999), “Adjusting for Nonignorable Drop-out Using Semiparametric Nonresponse Models (with discussion)”, *Journal of the American Statistical Association*, 94:1096–1146.

Tan Z (2007), “Comment: Understanding OR, PS and DR”, *Statistical Science*, 22:4:560-568.

Zeger SL, Liang KY, Albert PS (1988), “Models for Longitudinal Data: A Generalized Estimating Equation Approach”, *Biometrics*, 44:1049-1060.