Paper Number: RW05

Considerations in the Use of Propensity Scores in Observational Studies

Lawrence Rasouliyan, Estel Plana, Jaume Aguado
RTI Health Solutions, Barcelona, Spain

ABSTRACT

The use of propensity scores allows one to reduce the effects of confounding that can occur because of potential differences in the distribution of measured baseline characteristics between treatment groups in observational studies. Many considerations are important in the generation of the propensity score model, such as choosing baseline variables from a list of potential confounders, constructing an appropriate logistic regression model, assessing the balance of baseline variables between treatment groups, and evaluating the distribution of propensity scores between treatment groups. Additionally, once the propensity score model has been specified, several different methods can be used to incorporate the propensity score in the assessment of the treatment effect, including matching, stratification, inverse probability of treatment weighting, and covariate adjustment. These considerations will be discussed, and examples will be provided.

INTRODUCTION

When estimating the effect of a particular treatment on a particular outcome, the randomized controlled trial (RCT) has been well established as the "gold standard." In addition to these studies being designed and powered to test very specific hypotheses related to treatment, the process of randomization in RCTs makes certain that any observed treatment effect is not due to the influence of other extraneous variables that could potentially confound the relationship between treatment and outcome. Because these confounding variables, both measured and unmeasured, are approximately equally distributed across treatment groups, RCTs yield unbiased estimates of the treatment effect.

In observational or nonrandomized studies, treatment is determined by "naturalistic" processes (e.g., normal clinical practice) rather than by randomized assignment. Thus, patients who receive one particular treatment may differ in various underlying characteristics from patients who receive another treatment. Failure to adjust for patient characteristics that are associated with both the treatment and the outcome (i.e., confounding variables) may lead to biased results.

Propensity score (PS) methods offer a means by which one can reduce the effects of measured confounding variables in the interpretation of the treatment effect (Rosenbaum et al., 1983). Ultimately, a PS, which serves as an overall "balancing" score, is calculated for each patient based on his or her measured underlying characteristics (Austin, 2011). Outcomes are then compared for patients who have a similar distribution of these underlying characteristics across treatment groups. In doing so, the potential bias associated with the influence of confounding variables in the interpretation of the treatment effect is reduced.

THEORY AND ILLUSTRATIVE EXAMPLE

To illustrate the considerations in the use of PS methods, an example from an 8-year prospective, multicenter, observational study in oncology is used. The outcomes of interest are overall survival (OS), measured as time to death; progression-free survival (PFS), measured as time to progression or death; and overall response rate (ORR), measured as a dichotomous response to treatment.

To avoid any suggestion that the authors of this paper are making independent clinical inferences about a particular disease indication or about the performance of commercially available treatments, the generic term "cancer" is used to denote the disease, and the generic terms "blue pill" (BP) and "red pill" (RP) are used to denote the two possible therapeutic regimens. Furthermore, rather than using actual study data, a simulated data set is generated to mimic the general overall attributes of actual observational study results.

DATA SOURCE AND SIMULATION

Data were simulated for 3,255 cancer patients, of which 1,871 patients received BP and 1,384 patients received RP according to normal clinical practice. The study length was 8 years, and patients were followed from enrollment to end of study, death, or loss to follow-up.

Patient characteristics were simulated through random number generation using the RAND and RANDUNI functions in SAS. Continuous patient characteristic variables were simulated under the normal distribution with specified mean and standard

deviation. For categorical patient characteristic variables, each group was indicated if the randomly generated number under the uniform distribution was less than or equal to the specified percentage.

Time-to-event data related to OS and PFS were simulated from the Weibull distribution by using the RAND function in SAS. Weibull scale and shape parameters were varied by treatment group according to the risk profile of the patient so that higher risk patients were more likely to experience events at shorter times than lower risk patients. Similarly, ORR was simulated according to the risk profile of the patient. ORR was indicated if the randomly generated number under the uniform distribution was less than or equal to the specified percentage within the particular treatment group and risk stratum. Details on data simulation can be provided by request.

PATIENT CHARACTERISTICS

The initial set of patient characteristics variables to be considered in the propensity score model are as follows: age, sex, performance status (0 to 5 score used to quantify general well-being and activities of daily life), prognosis risk category, disease stage, number of extranodal sites, elevated lactate dehydrogenase (LDH) level, bone marrow involvement, geographic region, and center type. The distribution of these variables by treatment group is summarized in Table 1 in Appendix A.

Notable differences between treatment groups were observed for several patient characteristics. Patients who received RP were more likely to be younger, from Europe, have better performance status and prognosis risk category, have normal LDH levels, and be seen at a community center than patients who received BP. Given these differences, one could always argue that any treatment effect that may be observed could be due to the influence of these variables rather than the treatment itself. That is, it is not evenhanded to compare directly the effectiveness of BP to RP because the patient populations that receive each therapy are different from each other. Thus, it is important to adjust for these underlying differences between BP and RP patients to reduce the potential bias in estimating the treatment effect.

PROPENSITY SCORE DEFINITION

The PS is the probability of a patient receiving a certain treatment (versus receiving another treatment) given his or her underlying baseline characteristics. This probability is usually estimated from a multivariable logistic regression model given the dichotomous nature of receiving a treatment. In our example, we model the log odds of receiving BP as a function of the relevant independent variables as follows:

$$ln\left(\frac{P_{BP}}{1-P_{BP}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots \quad \text{(Equation 1)}$$

In the above equation, P_{BP} is the probability of receiving BP, β_0 is the intercept, and β_1, β_2, and β_3 represent the coefficients for patient characteristics X_1, X_2, and X_3, respectively. By definition, P_{BP} is equivalent to the PS. Solving for P_{BP}, the following formula is obtained for the PS:

$$P_{BP} = \frac{exp[\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots]}{1 + exp[\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots]} \quad \text{(Equation 2)}$$

Thus, once the final logistic regression model has been specified and the coefficients have been determined, the PS can be calculated for each patient by substituting the values of his or her underlying baseline characteristics. The PS ultimately summarizes all confounders into a single variable and is particularly useful for situations of rare outcomes. The question remains, however, which patient characteristics to select for inclusion in the PS model.

SELECTION OF PATIENT CHARACTERISTICS FOR PROPENSITY SCORE MODEL

The selection of covariates for inclusion in a multivariable logistic regression model is as much of an art as a science. Several factors should be considered, such as strength of association, linearity, plausibility, multicollinearity, and effect modification. Additionally, various statistically driven algorithms exist for selecting the final covariates.

The purpose of constructing a PS model is not to create a risk model. Rather, it is to control for confounding (variables associated with both treatment and outcome) when applied to estimate the treatment effect. Simulation studies have shown that variables that are unrelated to the treatment but are related to the outcome should always be included in the estimation of propensity scores (e.g., Brookhart et al., 2006). Including these variables increases the precision of the estimated effect of treatment without increasing bias. In contrast, including variables that are related to the treatment but not to the outcome can decrease precision of the estimated effect of exposure without decreasing bias. Hence, applying standard model-fitting strategies to produce the logistic regression model with the best fit may not be the most optimal approach to generate the PS model.

In our example, confounding variables would be those variables associated with treatment and also with any of the outcomes of interest (OS, PFS, and ORR). However, we want to make certain to include all relevant variables associated with any of the outcomes, regardless of their associations with treatment. Initially, we consider the entire set of characteristics presented in Table 1 (Appendix A), with the exception of prognosis risk category, which itself was computed as a linear combination of other patient characteristics.

One method of determining which patient characteristics are associated with the outcomes is to fit a series of univariable models for each outcome. For each outcome, each model consists of one patient characteristic as the independent variable. The resulting parameter estimates from each univariable model can then be evaluated to determine the degree to which that variable is related to the outcome of interest. In our analysis, we fit a series of univariable Cox proportional hazards models for OS and PFS (both time-to-event outcomes) and a series of univariable logistic regression models for ORR (dichotomous outcome). The SAS procedures mostly commonly used for fitting Cox proportional hazards models and logistic regression models are PROC PHREG and PROC LOGISTIC, respectively. The following SAS code demonstrates an example for each outcome for the geographic region variable (named REGION in the data set). These models are to be repeated for each individual patient characteristic:

```
* Univariable models for each outcome as a function of geographic region
  Models to be repeated for each patient characteristic individually.;

* Univariable Cox Regression Model Predicting OS;
proc phreg data=ad01;
  class region (ref="North America");
  model osmo*osevt(0) = region;
run;

* Univariable Cox Regression Model Predicting PFS;
proc phreg data=ad01;
  class region (ref="North America");
  model pfsmo*pfsevt(0) = region;
run;

* Univariable Logistic Regression Model Predicting ORR;
proc logistic data=ad01;
  class region (ref="North America");
  model orr (event = "1") = region;
run;
```

Parameter estimates, standard errors, and *P* values from this univariable model exercise are presented in Table 2 in Appendix A. Some variables (e.g., elevated LDH level) were very strongly associated with the outcomes, while others (e.g., sex) were not. With regard to which variables have associations “strong enough” to be considered for inclusion in the PS model, no strict guidelines have been established. One could consider, for example, setting a threshold for the parameter estimates, standard errors, *P* values, and/or some combination of these values. Additionally, the researcher may also want to weigh the advantages and disadvantages of including too many or too few variables for the particular question at hand. Clinical input and the available sample size to estimate the PS may also be important in making such decisions.

In our example, we include a covariate (patient characteristic) in the PS model if at least one of the following is true for at least one of the outcomes: (1) the absolute value of the parameter estimate is at least 0.2 for at least one level of the covariate or (2) the *P* value is less than 0.01 for at least one level of the covariate. Applying these criteria to Table 2 (Appendix A), the following variables are included: age group, geographic region, performance status, disease stage, extranodal sites ≥ 2, and elevated LDH level.

PROPENSITY SCORE MODEL CONSTRUCTION

The PS model ultimately is the multivariable logistic regression model predicting receipt of BP as a function of patient characteristics per Equation 1. To generate this model in SAS, PROC LOGISTIC is employed. All patient characteristics identified in the previous exercise are categorical in nature and appear in the CLASS statement in addition to the MODEL statement.

Additionally, rather than using the resulting coefficients and calculating the PS for every patient individually per Equation 2, the OUTPUT statement in PROC LOGISTIC can be used to create a new data set (specified using OUT= syntax) with the PS for each patient included as an additional variable (specified using PRED= syntax):

```
* PS Model: Multivariable Logistic Regression Predicting Receipt of Blue Pill;
proc logistic data=ad01;
  class agegrp (ref="<=50") region (ref="North America") perfstat (ref="0")
        stage (ref="III") extranod (ref=">=2") ldhlevel (ref="Elevated") /
        param=ref;
  model tx (event = "BP") = agegrp region perfstat stage extranod ldhlevel;
  output out=ad02 pred=propensity;
run;
```

The above SAS code generates the PS model and outputs a data set named AD02 that contains a variable named PROPENSITY, which represents the PS value for each patient. Additionally, the following SAS output is generated:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.1882	0.1860	40.7886	<.0001
agegrp	51 to 60	1	1.1653	0.1600	53.0770	<.0001
agegrp	61 to 70	1	2.1720	0.1610	181.9532	<.0001
agegrp	>=70	1	3.1191	0.2114	217.6932	<.0001
region	Asia	1	0.5041	0.1274	15.6471	<.0001
region	Europe	1	-0.4297	0.0835	26.4601	<.0001
perfstat	1	1	0.6027	0.0856	49.5625	<.0001
perfstat	>=2	1	1.2876	0.2182	34.8199	<.0001
stage	IV	1	0.1179	0.0811	2.1121	0.1461
extranod	<2	1	-0.0882	0.0849	1.0814	0.2984
ldhlevel	Normal	1	-0.3877	0.0925	17.5754	<.0001

Most of the variables are very strongly associated with receipt of BP, as evidenced by magnitude of the parameter estimates and the very small *P* values. Because these variables were also strongly associated with treatment (Table 1 in Appendix A), they are confounders. Although disease stage and extranodal sites do not appear to be associated with receipt of BP, they remain in the PS model because they were shown to be associated with the outcomes of interest.

PROPENSITY SCORE TRIMMING

After the PS model has been established, it is important to assess the degree of overlap, or common support, in the PS distributions of treatment groups. This assessment may be done by creating stacked histograms of the PS distribution by treatment. For our example, stacked histograms for BP and RP patients are presented in Figure 1 in Appendix A.

The degree of overlap of distributions between treatment groups is related to the comparability between the two populations. In the extreme case where a complete separation of distributions between treatment groups is observed, the differences between the two populations would be too great for comparability, and PS methods may not be appropriate. However, in most situations (including our example), a portion of each distribution falls within a region of overlap. Limited overlap in PS distributions can result in decreased precision of the treatment effect. If limited overlap is observed, then the PS model could be modified by adding or removing variables to increase overlap and comparability. If limited overlap persists, then the PS approach for the treatment comparison at hand may need to be reassessed all together.

To ensure comparability of patients, a process of PS trimming is often employed. In this process, patients with extreme values of PS are excluded from subsequent analyses. No strict definition of “extreme” is universally accepted, and thresholds can vary from study to study or from distribution to distribution. One rule of thumb, for instance, could be to exclude all patients who fall outside the region of overlap. If we were to apply this rule of thumb to our example, we would exclude all patients who had a PS less than the minimum observed PS value among RP patients, and we would exclude all patients who had a PS greater than the maximum observed PS value among BP patients.

In our example, however, we apply a slightly more conservative approach. The minimum PS threshold is set as the value of the 1st percentile observed among BP patients (0.176 in our example), and the maximum PS threshold is set as the value of the 99th percentile observed among RP patients (0.873 in our example). Thus, all patients who have a PS less than 0.176 or greater than 0.873 are excluded from subsequent analyses. These PS values have been designated in Figure 1 in Appendix A with vertical lines. It is important to reiterate that the percentile values used in our example are arbitrary, and the researcher

could use other percentile values to be more conservative (e.g., 2.5th and 97.5th percentiles) or less conservative (e.g., 0.5th and 99.5th percentiles), as appropriate.

Of the 3,255 patients in our example, 317 (9.7%) were excluded from analysis by the PS trimming process. The percentages of patients excluded due to PS trimming were similar among BP (10.2%) and RP (9.1%) patients. The patients who were not excluded from PS trimming comprise the PS-trimmed cohort (N = 2,945) for inclusion in subsequent evaluation of the treatment effect.

APPLICATION OF THE PROPENSITY SCORE

The most common approaches of applying the PS to the data to adjust for confounding are (1) stratification, (2) matching, (3) inverse probability treatment weighting (IPTW), and (4) covariate adjustment. The commonality in these approaches is that outcomes are ultimately compared across treatment groups for patients with similar PS values. Each of these approaches, however, handles this comparison differently. While some of these approaches may be more preferable than others, depending on the particular study, question at hand, or researcher applying the methods, all four approaches are discussed in the context of our example.

STRATIFICATION

In the stratification approach, patients with similar PS values are grouped together into mutually exclusive strata. Typically, the PS values across both treatment groups are ranked and categorized into quintiles or deciles (in our example, we use quintiles). If the PS truly is a "balancing score," then patients within any particular stratum would be similar in their baseline characteristics. Thus, any estimate of the treatment effect within a particular stratum should not be influenced by measured underlying differences between BP and RP patients.

The following SAS code can be used to categorize the patients in the PS-trimmed cohort into five mutually exclusive quintiles. In the data set STRAT02, a variable named PSQUIN is created that has integer values ranging from 1 to 5 denoting the PS quintile.

```
* Categorize PS-Trimmed Cohort into PS Quintiles;
proc rank data=ad02 out=strat01 groups=5;
  where (pstrimcohort eq 1);
  var propensity;
  ranks psrank;
run;

data strat02;
  set strat01;
  psquin = (psrank + 1);
  label psquin = "Propensity Score Quintile (1 to 5)";
run;
```

Once the patients have been grouped into PS quintiles, it is important to assess whether the covariates are balanced between BP and RP patients. Assessing balance between treatment groups usually involves calculating a standardized difference between the two groups for each patient characteristic of interest. If this magnitude of the standardized difference is less than a specified threshold, then balance is considered to have been achieved (Lanehart et al., 2012). Additionally, the standardized difference has a desirable property in that it is not influenced by sample size and allows for comparison of variables measured in different units (Austin, 2011).

For continuous variables, the standardized difference (d) is calculated as follows:

$$d = \frac{(X_{BP} - X_{RP})}{\sqrt{\frac{s_{BP}^2 - s_{RP}^2}{2}}} \qquad \text{(Equation 3)}$$

In Equation 3, X_{BP} and X_{RP} represent the sample means of the patient characteristic for BP and RP patients, respectively, and s^2_{BP} and s^2_{RP} represent the sample variances of the patient characteristic for BP and RP patients, respectively.

The standardized difference (d) for categorical variables is calculated as follows:

$$d = \frac{(p_{BP} - p_{RP})}{\sqrt{\frac{p_{BP}(1-p_{BP}) + p_{RP}(1-p_{RP})}{2}}} \quad \text{(Equation 4)}$$

In Equation 4, p_{BP} and p_{RP} represent the proportion of patients with the characteristic for BP and RP patients, respectively.

In our example, all patient characteristics are categorical variables, and we calculate the standardized difference within each quintile. For categorical variables, the standardized difference is typically calculated for each level of the variable. It is important to note that for dichotomous variables, the standardized difference for one level is simply the negative of that for the other level. For higher level categorical variables, such a relationship is not observed.

We set a threshold of 0.25 to represent adequate balance of the covariate between BP and RP patients. Therefore, if the magnitude of the standardized difference is less than 0.25 within any given PS quintile, then we consider that an achievement of adequate balance has been achieved for that particular patient characteristic.

The below SAS code can be used to calculate the standardized difference between treatment groups. In this code, the standardized difference is calculated for the variable REGION in the first PS quintile, and the resulting values are outputted into the data set BALANCE03. This exercise is repeated for each patient characteristic within each PS quintile. This code, of course, can be incorporated into a SAS macro for efficiency:

```
* Code to Calculate the Standardized Difference of Categorical Variables
  Variable = region, PS Quintile = 1;
ods output crosstabfreqs=balance01 (where=(_type_ eq "11"));
proc freq data=strat02;
  where (psquin eq 1);
  table region*tx / norow nopercent;
run;

proc transpose data=balance01 out=balance02 prefix=prob;
  by region;
  var colpercent;
run;

data balance03 (keep=region stddiff);
  set balance02;
  p1 = (prob1/100);
  p2 = (prob2/100);
  stddiff = ((p1-p2) / sqrt(((p1*(1-p1))+(p2*(1-p2)))/2));
  label stddiff = "Standardized Difference";
run;
```

The resulting standardized difference values for all patient characteristic variables of interest are included in Table 3 in Appendix A. The smallest standardized difference magnitudes were observed in PS quintile 1 for performance status of 1 (d = –0.003) and for disease stage (d = 0.005). The largest standardized difference magnitudes were observed in PS quintile 5 for center type (d = 0.215), PS quintile 1 for sex (d = 0.223), and PS quintile 2 for center type (d = 0.240). It is important to note that some age group categories did not have any patients in some PS quintiles, as indicated in the table. Because all of the standardized difference magnitudes were less than 0.25, we can conclude that the PS model achieved adequate balance in patient characteristics between BP and RP patients for the purposes of applying PS stratification.

However, if it were decided that an adequate balance of covariates was not obtained, then the PS model should be refined. Options for refining the model to achieve a better balance may include the addition of more variables, interaction terms, or higher order terms (e.g., polynomial or logarithmic values of continuous variables).

Once it has been decided that an adequate balance of covariates has been achieved, the relative treatment effect can be estimated. For the time-to-event outcomes (OS and PFS), a Cox regression model is generated where the time to outcome is modeled as a function of treatment group and PS quintile (both as CLASS variables). Similarly, for the dichotomous outcome (ORR), a logistic regression model is generated as a function of treatment group and PS quintile. The following SAS code can be used to estimate the relative treatment effect:

```
* PS Stratification;
* Relative Treatment Effect (Hazard Ratio): OS;
proc phreg data=strat02;
  class tx (ref="BP") psquin;
  model osmo*osevt(0) = tx psquin / rl;
run;

* Relative Treatment Effect (Hazard Ratio): PFS;
proc phreg data=strat02;
  class tx (ref="BP") psquin;
  model pfsmo*pfsevt(0) = tx psquin / rl;
run;

* Relative Treatment Effect (Odds Ratio): ORR;
proc logistic data=strat02;
  class tx (ref="BP") psquin / param=ref;
  model orr (event = "1") = tx psquin / rl;
run;
```

It is important to note that other methods that yield similar results could be used to estimate the treatment effect (Austin, 2014). For instance, one could use the STRATA statement to indicate PS quintile and the perform regressions as a function of treatment group only. Alternatively, one could generate separate models by PS quintile and calculate a weighted estimate across strata.

The following abbreviated output was generated indicating the OS hazard ratio (RP vs. BP):

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
tx	RP	1	-0.26008	0.09251	7.9041	0.0049	0.771	0.643	0.924

The PFS hazard ratio (RP vs. BP) was as follows:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
tx	RP	1	-0.37440	0.06213	36.3140	<.0001	0.688	0.609	0.777

The ORR odds ratio (RP vs. BP) was as follows:

Odds Ratio Estimates and Wald Confidence Intervals

Effect		Unit	Estimate	95% Confidence Limits	
tx	RP vs BP	1.0000	2.622	1.879	3.659

These results indicate that patients who were given RP had prolonged OS (HR, 0.771; 95% CI, 0.643-0.924) and prolonged PFS (HR, 0.688; 95% CI, 0.609, 0.777) in addition to a higher likelihood of achieving ORR (OR, 2.62; 95% CI, 1.88-3.66) relative to patients who were given BP. Thus, based on this analysis, RP appears to have an advantageous effect relative to BP across all outcomes of interest after adjustment through PS stratification.

However, it is important to state that these relative treatment effects estimated in the above exercise are conditional (average effect at the individual level) rather than marginal (effect at the population level) in nature (Austin, 2014).

MATCHING

In the matching approach, each RP patient is matched to a BP patient who has a similar propensity score. Ultimately, outcomes are compared by treatment group while adjusting for these matched pairs. Various algorithms for matching patients are available. In our example, we apply the greedy matching algorithm described by Parsons (2004) with a 1:1 matching ratio. We denote RP patients as “cases” and BP patients as “controls” in the algorithm, given that our study has more BP patients than RP patients.

Upon application of the matching algorithm, 960 RP patients were matched to 960 BP patients, yielding a total matched sample size of 1,920 patients. Thus, from the PS-trimmed cohort of 2,945 patients, an additional 1,025 patients were excluded from the matching analysis due to lack of suitable matches. The output data set AD_MATCHING was created that contains a variable MATCHID to denote matched pairs.

To assess how well the PS matching process balanced the covariates, the standardized difference is computed on the matched sample using Equations 3 and 4. Results are included in Table 3 in Appendix A.

The smallest standardized difference magnitudes were observed for North American geographic region (d = 0.002), sex (d = 0.004), and disease stage (d = 0.002). With the exception of center type (d = 0.191), all variables had standardized difference magnitudes of less than 0.01, indicating a high degree of balance of patient characteristics across treatment groups for the application of PS matching.

To estimate the relative treatment effect for OS and PFS, one could perform a Cox regression stratified on matched pairs. However, a less biased approach to estimating the marginal treatment effect is to use a robust variance estimator that accounts for clustering within matched pairs (Austin, 2014). The SAS code for our example is as follows:

```
* PS Matching Accounting for Clustering Within Matched Pairs;
* Relative Treatment Effect (Hazard Ratio): OS;
proc phreg data=ad_matching covs(aggregate);
  id matchid;
  class tx (ref="BP");
  model osmo*osevt(0) = tx / rl;
run;

* Relative Treatment Effect (Hazard Ratio): PFS;
proc phreg data=ad_matching covs(aggregate);
  id matchid;
  class tx (ref="BP");
  model pfsmo*pfsevt(0) = tx / rl;
run;
```

For the ORR outcome, PROC GENMOD can be used to perform the logistic regression while accounting for clustering within matched pairs as follows:

```
* PS Matching Accounting for Clustering Within Matched Pairs;
* Relative Treatment Effect (Odds Ratio): ORR;
proc genmod data=ad_matching desc;
  class matchid;
  model orr = tx / dist=bin link=logit;
  repeated subject=matchid / type=un corrw covb;
  estimate 'RP vs. BP' tx 1 /exp;
run;
```

The following abbreviated output was generated indicating the OS hazard ratio (RP vs. BP):

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
tx	RP	1	-0.22413	0.10084	0.949	4.9402	0.0262	0.799	0.656	0.974

The PFS hazard ratio (RP vs. BP) was as follows:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
tx	RP	1	-0.41658	0.07008	0.987	35.3330	<.0001	0.659	0.575	0.756

The ORR odds ratio (RP vs. BP), denoted in the output as "Exp (RP vs. BP)," was as follows:

Contrast Estimate Results

Label	Mean Estimate	Mean Confidence Limits		L'Beta Estimate	Standard Error	Alpha	L'Beta Confidence Limits	
RP vs. BP	0.7234	0.6442	0.7906	0.9612	0.1874	0.05	0.5939	1.3286
Exp(RP vs. BP)				2.6148	0.4901	0.05	1.8110	3.7756

Similar to the results produced using the PS stratification approach, the results from PS matching indicated that RP appears to be advantageous relative to BP across all outcomes. Patients who were given RP had prolonged OS and PFS (as evidenced by the hazard ratios less than 1) and had a higher likelihood of achieving ORR (OR, 2.61; 95% CI, 1.81-3.78) than patients who were given BP.

INVERSE PROBABILITY TREATMENT WEIGHTING (IPTW)

Another means of applying the PS in adjustment for confounding is IPTW. In this approach, each patient is weighted by the inverse of the probability of receiving the treatment that he or she actually received. In the context of our example, BP patients would be assigned a weight equal to the inverse of the PS (remembering that the PS itself is the probability that the patient received BP), while RP patients would be assigned a weight equal to the inverse of 1 – PS. However, to obtain appropriate estimates of the variance, stabilized weights are commonly used (Xu et al., 2010). For each patient, the stabilized weight is calculated by multiplying his or her original weight by the proportion of patients who received the treatment that he or she received. The following SAS code was used to compute the stabilized weights in our example and assigns these values to the variable STWEIGHT:

```
* Calculating Weights According to IPTW;
data ad_iptw;
  set saslib.ad02;
  if (pstrimcohort eq 1);
  * pBP = Proportion of BP patients in study = 57.18%;
  pBP = 0.5718;
  * BP patients;
  if (tx eq 1) then stweight = (pBP / propensity);
  * RP patients;
  else if (tx eq 2) then stweight = ((1-pBP) / (1-propensity));
run;
```

To assess the balance of patient characteristics, Equations 3 and 4 can be applied to continuous and categorical variables, respectively, while accounting for the weights. The corresponding SAS code would be the same as the previous examples, except with the addition of the WEIGHT statement in PROC FREQ. The remainder of the code would not change. Using the previous example code of the REGION variable, the first part of the code was as follows.

```
* Code to Calculate the Standardized Difference of Categorical Variables
  In IPTW approach;
ods output crosstabfreqs=ad_iptw (where=(_type_ eq "11"));
proc freq data=strat02;
  weight stweight;
  table region*tx / norow nopercent;
run;
```

The standardized difference results based on IPTW are included in Table 3 in Appendix A. The smallest standardized difference magnitudes were observed for performance status of 1 (d = –0.003) and elevated LDH level (d = 0.004). With the exception of center type (d = 0.175), all variables had standardized difference magnitudes of less than 0.01, indicating a high degree of balance of patient characteristics across treatment groups for the application of IPTW.

To estimate the relative treatment effect, one could perform Cox regression (for OS and PFS) and logistic regression (for ORR) models while adjusting for stabilized weights as follows:

```
* IPTW with Stabilized Weights;
* Relative Treatment Effect (Hazard Ratio): OS;
proc phreg data=ad_iptw;
   class tx (ref="BP");
   model osmo*osevt(0) = tx / rl;
   weight stweight / normalize;
run;

* Relative Treatment Effect (Hazard Ratio): PFS;
proc phreg data=ad_iptw;
   class tx (ref="BP");
   model pfsmo*pfsevt(0) = tx / rl;
   weight stweight / normalize;
run;

* Relative Treatment Effect (Odds Ratio): ORR;
proc logistic data=ad_iptw;
   class tx (ref="BP") / param=ref;
   model orr (event="1") = tx / rl;
   weight stweight / normalize;
run;
```

The following abbreviated output was generated indicating the OS hazard ratio (RP vs. BP):

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
tx	RP	1	-0.19501	0.08476	5.2939	0.0214	0.823	0.697	0.972

The PFS hazard ratio (RP vs. BP) was as follows:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
tx	RP	1	-0.34704	0.05724	36.7541	<.0001	0.707	0.632	0.791

The ORR odds ratio (RP vs. BP) was as follows:

Odds Ratio Estimates and Wald Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
tx RP vs BP	1.0000	2.639	1.942	3.586

The results from IPTW yielded relative treatment effects in the same direction as PS stratification and matching, suggesting an advantageous effect of RP relative to BP. However, all hazard ratios and odds ratios tended to be closer to the null value.

COVARIATE ADJUSTMENT

The most elementary approach in applying the PS to adjust for differences in patient characteristics is simply to include the PS as a continuous variable in the model estimating the treatment effect. In our example, we perform either a Cox regression (for OS and PFS) or logistic regression (for ORR) while including the variable PROPENSITY as a covariate in the model, as follows:

```
* Covariate Adjustment of PS;
* Relative Treatment Effect: OS;
proc phreg data=ad02;
   class tx (ref="BP");
   model osmo*osevt(0) = tx propensity / rl;
run;

* Relative Treatment Effect: PFS;
proc phreg data=ad02;
   class tx (ref="BP");
   model pfsmo*pfsevt(0) = tx propensity / rl;
run;

* Relative Treatment Effect: ORR;
proc logistic data=ad_matching;
   class tx (ref="BP") / param=ref;
   model orr (event = "1") = tx  propensity / rl;
run;
```

The following abbreviated output was generated indicating the OS hazard ratio (RP vs. BP) for the parameter TX:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
tx	RP	1	-0.23289	0.09266	6.3169	0.0120	0.792	0.661	0.950
propensity		1	1.88894	0.26244	51.8040	<.0001	6.612	3.953	11.060

The PFS hazard ratio (RP vs. BP), indicated for the parameter TX, was as follows:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
tx	RP	1	-0.37662	0.06216	36.7105	<.0001	0.686	0.607	0.775
propensity		1	1.05791	0.16870	39.3271	<.0001	2.880	2.069	4.009

The ORR odds ratio (RP vs. BP), indicated for the parameter TX, was as follows:

Odds Ratio Estimates and Wald Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
tx RP vs BP	1.0000	2.606	1.869	3.633
propensity	1.0000	0.239	0.104	0.549

It is important to note that covariate adjustment of the PS as a continuous variable assumes a linear relationship between the propensity score and outcome of interest, which may not be true. Additionally, balance diagnostics of patient characteristics through the assessment of standardized differences are not directly applicable to this approach of using the PS as a continuous covariate. Given these limitations, covariate adjustment of the PS as a continuous variable is increasingly less common in these types of analyses.

COMPARISON OF APPROACHES

In addition to the four approaches used to adjust for PS, crude results were also generated on the original simulated data. In these results, univariable Cox regression (for OS and PFS) and logistic regression (for ORR) were performed modeling each outcome as a function of treatment only. The crude treatment effect for all outcomes along with the PS-adjusted results from all four approaches are presented together in Figure 2 in Appendix A. Crude treatment effect estimates were further away from the null than all estimates derived from PS methods.

CONCLUSION

Propensity score methods are a powerful tool for treatment comparison across two populations that may differ in underlying characteristics. Through the analysis of a simulated oncology study (in which the results mimicked those of an actual prospective, multicenter observational study), topics explored included assessment of patient characteristics, selection of variables for inclusion in the propensity score model, application of four approaches to adjust for propensity score, assessment of covariate balance between treatment groups, and estimation of the treatment effect while adjusting for propensity score. It is important to state, however, that propensity score methods adjust only for measured confounding variables. The influence of unmeasured confounders on the treatment effect after adjustment with propensity score methods cannot be directly quantified. Thus, the potential for residual bias is always present.

The results of our simulation indicated that for overall survival and progression-free survival, stratification, matching, and covariate adjustment yielded similar point estimates of the hazard ratio, while the point estimates for inverse probability treatment weighting tended to be closer to the null. For overall response rate, point estimates of the odds ratio were similar across all four approaches of propensity score adjustment. Crude point estimates of the treatment effect across all outcomes were further away from the null than all four approaches of propensity score adjustment, indicating that the application of propensity score methods attenuated the crude observed treatment effect.

REFERENCES

Austin PC. An Introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res. 2011 May;46(3):399-424.

Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. Stat Med. 2014 Mar 30;33(7):1242-58.

Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol. 2006 Jun 15;163(12):1149-56.

Lanehart RE, Rodriguez de Gil P, Kim ES, Bellara AP, Kromrey JD, Lee RS. Propensity score analysis and assessment of propensity score approaches using SAS® procedures [Paper 314-2012]. SAS Global Forum; 2012. Available at: http://support.sas.com/resources/papers/proceedings12/314-2012.pdf. Accessed August 30, 2016.

Parsons LS. Performing a 1:N case-control match on propensity score [Paper 165-29]. SAS Users Group International (SUGI) 29; 2004. Available at: http://www2.sas.com/proceedings/sugi29/165-29.pdf. Accessed August 30, 2016.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.

Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. Value Health. 2010 Mar-Apr; 13(2): 273–277.

CONTACT INFORMATION

The authors welcome questions and comments. Please direct inquiries to

Lawrence Rasouliyan | Director, Biostatistics | RTI Health Solutions | Email: lrasouliyan@rti.org

APPENDIX A – TABLES AND FIGURES

Table 1. Patient Characteristics by Treatment Group

Characteristic	Blue Pill N = 1871	Red Pill N = 1384	*P* value*
Age group, n (%)			
≤ 50	59 (3.2%)	234 (16.9%)	<0.001
51 to 60	576 (30.8%)	684 (49.4%)	
61 to 70	925 (49.4%)	412 (29.8%)	
> 70	311 (16.6%)	54 (3.9%)	
Sex, n (%)			
Male	834 (44.6%)	593 (42.8%)	0.326
Female	1037 (55.4%)	791 (57.2%)	
Geographic region, n (%)			
North America	846 (45.2%)	553 (40.0%)	<0.001
Europe	705 (37.7%)	702 (50.7%)	
Asia	320 (17.1%)	129 (9.3%)	
Performance status, n (%)			
0	1060 (56.7%)	1001 (72.3%)	<0.001
1	688 (36.8%)	352 (25.4%)	
≥ 2	123 (6.6%)	31 (2.2%)	
Prognosis risk category, n (%)			
Low risk	297 (15.9%)	228 (16.5%)	<0.001
Medium risk	684 (36.6%)	656 (47.4%)	
High risk	890 (47.6%)	500 (36.1%)	
Disease stage, n (%)			
III	1210 (64.7%)	851 (61.5%)	0.063
IV	661 (35.3%)	533 (38.5%)	
Extranodal sites, n (%)			
≥ 2	591 (31.6%)	417 (30.1%)	0.374
< 2	1280 (68.4%)	967 (69.9%)	
LDH level, n (%)			
Elevated	498 (26.6%)	290 (21.0%)	<0.001
Normal	1373 (73.4%)	1094 (79.0%)	
Bone marrow involvement, n (%)			
Yes	1165 (62.3%)	900 (65.0%)	0.106
No	706 (37.7%)	484 (35.0%)	
Center type, n (%)			
Academic	400 (21.4%)	197 (14.2%)	<0.001
Community	1471 (78.6%)	1187 (85.8%)	

LDH = lactate dehydrogenase.

* *P* values derived from the chi-square test.

Table 2. Univariable Models to Determine Association Between Candidate Patient Characteristics and Outcomes

Variable	OS			PFS			ORR		
	β	SE	P Value	β	SE	P Value	β	SE	P Value
Age group									
≤ 50	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
51 to 60	0.291	0.198	0.141	0.038	0.109	0.728	0.411	0.118	0.001
61 to 70	0.922	0.190	<0.001	0.511	0.106	<0.001	-0.189	0.105	0.070
> 70	1.057	0.207	<0.001	0.393	0.123	0.001	-0.806	0.128	<0.001
Sex									
Male	0.049	0.078	0.536	0.002	0.053	0.970	0.021	0.062	0.734
Female	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Geographic region									
North America	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Europe	-0.041	0.085	0.628	-0.058	0.057	0.307	0.093	0.085	0.274
Asia	0.167	0.114	0.144	0.138	0.078	0.076	-0.298	0.105	0.004
Performance status									
0	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
1	0.059	0.086	0.491	0.092	0.056	0.104	0.019	0.106	0.859
≥ 2	0.679	0.142	<0.001	0.423	0.108	<0.001	-0.334	0.160	0.036
Disease stage									
III	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
IV	0.349	0.086	<0.001	0.335	0.057	<0.001	-0.217	0.067	0.001
Extranodal sites									
≥ 2	0.209	0.082	0.011	0.152	0.055	0.006	-0.201	0.063	0.001
< 2	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
LDH level									
Elevated	0.673	0.081	<0.001	0.335	0.057	<0.001	-0.417	0.063	<0.001
Normal	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Bone marrow involvement									
Yes	-0.019	0.081	0.813	0.001	0.054	0.986	-0.018	0.064	0.780
No	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Center type									
Academic	0.187	0.096	0.051	0.129	0.065	0.047	0.068	0.082	0.405
Community	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref

β = model coefficient; LDH = lactate dehydrogenase; ORR = overall response rate; OS = overall survival; PFS = progression free survival; Ref = referent group; SE = standard error.

Note: OS defined as time to death; PFS defined as time to progression or death; ORR defined as dichotomous response to treatment.

Note: Each model for OS and PFS was fit with Cox regression, and each model for ORR was fit with logistic regression.

Figure 1. Stacked Histograms Depicting Distribution of Propensity Scores by Treatment Group

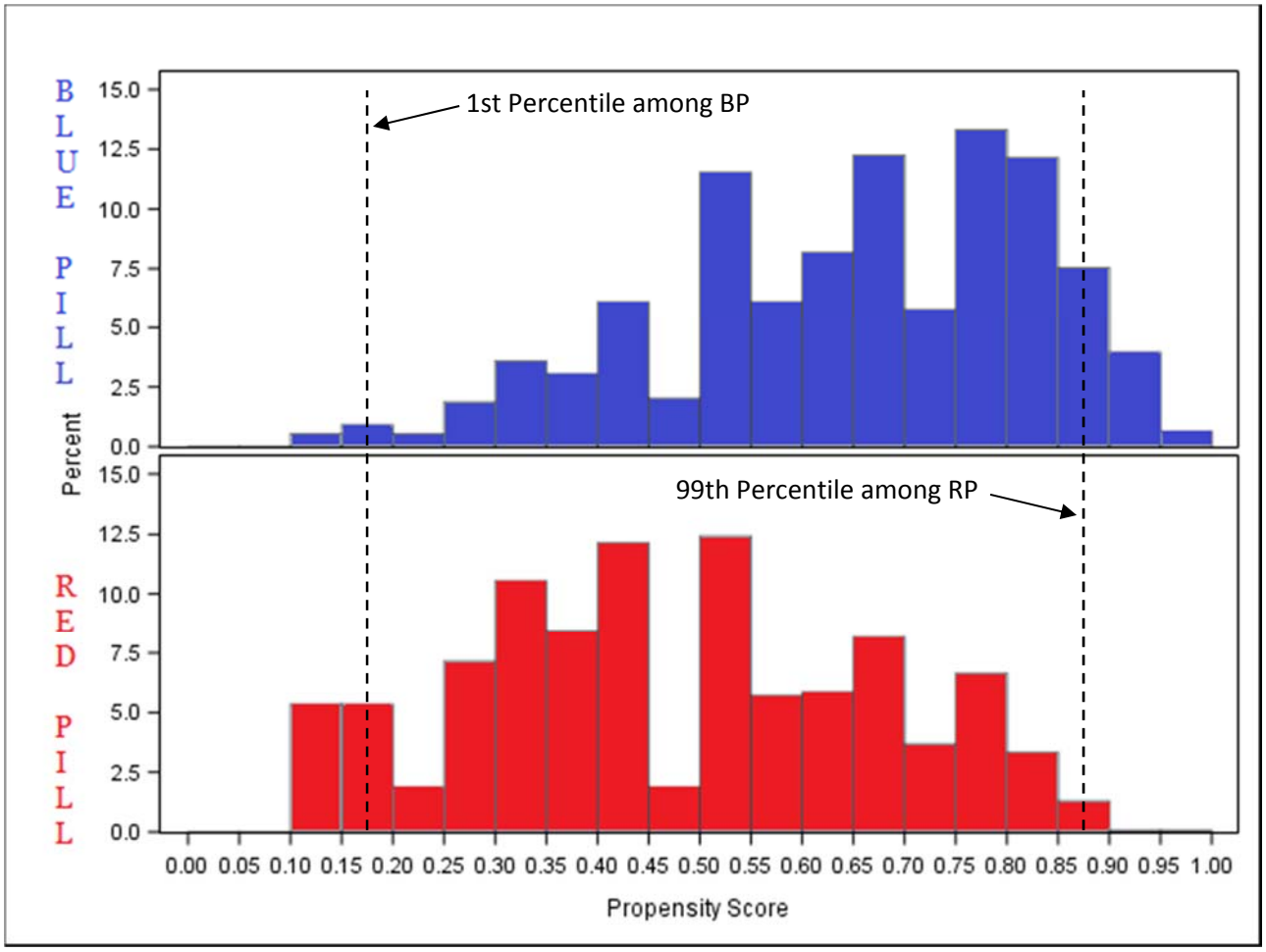
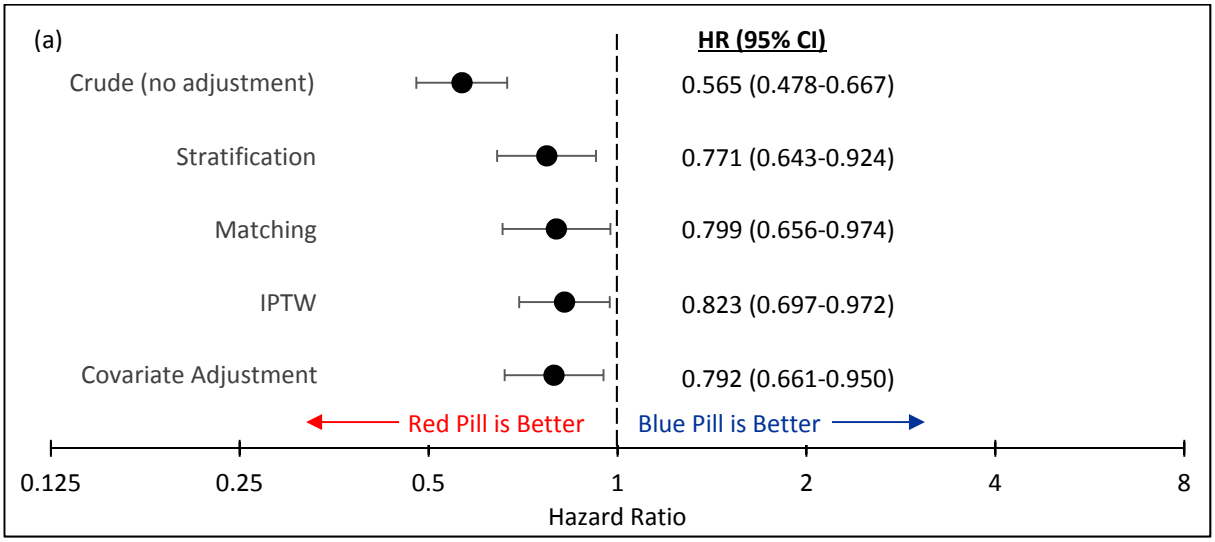


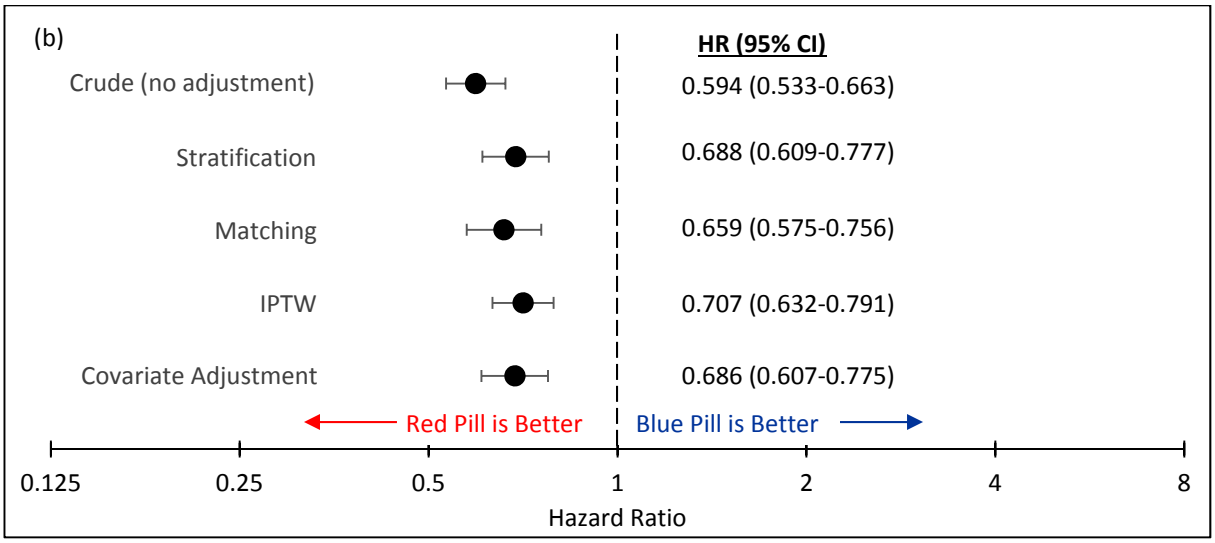
Table 3. Standardized Differences between Blue Pill and Red Pill Patients Using Various Propensity Score Approaches

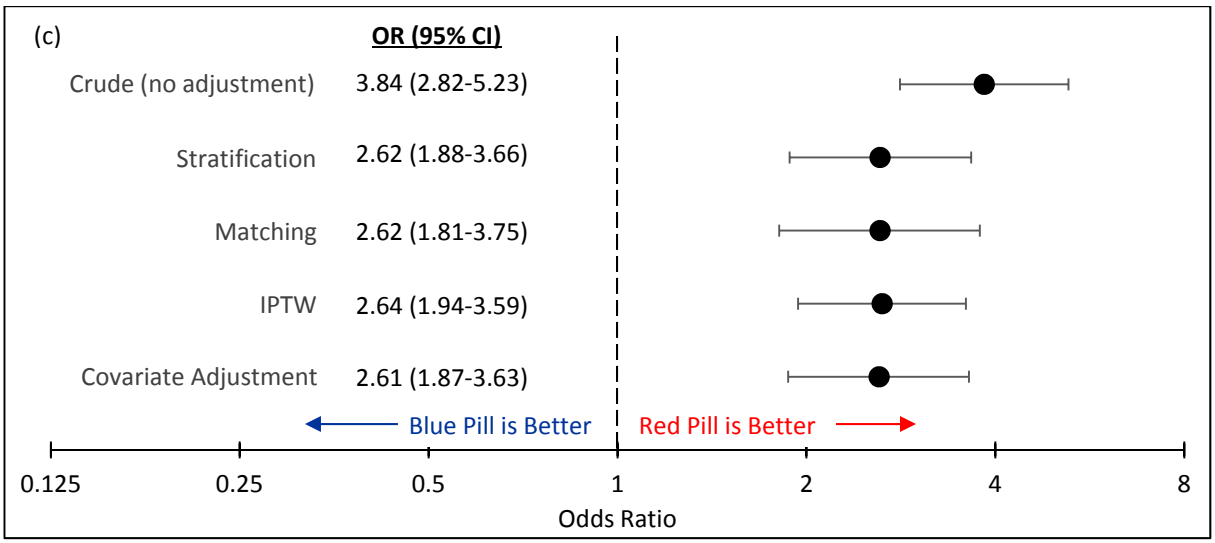
Variable	Stratification Quintile					Matching	IPTW
	1	2	3	4	5		
Age group							
≤ 50	-0.110	-0.162	0.107	--	--	0.005	0.017
51 to 60	0.110	-0.116	-0.027	0.058	-0.109	-0.002	0.006
61 to 70	--	0.155	0.016	0.069	-0.151	0.011	-0.017
> 70	--	--	--	-0.218	0.204	-0.026	0.006
Sex							
Male	0.223	-0.176	-0.014	0.066	0.026	-0.004	0.019
Female	-0.223	0.176	0.014	-0.066	-0.026	0.004	-0.019
Geographic region							
North America	0.057	-0.067	0.116	-0.065	-0.025	0.002	0.009
Europe	-0.042	0.024	-0.130	0.017	0.109	-0.021	0.006
Asia	-0.028	0.075	0.030	0.071	-0.079	0.029	-0.022
Performance status							
0	0.018	0.066	0.015	-0.165	0.146	-0.015	0.017
1	-0.003	-0.052	-0.035	0.149	-0.120	0.032	-0.003
≥ 2	-0.099	-0.114	0.049	0.049	-0.036	-0.048	-0.034
Disease stage							
III	-0.005	-0.182	-0.210	-0.058	0.065	0.004	-0.004
IV	0.005	0.182	0.210	0.058	-0.065	-0.004	0.004
Extranodal sites							
≥ 2	0.025	-0.111	0.129	-0.033	-0.044	-0.029	-0.009
< 2	-0.025	0.111	-0.129	0.033	0.044	0.029	0.009
LDH level							
Elevated	-0.146	0.076	0.028	0.068	0.199	-0.013	0.004
Normal	0.146	-0.076	-0.028	-0.068	-0.199	0.013	-0.004
Bone marrow involvement							
Yes	-0.036	-0.103	-0.079	-0.203	0.052	-0.052	-0.073
No	0.036	0.103	0.079	0.203	-0.052	0.052	0.073
Center type							
Academic	0.054	0.240	0.166	0.169	0.215	0.191	0.175
Community	-0.054	-0.240	-0.166	-0.169	-0.215	-0.191	-0.175

IPTW = inverse probability treatment weighting; LDH = lactate dehydrogenase.

Figure 2. Comparison of Approaches: (a) OS, (b) PFS, (c) ORR







CI = confidence interval; HR = hazard ratio; IPTW = inverse probability treatment weighting; OR = odds ratio; ORR = overall response rate; OS = overall survival; PFS = progression-free survival.