COMPARING WEIGHTING METHODS IN PROPENSITY SCORE ANALYSIS

Michael A. Posner, Ph.D., Department of Mathematical Sciences, Villanova University
Arlene S. Ash, Ph.D., Health Care Research Unit, Boston Medical Center

Michael A. Posner is Assistant Professor, Department of Mathematical Sciences, Villanova University, Villanova, PA 19085. Arlene. S. Ash is Research Professor, Health Care Research Unit, Boston Medical Center, Boston, MA 02118. The authors thank Stefan Kertesz and Boston Health Care for the Homeless Program (BHCHP) for use of the Respite dataset.

Abstract

The propensity score method is frequently used to deal with bias from standard regression in observational studies. The propensity score method involves calculating the conditional probability (propensity) of being in the treated group (of the exposure) given a set of covariates, weighting (or sampling) the data based on these propensity scores, and then analyzing the outcome using the weighted data. I first review methods of allocation of weights for propensity score analysis and then introduce weighting within strata and proportional weighting within strata as alternative weighting methods. These new methods are compared to existing ones using empirical analysis and a data set on whether sending patients to a respite unit prevents readmission or death within ninety days. Simulations are then described and discussed to compare the existing and new methods.

INTRODUCTION

Research often involves determining the effect of an intervention or treatment on an outcome of interest. Randomized controlled trials (RCTs) are the gold standard in scientific research. RCTs involve randomizing subjects to a treatment arm with the goal of eliminating biases by theoretically placing even distributions of subjects by all variables, both measured and unmeasured, in each group. Through this design, they provide strong internal validity. In some situations, however, RCTs are not feasible, ethical, or readily available and observational studies take their place. In some situations, the randomization may fail (eg. patients do not adhere to study protocols). Observational studies provide an alternative to randomized controlled trial. They have strong external validity and allow for generalizability to an entire population, rather than the subset of participants in a trial. Observational studies can be performed in situations when RCTs are unfeasible or unethical as well (parachute article). In addition, RCTs often take years of time and cost millions of dollars to complete, while observational studies are cheaper and faster. So why aren't observational studies more frequently utilized as research tools? Because they are susceptible to bias when models are misspecified and covariates are not evenly distributed across treatment groups (Posner & Ash, in progress).

The propensity score method is frequently used to deal with bias from standard regression in observational studies. The propensity score method involves calculating the conditional probability (propensity) of being in the treated group (of the exposure) given a set of covariates, weighting (or sampling) the data based on these propensity scores, and then analyzing the outcome using the weighted data. [add lots of citations]

NEED LOTS MORE ON PROPENSITY SCORES

This article focuses on the crucial step of determining weights. First, we review methods of allocation of weights for propensity score analysis and then introduce weighting within strata and proportional weighting within strata as alternative weighting methods. We then compare these new methods to existing ones using empirical analysis and a data set on whether sending patients to a respite unit prevents readmission or death within ninety days. Simulations are then described and discussed to compare the existing and new methods. There is also a summary and discussion.

BACKGROUND – METHODS OF SUB-SAMPLING

The propensity score has a number of properties. It is a balancing score, meaning that assignment to treatment is independent of the covariates conditional on the propensity score. Under the assumption of strong ignorability (define this), the outcome is independent of the treatment conditioned on the covariates. Thus, the expected value of the average treatment effect, the difference between the treated and the control data, is the expected value of the average treatment effect conditioned on the propensity score.

Once the propensity scores are calculated, the analyst has a number of options of how to sample or weight the data in order to determine the average treatment effect. The method of selecting an appropriate set of data that is similarly distributed on covariates is a crucial step in the propensity score method. There are four commonly used methods for selecting the sample or weighting the data: random selection within strata, matching, regression adjustment, and weighting based on the inverse of the propensity score. We introduce another method of weighting that provides an alternative to weighting by the inverse propensity score that is less susceptible to extreme weights and has a higher coverage probability of the true value, according to simulations.

RANDOM SELECTION (OR SAMPLING) WITHIN STRATA

Random selection within strata was proposed by Rosenbaum and Rubin (1983) in their paper that introduced propensity scores. In this paper, they presented the propensity score as a way to summarize numerous variables into a scalar balancing score – the propensity of being in the treated group. This score could much more readily be used instead of the vector of variables, including being used to stratify the data in quintiles. Cochran (1968) had calculated that stratification based on five strata on a covariate eliminates 90% of bias in observational studies and Rosenbaum and Rubin followed his logic and argument by suggesting splitting the propensity score into quintiles in order to reduce bias.

As in all the methods, the probability of being in the treated group, conditioned on the covariates, is first calculated. This is typically accomplished with a logistic or multinomial model using all covariates. In random sampling within strata, all observations are ranked on their propensity score, and the data are then divided into quantiles of the propensity score. Within each stratum, equal sample sizes in the treatment and control groups are selected. Thus, if the treatment group is larger, a subset of treated observations in that stratum is randomly chosen so that the sample size equals that of the control group, and vice-versa if the control group is larger. Inferences will therefore be made only in the space where the distributions of the two groups overlap. If the distributions do not overlap in a region of the space, the data should be excluded.

In the context of weighting, this method assigns weights of 1 or 0 to each observation. If a given observation is in the selected sample, it gets a weight of 1, while if it is not, a weight of 0 is assigned to it. A weighted least square regression will result in the same estimates as if reduced sample size ordinary least square regression had been applied.

Random selection within strata has the advantage of simplicity in application, but poses some limitations. First, it can exclude a substantial amount of data if there are strata that have particularly small numbers of observations in one group or the other, which may create power problems. For example, if you have 100 people in the lowest quintile based on propensity score, and 3 in the treated group while 97 are in the control group, this method would select the 3 treated observations as well as a random sample of 3 out of the 97 in the control group, eliminating 94 observations (or 94% of the sample from this quintile). Clearly, this would reduce the power and precision of the analysis. Second, since it is based on random selection, two researchers using this method may identify different analytic samples via randomization and thus obtain different results, violating the scientific principle of replicability.

There is an added benefit that many researchers have employed from this method. The effect size of exposure on outcome within strata can be examined to determine whether there is a differing effect across groups who are differing in their propensity of being in the treated group.

Stratification methods as described here have been used by many researchers (Rosenbaum and Rubin, 1984, Fiebach, et. al., 1990, Czajka, et. al., 1992, Hoffer, Greeley, and Coleman, 1985, Lavori, Keller, and Endicott, 1988, Stone, et. al., 1995, Lieberman, et. al., 1996, Gum, et. al., 2001 to list a few).

MATCHING

There are several propensity score approaches that use matching, three of which are considered here – a greedy algorithm, nearest neighbor matching, and nearest neighbor matching within calipers. These methods call for matching one treated observation for each control observation (or vice-versa, depending on which group has the smaller number of observations). For each treated observation, an algorithm is used to identify a control that has a similar propensity score. Rosenbaum (2002, section 10.3) discusses optimal matching techniques that expands on the 1:1 matching by involving k:1 matching, either for a pre-specified value of k and for varying values of k.

Rosenbaum and Rubin (1985) suggest that the logit of the propensity score is better to use for matching than the propensity score itself. This method linearizes distances from the 0-1 interval. This suggestion incorporates the fact that differences in probabilities of a fixed size are more important when the probabilities are close to 0 or 1. For example, a 0.01 difference between 0.01 and 0.02 represents doubling the likelihood for an individual, while the same difference between 0.50 and 0.51 is only a 2% increase.

The matching method originally proposed was nearest neighbor matching. In this strategy, all possible pairs of treated and control observations are considered and the pairs that produce the minimal distance in their propensity scores is used. Either Euclidean or Mahalanobis distance are typically employed for this. Euclidean distance is the geometric distance between two observations $\left(\sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}\right)$. Mahalanobis distance scales the distance to the variance in each observation based on the covariance matrix. $[(X_1-X_2)^T C^{-1}(X_1-X_2)$, where C is the covariance matrix of covariates X_1 and $X_2]$. Thus, the metric is weighted by the variance in each direction. If, for example, the variance of X_2 is twice the variance of X_1, then an observation needs to be twice as far in order to be equidistant in the Mahalanobis distance. One way to think about this is to imagine a car that has flat terrain east and west of it, and rocky terrain north and south of it. The distance that the car can travel in one hour is different depending on which direction it goes. A one hour trip north will not get you as far as a one hour trip west. In this example, Mahalanobis distance is analogous to the time it takes to get there – you have not traveled as far north, but it took an hour to get there, so it is considered equidistant to a one hour trip west. Note that if the data are standardized, Mahalanobis and Euclidean distance are identical.

The simplest, least efficient of these matching protocols is the "greedy algorithm". This method was implemented by Parsons (2001) and discussed in Rosenbaum (2002). For each observation in the smaller of the two groups, treatment or control, identify the observation from the other group whose propensity score (or logit thereof) is closest. After matching this pair, remove these observations from the pool of observations and move on to the next one, repeating this process until there are no more observations to match. Programming this algorithm is simpler, but can result in matching sub-optimal pairs together which are quite distant from each other. In addition, since the matches are chosen sequentially, the order of the data matters since you exclude each pair once you have matched them. Rearranging the data can result in dramatically different sets of matched pairs. This is not a desirable property.

Lastly, matching within calipers was proposed to protect against a treated and control observation that are not similar to each other in their propensity score being matched solely due to no other observation being a closer match (this may occur even when the greedy algorithm is not used). In particular, extreme observations which are different in covariates from all observations in the other treatment group should be excluded from the analysis. In this method, a limit is set, and if there are no observations in the other group within that range, the observation is dropped from analysis. Rosenbaum and Rubin (1985) suggested using a quarter standard deviation of the logit of the propensity score as the caliper width. Matching within calipers is one of the more frequently used methods for propensity score matching.

Matching has three benefits, according to Rosenbaum and Rubin (1983):

1. Matched treated and control pairs provide a simple representation of the data for researchers,
2. The variance of the estimate of the average treatment effect will be lower in matched samples than in random samples. This is due to more similar distributions of the observed covariates, and
3. Model-based methods are more robust to departures from underlying model assumptions.

REGRESSION ADJUSTMENT USING THE PROPENSITY SCORE

A third method is regression adjustment, also proposed in the initial paper by Rosenbaum and Rubin (1983). In this method the propensity score is calculated, as before, and is simply used as an additional covariate in the outcome model. Roseman (1994) shows that this method reduces bias in a manner similar to those previously discussed. Regression adjustment methods were used by Berk and Newton (1985), Berk, Newton, and Berk (1986) and Muller, et. al. (1986).

It is unclear, however, how this method really fixes the problem of bias from standard regression. The effect of adding a propensity score covariate in the outcome model is essentially to allow the treatment effect to vary with the propensity of being in the treated group. In the following example, let X be a covariate (or covariates), β_T be the constant treatment effect, T be an indicator of treatment (1 if treatment, 0 if control), β_0 and β_1 be the intercept and slope, respectively, p(Z) be the propensity score (which is dependent on the vector Z, which may or may not contain some of X), Y be the outcome, and β_{PS} be the slope for the propensity score term. In addition, D'Agostino (1998) states that this method fails when the discriminant is a non-monotone function of the propensity score, or if the variance between treatment groups is unequal.

A typical regression model will be:

$$Y = \beta_0 + \beta_1 X + \beta_T T + \varepsilon$$

while the model including the propensity score will be:

$$Y = \beta_0 + \beta_1 X + \beta_T T + \beta_{PS}\, p(Z) + \varepsilon$$

In the second model, the effect of β_T will be diluted by the presence of p(Z) in the model. In particular, p(Z) will likely be high when T=1, so the effect of β_T will be much less than in the first model. Thus, if β_T is used as an estimate of the effect of being in the treated group, this effect will be underestimated.

$$\hat{u} = X_1\beta' + X_2\gamma' + \eta$$
$$y = X_1\beta + X_3\gamma + \hat{u}\delta + \varepsilon$$

$$y = X_1\beta + X_3\gamma + \delta(X_1\beta' + X_2\gamma' + \eta) + \varepsilon$$
$$y = X_1(\beta + \delta\beta') + X_3\gamma + X_2\gamma' + \delta\eta + \varepsilon$$
$$E[y] = X_1(\beta + \delta\beta') + X_3\gamma + X_2\gamma'$$
$$Var(y) = Var(\delta\eta + \varepsilon)$$
$$= \delta^2\sigma_\eta{}^2 + \sigma_\varepsilon{}^2 + 2\delta Cov(\eta, \varepsilon)$$

WEIGHTING BY THE INVERSE PROPENSITY SCORE

A fourth method of quasi-randomization was proposed by Imbens (2000) and further discussed by Hirano and Imbens (2001) and is similar to one proposed independently by Robins and Rotnitzky (1995) in the context of marginal structure models for time-dependent treatment. Here, the inverse of the propensity score is used to weight each observation in the treated group, and one minus the inverse of the propensity score (i.e., the propensity of NOT being in the treated group) in the controls. Weighting has the nice property of including all the data (unless weights are set to 0) and does not depend on random sampling, thus providing for replicability.

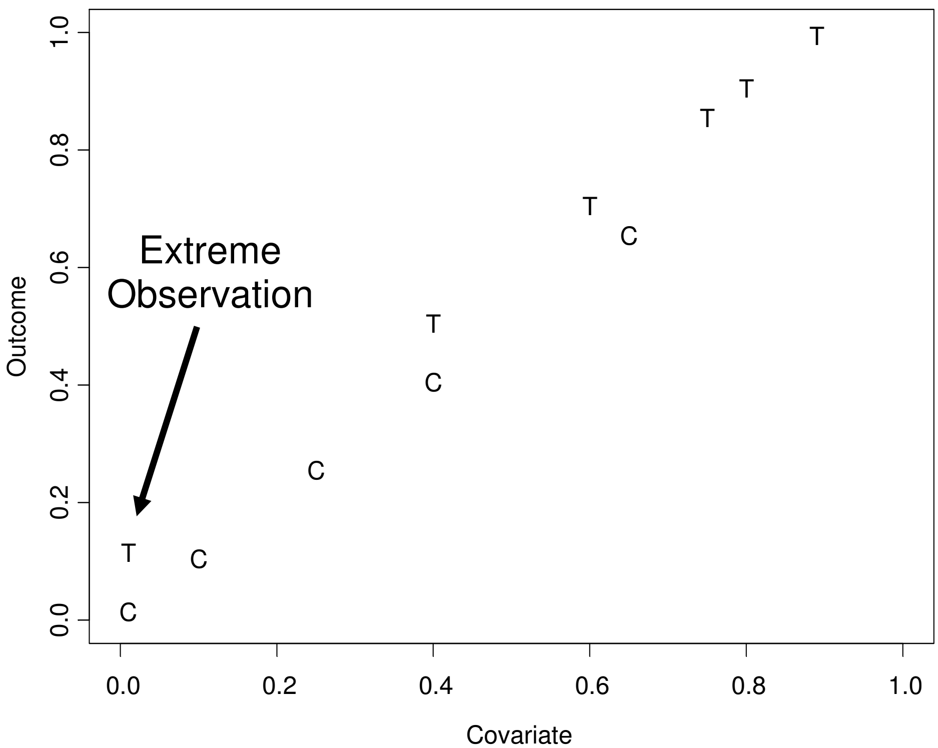
Imbens has shown that weighting based in the inverse of the propensity score produces unbiased estimates by the following: We wish to estimate the average treatment effect, $E[Y_T\text{-}Y_C]$ where Y_T is the outcome for the treated observations and Y_C is the outcome for the control observations. This can be separated to $E[Y_T]$ $-E[Y_C]$. We actually want to examine the effects conditional on their observed covariates, so $E[Y_T|X]$ $-E[Y_C|X]$ is what we wish to estimate. The following equation is a modification of Imbens' equation and shows that weighting by the inverse of the propensity score, (p(x,T)), where T is an indicator of treatment (1=treatment, 0=control), X is the vector of independent variables, and Y is the outcome, produces an unbiased estimate of the true treatment effect. The same result holds for the controls as well.

$$E\left[\frac{Y\times T}{p(x,T)}\right]=E\left[E\left[\frac{Y\times T}{p(x,T)}\middle|X\right]\right]=E\left[E\left[\frac{Y\times T}{p(x,T)}\middle|X,T=1\right]P(T=1|X)\right]=$$

$$E\left[E\left[\frac{Y_t}{p(x,t)}\middle|X\right]p(x,t)\right]=E\big[E[Y_t|X]\big]=E[Y_t]$$

While this method can be shown to have nice mathematical properties, it does not work well in practice. Consider a lone treated observation that happens to have a very low probability of being treated (see figure 1 – the treated observation, "T", in the lower left hand corner of the graph). The value of the inverse of the propensity score will be extremely high, asymptotically infinity. The effect size obtained will be dominated by this single value, and any fluctuations in it will produce wildly varied results, which is an undesirable property.

FIGURE 1: EXAMPLE OF EXTREMELY INFLUENTIAL OBSERVATION IN INVERSE PROPENSITY WEIGHTING SCHEME



WEIGHTING WITHIN STRATA AND PROPORTIONAL WEIGHTING WITHIN STRATA

The methods that I propose for propensity score weighting are weighting within strata and proportional weighting within strata. The latter is only useful in the presence of polychotomous exposure groups. There are five steps to calculate the weights: 1) Calculate a propensity score for each observation, 2) Sort data into quantiles of the propensity score, 3) Calculate the number of treated and control observations in each quantile, then 4) Assign a weight to each observation within each group (treated or control) of each quantile that is the reciprocal of the proportion of observations in that quantile group (treated or control) relative to the total number of observations in that quantile, and 5) Multiplied by the number of groups (to scale appropriately). An example of this is presentedlater. Once this has been accomplished, perform a weighted least squares regression of the outcome of interest using the calculated weights.

Proportional weighting within strata follows the same five steps as weighting within strata and adds a sixth step. The last step is to rescale the weights so that the sum of weights given to each treatment group is equal to the original sample size in that group. As an example, if there are 100 observations in group 1, 100 in group 2, and 400 in group 3, all weights for groups 1 and 2 are decreased by multiplying by 100/200 and all control observations weights are inflated by multiplying by 400/200. The value 200 is obtained from assigning equal weights to each group (600 observations divided by 3 groups), and is the total of the weights assigned within each group by weighting within strata. The advantage of this is that it reflects the actual amount of information present from the treatment and the control groups.

An illustration of these methods is given in the following two sections. These methods share the virtues of all weighting methods in that they do not involve random selection and include all data in analyses (unless, as in the sampling schemes you assign weights of 0 to some observations). Evaluation of these methods is accomplished through simulations and through a data set on whether sending patients to a respite unit prevents readmission or death within ninety days followed by simulations to compare various methods of sample selection or weighting in propensity score analysis.

COMPARING WEIGHTING SCHEMES – EMPIRICAL RESULTS

The following examples demonstrate results obtained by the different methods under varying assumptions. Subjects are members of one of three covariate groups, called low, moderate, and high, and either receive a treatment or serve as a control. The numbers were selected so that those in the high covariate group have a high chance of being in the treatment group, while those in the low covariate group have a small chance of being in the treatment group. For example, in table 1a, the probability of being in the treated group is 25% (30/120) for the low group, 67% (200/300) for the moderate group, and 94% (170/180) in the high group. Propensity score analysis is performed in four ways – random selection within strata, weighting within strata, proportional weighting within strata, and inverse propensity weighting. The percents effectiveness are different for each cell of the table, and given in the next part of table 1, listed as “true trt” and “true control”. The results are presented in tables 1a – 1c to illustrate the differences between the weighting schemes. The different tables present different treatment effect sizes and covariate distributions.

For example, in table 1a, there is a 10% treatment effect in the low group, a 5% effect in the moderate group, and a 0% treatment effect in the high group. Typically, a single treatment effect estimate is made (whether or not this is an appropriate decision is left to the analysts and evaluators of each study and is intentionally omitted here). If this were the case, the true treatment effect, should be 4.5%, which is an average of the three effect sizes, weighted to the sample size in the group. The naïve estimate, which does not separate by the covariate groups, would estimate a –0.5% effect size. This is an example of Simpson’s paradox, where summarizing over a variable masks the actual effect.

The randomization within strata produces an estimate of 5.7%, which has a 1.2% bias from the true effect. In addition, note that if there were numerous covariates, the randomization process would produce different results in different instances of the randomization. The weighting within strata, proportional weighting within strata, and inverse probability weighting all estimate the true effect (4.5%). In this example, with only one covariate, weighting within strata and inverse probability weighting wind up with the same results.

The proportional weighting within strata rescales the results by matching the sample size for treatment and control groups. This has the important feature of not artificially deflating the variances. It is clear that if you have N observations, with equal variances (or unknown variances, as in most situations), the minimal standard error will be obtained by allocating N/2 observations to each treatment group. From this, we note that the weighting within strata and inverse probability weighting may have artificially deflated variance estimates. Hirano and Imbens (2001) discuss this problem in their paper.

In addition, note that the inverse probability weighting produces weights that vary within strata, while these two methods do not. Recall the problem of one extreme observation having a large influence on the results, as shown in figure 2. These two methods protect against this more than the inverse probability weighting does, since the effect of one extreme observation would be no larger than any other observation in the same quantile group based on the propensity score.

TABLE 1a. DIFFERENCE IN ESTIMATED EFFECTS FROM NEWLY PROPOSED WEIGHTING SCHEMES

Raw Data	Low	Mod	High			
Treatment	30	200	170	400		56.5%
Control	90	100	10	200		57.0%
	120	300	180	600	**Crude Diff**	**-0.5%**

True Trt	70%	60%	50%			
True Control	60%	55%	50%		**True Diff**	**4.5%**
Diff	10%	5%	0%			
Random w/in Strata	30	100	10	140		61.4%
	30	100	10	140		55.7%
	60	200	20	280	**Difference**	**5.7%**
Weight w/in Strata	60	150	90	300		59.0%
	60	150	90	300		54.5%
	120	300	180	600	**Difference**	**4.5%**
Proportional w/in Strata	80	200	120	400		59.0%
	40	100	60	200		54.5%
	120	300	180	600	**Difference**	**4.5%**
Inverse Probability	60	150	90	300		59.0%
	60	150	90	300		54.5%
	120	300	180	600	**Difference**	**4.5%**

TABLE 1b. DIFFERENCE IN ESTIMATED EFFECTS FROM NEWLY PROPOSED WEIGHTING SCHEMES

Raw Data	Low	Mod	High			
Treatment	30	200	170	400		63.3%
Control	90	100	10	200		57.0%
	120	300	180	600	**Crude Diff**	**6.3%**
True Trt	70%	65%	60%			
True Control	60%	55%	50%		**True Diff**	**10.0%**
Diff	10%	10%	10%			
Random w/in Strata	30	100	10	140		65.7%
	30	100	10	140		55.7%
	60	200	20	280	**Difference**	**10.0%**
Weight w/in Strata	60	150	90	300		64.5%
	60	150	90	300		54.5%
	120	300	180	600	**Difference**	**10.0%**
Proportional w/in Strata	80	200	120	400		64.5%
	40	100	60	200		54.5%
	120	300	180	600	**Difference**	**10.0%**
Inverse Probability	60	150	90	300		64.5%
	60	150	90	300		54.5%
	120	300	180	600	**Difference**	**10.0%**

In table 1b, the treatment effect is changed to 10% for each group. This still results in a biased crude estimate, but all the methods of propensity score adjustment lead to same conclusion – an unbiased estimate of a 10% treatment effect.

TABLE 1c. DIFFERENCE IN ESTIMATED EFFECTS FROM NEWLY PROPOSED WEIGHTING SCHEMES

Raw Data	Low	Mod	High			
Treatment	48	150	108	306		58.0%
Control	72	150	72	294		55.0%
	120	300	180	600	**Crude Diff**	**3.0%**
True Trt	70%	60%	50%			
True Control	60%	55%	50%		**True Diff**	**4.5%**
Diff	10%	5%	0%			
Random w/in Strata	48	150	72	270		59.1%
	48	150	72	270		54.6%
	96	300	144	540	**Difference**	**4.6%**
Weight w/in Strata	60	150	90	300		59.0%
	60	150	90	300		54.5%
	120	300	180	600	**Difference**	**4.5%**
Proportional w/in Strata	61	153	92	306		59.0%
	59	147	88	294		54.5%
	120	300	180	600	**Difference**	**4.5%**
Inverse Probability	60	150	90	300		59.0%
	60	150	90	300		54.5%
	120	300	180	600	**Difference**	**4.5%**

Table 1c illustrates that with a smaller difference in the distribution of the covariate (low, mod, high), the bias is still present, but is reduced. Here, the percent difference between covariates is reduced from 25% low, 67% moderate, and 94% high to 40% low, 50%

moderate, and 60% high. From this, we see that greater covariate imbalance leads to more biased results for the crude analysis.

SIMULATION METHODS

Simulations were performed to compare the different methods of sample selection or weighting. The exposure (E) is a dichotomous variable, the outcome (O) is a continuous variable, and the confounder (C) is also continuous. For example, consider an observational unit being a neighborhood around a hospital. We could consider the exposure to be whether a hospital provides mammography services (yes/no), the outcome to be rate of early detection of breast cancer in the hospital's potential patients (a continuous measure), and the covariate to be percentile of income by zip code (also continuous).

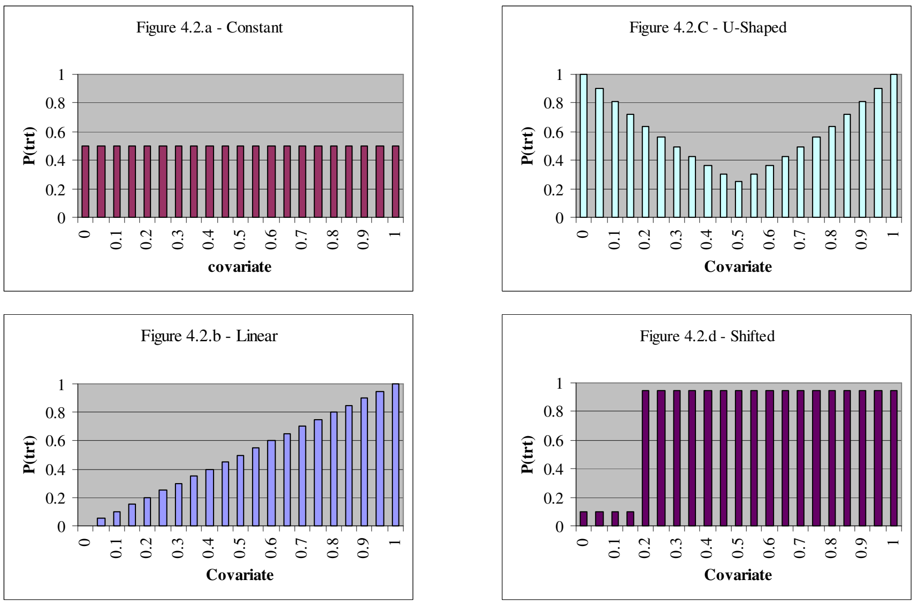
The data were generated according to the following:

1) The covariate (C) was generated as a continuous uniform (0,1) variable.

2) Exposure (E) was generated using four different methods. A graphical representation of these distributions can be found in figure 3. The first is an even distribution, and assigns the probability of being in the exposed group to be 50%, regardless of the value of the covariate. The second is a linear distribution that assumes that the probability of being exposed is equal to the covariate. For example, if C=0.7, then the individual was given a 70% probability of being exposed. Thus, those with high values of the covariate are very likely to be exposed, while those with low values are very unlikely to be unexposed. While this forces a differing covariate distribution, the symmetric nature of it means that the residuals for users and non-users will both have a mean of 0 (this will be discussed further later). The third method for generating exposure is a U-shaped curve where the likelihood of being exposed is high on the low and high range of the "U", while an observation is given a low probability of being exposed in the middle range of the "U". Contextually, this may be seen in the situation where those of high income can afford assistance or are more educated and low income people are offered incentives to get mammography, while the middle income group is left out. This type of relationship was seen in an investigation we conducted on mammography rates in Haitians in Boston (David, et. al., 2005). The fourth method is a shifted distribution, that

considers the situation where there are very few unexposed who exist on the low end of the covariate, while those in the exposed group are plenty and congregate more heavily on the high and middle values of income (Underrepresented). This situation has arisen in research on gender bias in compensation (Ash, 1986).

FIGURE 2. DIFFERING COVARIATE DISTRIBUTIONS USED IN THE SIMULATION STUDY



3) The outcome has one of two relationships with the covariate, either linear or non-linear (fourth root). In the former case, the model is correctly specified, while in the latter case, the model is misspecified. (Note that this should not be confused with the relationship between the exposure and the outcome – see #4 below)

4) The desired effect size (β_{ue}) is then added to the outcome for those in the exposed group (i.e. if the effect size states that those in the treated group are, on average, 1 unit higher than the control units, then treatment groups had 1 unit added to their value).

5) Seven methods of sample selection/weighting are then applied to the data and bias (difference in estimated relationship between exposure and outcome and the actual relationship) is calculated. The first two methods, crude analysis (no covariate adjustment) and standard regression, do not employ propensity score methods. The latter five, random selection within strata, propensity score regression, weighting by the inverse

propensity score, weighting within strata, and nearest neighbor matching via a greedy algorithm, use propensity score methods to address issues of bias from standard regression. Weighting within strata is the new method presented in this article. Proportional weighting within strata, the other method introduced in the same section, is only helpful when doing polychotomous propensity score analysis, thus is it not included here. The relationship between the covariate and the outcome is modeled linearly, to incorporate model misspecification when the actual distribution is non-linear.

7) Each scenario, with different parameter values, is replicated 500 times, so that a sampling distribution is obtained. Note that for propensity score using random selection within strata (PSRSWS) and propensity score using a greedy matching algorithm (PSGrd), there were data anomalies where the randomly selected sample size was larger than original, so that observation was excluded. Thus, some scenarios have fewer than 500 replicates.

The following summary measures were calculated:

Distshape = Linear (L) or Non-Linear (N) model between the covariate and outcome. The linear model is correctly specified, while the non-linear (fourth root) model is misspecified, since a linear relationship is used to model it.

Datadist = One of the four distributions for the covariate (C=Constant, L=linear, U=U-shaped, S=shifted), as described above

Nobser = # of observations, either 100 or 1,000

Seu = Standard deviation from the true model (taking values 0.01 or 5)

β_{ue} = Effect size for user (taking values 0, 0.1, or 2)

Type = Crude, Standard Regression (SR), Random Selection Within Strata (PSRSWS), Propensity Score Regression (PSReg), Weighting by the Inverse Propensity Score (PSWIP), the new method of Weighting within Strata (PSWWS), and Nearest Neighbor Matching using a Greedy Algorithm (PSGrd)

Minobs = smallest number of observations selected (relevant only for PSRSWS and PSGrd methods, since sample size reduction was done)

Maxobs = largest number of observations selected (relevant for PSRSWS and PSGrd only)

Meanbias = the average bias (estimated effect minus true effect) of the 500 replicates

RelBias = relative bias = (meanbias – β_{ue}) / β_{ue}

StdBias = standard deviation of the bias of the 500 replicates

MSE = mean square error = $(\text{bias})^2$ + variance

p5bias = 5th percentile of bias

p95 bias = 95th percentile of bias

piwidth = prediction interval width = p95bias – p5 bias

covprob = coverage probability = the percent of times the true value is within the 95% confidence interval of the effect size estimate

Parameters are arbitrarily chosen in order to compare biases. A sample of size larger than 1000 was considered, but rejected, due to processing time constraints (due to the greedy matching algorithm, which compares all possible pairs of data).

SIMULATIONS RESULTS

The following table (2) presents a subset of the results for several scenarios which have been numerically labeled. Figure 3 presents a comparison of the coverage probabilities graphically. This output is for 1000 observations, se=0.01, and β_{ue}=0.1. The complete set of results can be found in appendix B. Let us focus on the coverage probability (the chance that the 95% confidence interval contains the true estimate). In theory, this value should be 95% for unbiased estimates. For example, scenario 49, presents a correctly specified (distshape=L) model with constant probability of being in the treated group (datadist=C). The coverage probabilities are all very close to 95%. Scenario 52 represents a correctly specified model (distshape=L) but differing distribution of the covariate between treatment groups (datadist=S). Here, notice that the crude estimate is quite biased (coverage probability of 0% and mean bias of 0.4). The other methods seem to capture the correct coverage value, though PSWWS is lower than others. In scenario 93, the model is incorrectly specified (distshape=N), but the data distribution is constant (datadist=C). Under this scenario, the coverage probabilities are all close to 95%, except for nearest neighbor matching, which falls to 55%. In scenario 96, there are both model misspecification (distshape=N) and uneven covariate distribution between treatment groups (datadist=S), the two conditions requiring propensity score analysis. Here, note that the crude coverage probability is 0%, as seen before, but the standard regression coverage probability is also 0%, meaning that standard regression techniques do not appropriately deal with the bias in these situations. PSRWS, PSReg, and PSWWS result in estimates of effect that are the least biased among all the methods. Figure 4 presents the mean bias from each result, analogous to the coverage probabilities shown in figure 3. From this, we see that the primary reason for the coverage probabilities to be low is a large mean bias. Results were similar in comparing other scenarios.

TABLE 2. COMPARING METHODS OF SAMPLE WEIGHTING/SELECTION FOR PROPENSITY SCORE METHODS VIA SIMULATION

Comment [MAP1]: UPDATE THIS TABLE WITH CORRECT RESULTS

Dist Shape	Data Dist	Min Obs	Max Obs	Type of Analysis	MSE	Mean Bias	St.Dev. Bias	Coverage Probability
Linear	Constant	1000	1000	0-Crude	0.000	-0.002	0.019	93%
Linear	Constant	1000	1000	1-SR	0.000	0.000	0.001	96%
Linear	Constant	894	990	2-PSRSWS	0.000	0.000	0.001	96%
Linear	Constant	1000	1000	3-PSReg	0.000	0.000	0.001	96%
Linear	Constant	1000	1000	4-PSWIP	0.000	0.000	0.001	96%
Linear	Constant	1000	1000	5-PSWWS	0.000	0.000	0.001	96%
Linear	Constant	906	1000	6-PSGrd	0.000	0.000	0.001	96%
Linear	Shifted	1000	1000	0-Crude	0.158	0.397	0.017	0%
Linear	Shifted	1000	1000	1-SR	0.000	0.000	0.001	95%
Linear	Shifted	82	192	2-PSRSWS	0.000	0.000	0.002	94%
Linear	Shifted	1000	1000	3-PSReg	0.000	0.000	0.001	95%
Linear	Shifted	1000	1000	4-PSWIP	0.000	0.000	0.001	88%
Linear	Shifted	1000	1000	5-PSWWS	0.000	0.000	0.001	62%
Linear	Shifted	360	522	6-PSGrd	0.000	0.000	0.001	95%
Non-Linear	Constant	1000	1000	0-Crude	0.000	0.000	0.011	94%
Non-Linear	Constant	1000	1000	1-SR	0.000	0.000	0.004	94%
Non-Linear	Constant	874	986	2-PSRSWS	0.000	0.000	0.003	99%
Non-Linear	Constant	1000	1000	3-PSReg	0.000	0.000	0.003	94%
Non-Linear	Constant	1000	1000	4-PSWIP	0.000	0.000	0.004	94%
Non-Linear	Constant	1000	1000	5-PSWWS	0.000	0.000	0.003	99%
Non-Linear	Constant	906	1000	6-PSGrd	0.000	0.001	0.008	55%
Non-Linear	Shifted	1000	1000	0-Crude	0.069	0.263	0.012	0%
Non-Linear	Shifted	1000	1000	1-SR	0.006	0.076	0.006	0%
Non-Linear	Shifted	88	198	2-PSRSWS	0.000	0.005	0.010	92%
Non-Linear	Shifted	1000	1000	3-PSReg	0.000	0.000	0.005	86%
Non-Linear	Shifted	1000	1000	4-PSWIP	0.006	0.074	0.010	0%
Non-Linear	Shifted	1000	1000	5-PSWWS	0.000	0.003	0.005	72%
Non-Linear	Shifted	364	530	6-PSGrd	0.004	0.059	0.006	0%

FIGURE 3: COVERAGE PROBABILITIES FROM SELECT SIMULATION RESULTS (AS PRESENTED IN TABLE 2)

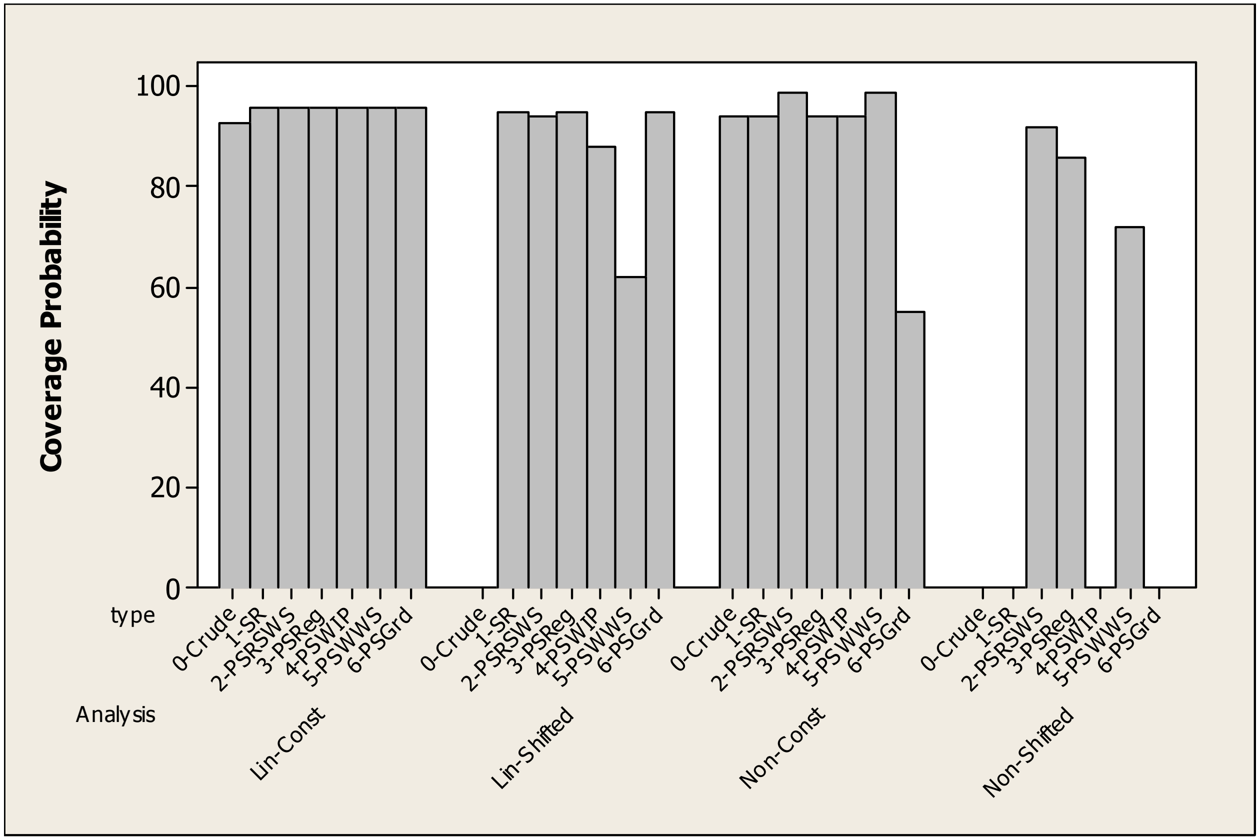
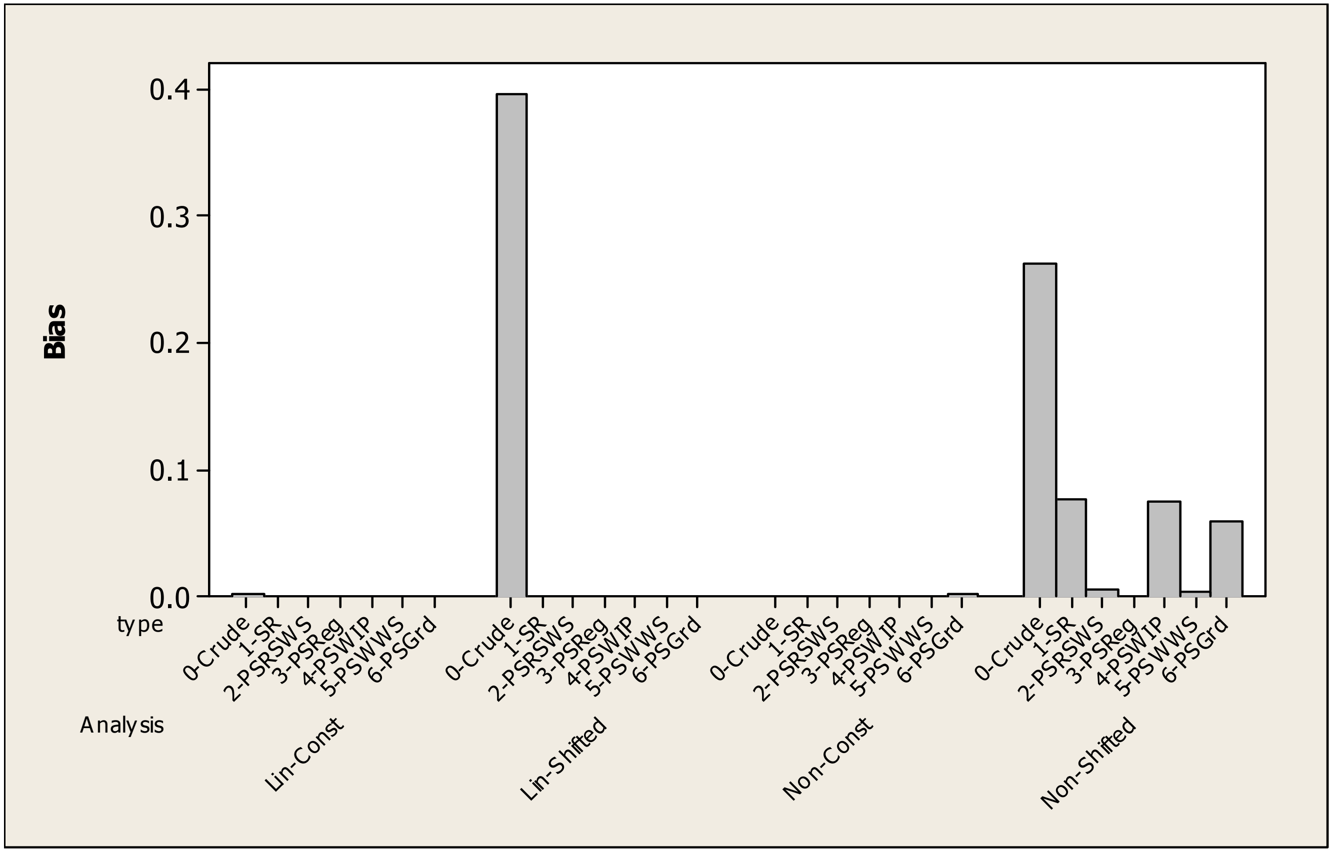


FIGURE 4: BIAS FROM SELECT SIMULATION RESULTS (AS PRESENTED IN TABLE 2)



Bias
0.4
0.3
0.2
0.1
0.0
type
0-Crude
1-SR
2-PSRSWS
3-PSReg
4-PSWIP
5-PSWWS
6-PSGrd
Analysis
Lin-Const
Lin-Shifted
Non-Const
Non-Shifted

DIFFERENCE IN WEIGHTING SCHEMES – RESPITE DATA SET

RESPITE DATA SET

I consider an example of an observational study where propensity scores can be effectively used to address issues of bias from standard regression. The respite unit is a place where homeless patients can be discharged from the hospital and placed when going back on the streets puts them at higher risk of readmission. This is viewed as a cost-saving measure to the hospital. These data are presented in Kertesz, et. al. (2005).

We used administrative data to identify a retrospective cohort of homeless persons 18 or older who survived a non-maternity, medical-surgical hospital admission to Boston Medical Center between July 1, 1998 and June 30, 2001. We identified as homeless patients those who used the Boston Health Care for the Homeless Program (BHCHP) for at least one outpatient clinical encounter within 365 days of an inpatient admission to Boston Medical Center. Administrative data provided by Boston Medical Center identified 1029 candidate patients with BHCHP as their possible primary care site. Review of BHCHP databases confirmed 858 subjects with record of a BHCHP outpatient visit within 365 days of the index admission. We then obtained from Boston Medical Center's Medical Information System (MIS) *all* hospital and hospital-based ambulatory encounters from 1/1/1998 (6 months prior to 7/1/1998) to 6/1/2002 (11 months after 6/30/2001). Of the 858 subjects, 14 were only hospitalized for childbirth, 35 did not survive their only hospitalization, and 3 could not be matched (likely due to changes in MIS data system, or to miscoding), leaving 806 to be assessed for discharge disposition. We assessed each subject for readmission occurring within 90 days of hospital discharge. Death was ascertained from BHCHP's internal Homeless Death Database and the Massachusetts Registry of Vital Records and Statistics (1998-2001). We captured ICD-9 diagnoses from all Boston Medical Center encounters occurring during the index admission and the 6 preceding months, including those at the BHCHP hospital-based clinic, the emergency department, other outpatient services (e.g. specialty clinics) and inpatient admissions. We combined the MIS-derived data and information from BHCHP's Respite program to assign each subject to one of four mutually exclusive discharge dispositions. The respite group was defined as all patients who were admitted to respite within one day of hospital discharge. Non-respite homeless patients identified in the MIS data set as discharged to their own care were called "home."   Non-respite patients with discharge status indicating supervised recuperative care, e.g. skilled nursing facilities, chronic care hospitals, or home health care, were called "other." Those who left against medical advice (AMA) were called "AMA".

The study's key endpoint was readmission or death occurring within 90 days from hospital discharge. This endpoint, used previously, properly treats death as an adverse outcome. We allowed a one-day "window" to detect admissions to respite because 12 patients were referred to respite one day *after* hospital discharge, typically by homeless-experienced clinicians acting to correct what they may have considered inappropriate discharges to shelters or streets. Subjects readmitted to Boston Medical Center on the day of or day after discharge (n=22) did not have this opportunity for post-discharge referral to respite. Because inclusion of 22 early-readmitted subjects (only 2 of whom went to respite) could bias results in favor of respite, we conducted our main analysis excluding this group (reducing the sample to n=784), but confirmed in sensitivity analysis that results were more favorable to respite when the 22 were included. 41 patients who left the hospital against medical advice were also excluded from the analysis (n=743).

The reason why someone is sent to the respite unit is associated with patient characteristics including history of substance abuse and comorbidities, age, and race. Thus, analyses were done using propensity scores in order to protect against any model misspecification that may be present. Finally our clinical experience at the source hospital was potentially reassuring because respite referrals generally reflected concern that discharge to streets/shelters would result in early readmission (e.g. respite patients were possibly a "bad prognosis" subgroup). We estimated that the respite group could be informatively compared to patients discharged to other settings, acknowledging at worst, a potential bias against finding the hypothesized reduction in early hospital readmission.

Case-mix adjustment variables, drawn from the literature on readmission prediction, included age, sex, race/ethnicity, length of the index hospital admission, the presence of drug and alcohol abuse diagnostic codes during the admission or the preceding 6 months, and illness burden. Illness burden was measured (and adjusted for) using the Diagnostic Cost Groups/Hierarchical Condition Categories (DCG/HCC) risk score, calculated from all diagnoses during the index admission and the prior 6 months of inpatient and outpatient care at Boston Medical Center, including onsite primary care services from BHCHP. The DCG/HCC method generates a numerical estimate for expected health service utilization, and has been applied to prediction of mortality, veterans' service utilization, and Medicare costs. We implemented DCG/HCC scoring through DxCG™ 6.1 for Windows software, applying a model calibrated to Massachusetts Medicaid patients for the years 2000-2001.

These analyses were performed using respite group as both as a dichotomous (respite vs. home, excluding other) and polychotomous (respite vs. home vs. other) treatment variable.

COMPARING WEIGHTING METHODS IN RESPITE DATA SET DICHOTOMOUS OUTCOME

In weighting within strata, the weights are calculated based on the distribution of treated and control observations within each stratum. Like other propensity score methods, the data are split into strata based on propensity scores. Then a weight is assigned using the distribution of observations within the stratum. In weighting within strata, the total weighted sample size for that stratum is split between the treated and the control groups.

In the respite data set, the probability that the person was sent to the respite unit was calculated from a logistic regression model based on all covariates – age (young, middle, old), race (White, Black, Hispanic/Other), whether they had a history of alcohol abuse, whether they had a history of drug abuse, and a measure of their comorbidity burden (DCG score). This propensity score was then used to calculate quintiles. The following table presents the distribution of data from the respite data set:

TABLE 3. QUINTILES OF PROPENSITY SCORES FOR RESPITE DATA SET

Quintile	Respite	Home	TOTAL
1	8	104	112
2	25	90	115
3	27	85	112
4	24	69	93
5	52	85	137
TOTAL	136	433	569

Note that the uneven size of quintiles is due to categorical variables assigning similar probabilities to numerous people.

We can see that those who were predicted not to be sent to respite (quintile 1, which is the lowest propensity score) are least likely to actually have been sent to the respite unit. [8/112 = 7% in quintile 1 vs. 22% in quintile 2 vs. 24% in quintile 3 vs. 26% in quintile 4 vs. 38% in quintile 5].

In the first quintile, there are 8 in the respite group and 104 in the home group, for a total of 112 people. Thus, we would assign weights to each observation so that there is a weighted total of 56 per group, evenly dividing the 112 observations. This means that the weights for each respite observation is 56 / 8 = 7 and the weight for each home observation is 56 / 104 = 0.54. Table 4 displays the weights for all strata.

Proportional weighting within strata follows a similar idea to weighting within strata except that rather than splitting the weights between the groups, they are assigned proportional to the overall sample size in the groups. For example, using the data again from table 3, rather than assigning a weighted sample size of 56 to each group in the first strata, the respite group would get 136 / 569 = 23.9% of the weights and the home group would get 433 / 569 = 76.1% of the weights. Thus, weights are chosen to produce 112 * .239 = 26.8 weighted observations in the respite, and 85.2 weighted observations in the home group in the first strata. Note that these methods converge to the same result when you have equal total sample size in the groups.

Table 4 presents the sample sizes, weights within strata, proportional weights within strata, and inverse propensity weights for the respite data:

TABLE 4. WEIGHTS FOR RESPITE DATA USING THREE METHODS – WEIGHTING WITHIN STRATA, PROPORTIONAL WEIGHTING WITHIN STRATA, AND INVERSE PROPENSITY SCORE WEIGHTING

		Weights per Observation			Total Weights		
Quintile	**Method**	**Respite**	**Home**	**TOT**	**Respite**	**Home**	**TOT**
1	n	8	104	112	8	104	112
	Weighting w/in Strata	7.0	0.5	1.0	56	56	112
	Prop'l Wtg w/in Strata	3.4	0.8	1.0	27	85	112
	Inv. Propensity Wtg*	4.6	0.6	0.9	37	58	95
2	n	25	90	115	25	90	115
	Weighting w/in Strata	2.3	0.6	1.0	58	58	115
	Prop'l Wtg w/in Strata	1.1	1.0	1.0	28	87	115
	Inv. Propensity Wtg*	3.0	0.6	1.1	75	55	130
3	n	27	85	112	27	85	112
	Weighting w/in Strata	2.1	0.7	1.0	56	56	112
	Prop'l Wtg w/in Strata	1.0	1.0	1.0	27	85	112
	Inv. Propensity Wtg*	2.2	0.7	1.0	58	56	115
4	n	24	69	93	24	69	93
	Weighting w/in Strata	1.9	0.7	1.0	47	46	93
	Prop'l Wtg w/in Strata	0.9	1.0	1.0	22	71	93
	Inv. Propensity Wtg*	1.8	0.7	1.0	44	48	92
5	n	52	85	137	52	85	137
	Weighting w/in Strata	1.3	0.8	1.0	69	69	137
	Prop'l Wtg w/in Strata	0.6	1.2	1.0	33	105	137
	Inv. Propensity Wtg*	1.3	0.8	1.0	68	70	137
TOTAL	n	136	433	569	136	433	569
	Weighting w/in Strata	2.1	0.7	1.0	284	286	569
	Prop'l Wtg w/in Strata	1.0	1.0	1.0	136	433	569
	Inv. Propensity Wtg*	2.1	0.7	1.0	282	286	569

* Value for inverse propensity weighting are means, since each

observation can have a different weight. For example, in the first quintile, the weights range from 3.8 to 6.6.

Table 5 displays the distribution of covariates before and after propensity score methods have been imposed, as well as the odds ratios of early readmission or death for those in the respite group relative to those in the home group. Age, race, and drug abuse (DA) all appear to be associated with assignment to respite ($p<0.05$). However, all methods of propensity score matching and weighting (random selection within quintile, weighting within strata, proportional weighting within strata, and inverse propensity weighting) result in analytic samples that no longer have these associations present.

The naïve estimate of the adjusted odds ratio is 0.72, or the odds of early readmission or death is 28% lower for someone in the respite group relative to someone in the home group, controlling for other factors in the model (all covariates). The increase of the odds ratio in the random selection within strata method is potentially due to the exclusion of data. The weighting methods all converge on a similar result of 0.65-0.66. Note that the variance in the weighting within strata and inverse probability weighting is smaller than that of the proportional weighting within strata, as described above.

As a test, we produced the unweighted results within each quintile of propensity score. The adjusted odds ratios of early readmission or death for those in the respite group relative to those in the home group separately for each quintile are 0.54*, 0.35, 0.75*, 0.69, and 0.79, respectively. Note that the ones with an asterisk (*) produced unstable results due to small sample sizes in sub-groups. The crude odds ratios by quintiles are 0.64, 0.35, 0.81, 0.72, 0.84. From this, we see that the 0.87 odds ratio from the random selection within strata propensity score analysis appears to be over-estimated.

TABLE 5. COMPARISON OF DATA DISTRIBUTION AND RESULTS FROM NAÏVE ANALYSIS AND FOUR METHODS OF PROPENSITY SCORE ANALYSIS – DICHOTOMOUS OUTCOME

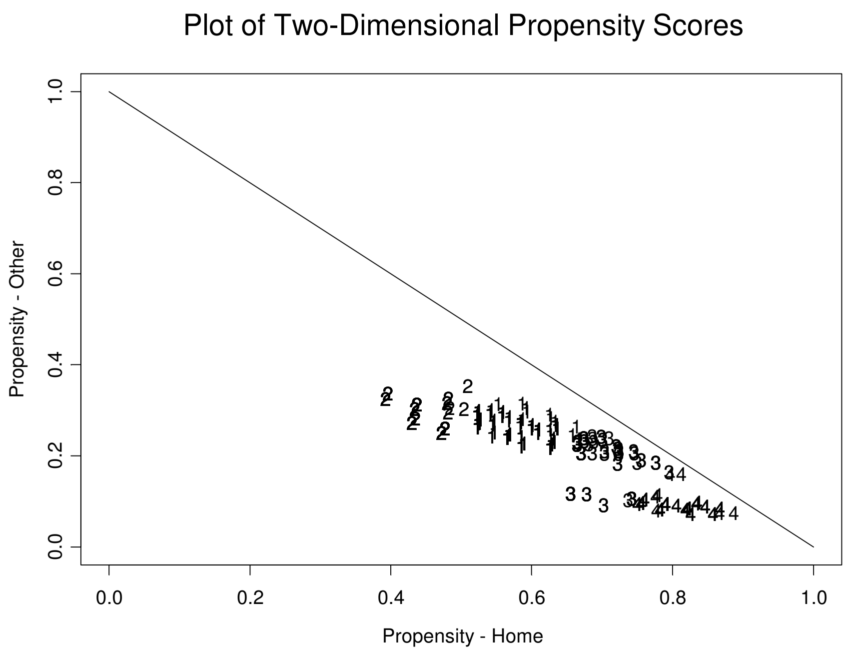
	Pre-		Random Selection Within Strata		Weighting Within Strata		Proportional Weighting Within Strata		Inverse Propensity Weighting	
	Resp	Home	Resp	Home	Resp	Home	Resp	Home	Resp	Home
n	136	433	136	136	284.5	284.5	136	433	281.8	287.2
Age										
Young	19%	31%	19%	21%	26%	28%	26%	28%	26%	28%
Mid	55%	51%	55%	54%	52%	52%	52%	52%	54%	52%
Old	26%	19%	26%	24%	21%	20%	21%	20%	20%	20%
p-value	0.02		0.89		0.89		0.92		0.87	
Race										
White	56%	39%	56%	54%	43%	43%	43%	43%	43%	43%
Black	35%	44%	35%	32%	42%	42%	42%	42%	43%	42%
HispOth	10%	17%	10%	13%	15%	15%	15%	15%	14%	15%
p-value	0.001		0.63		0.99		0.996		0.91	
AA	34%	31%	34%	38%	32%	31%	32%	31%	31%	31%
p-value	0.50		0.53		0.84		0.87		0.87	
DA	8%	19%	8%	7%	17%	16%	17%	16%	15%	16%
p-value	0.002		0.82		0.90		0.92		0.58	
DCG										
Low	10%	17%	10%	7%	11%	16%	11%	16%	13%	16%
Mid	69%	65%	69%	74%	71%	66%	71%	66%	69%	66%
High	21%	18%	21%	19%	18%	18%	18%	18%	18%	18%
p-value	0.14		0.63		0.22		0.36		0.62	
Odds Ratio	0.72 (0.43, 1.20)		0.87 (0.47, 1.62)		0.65 (0.42, 0.997)		0.66 (0.39, 1.10)		0.65 (0.42, 0.998)	

COMPARING WEIGHTING METHODS IN RESPITE DATA SET POLYCHOTOMOUS OUTCOME

The respite data set was then examined using polychotomous treatment groups – including the “other” group, so that we are comparing respite, home, and other. The steps are similar to a dichotomous analysis. First, a multinomial regression was done (instead of a logistic regression) to calculate the probability of being in each group. Next, a cluster analysis was done to assign each observation to a cluster, analogous to splitting the data into quantiles. Note that five clusters produced a group where there were no observations in the respite group, so four clusters were used. This decision will be left for future examination. Weighting the data was done based on the sample size in each treatment group within that stratum.

Figure 5 presents the results of the clustering, where the numbers on the graph (1-4) represent data points that are allocated to that numbered cluster. The value on the plot represents the propensity of being in the home group (X axis) and other group (Y axis). Thus, the propensity of being in the respite group is 1 – x – y. The solid line represents the sum of X and Y equal to one so the closer to the line the less likely they were to be sent to the respite unit. So, for example, we see that cluster 4 is unlikely to go to the respite unit, while cluster 1 and 2 appear to be most likely. Also, note that the odds ratios produced here are the odds or being readmitted or dying within 90 days for home or other group relative to respite group, the reverse of what we saw in the dichotomous analysis.

FIGURE 5: CLUSTERS BASED ON THE PROPENSITY SCORE



Tables 6 and 7 present summaries of the variables used in the analysis. The original data were broken down by cluster and treatment group. In addition, p-values for the chi-squared test demonstrate that the independent variables are associated with the treatment group, yet this association is no longer present in the matched data set (table 7). Notice that the confidence interval for the odds ratio in the proportional weight group is wider than the cluster weight or weighting by the inverse of the propensity score group. These results are similar to what was found in the dichotomous situation.

TABLE 6. COMPARISON OF WEIGHTS FOR FOUR METHODS OF PROPENSITY SCORE ANALYSIS

		Weights per Observation				Total Weights			
	Method	**Respite**	**Home**	**Other**	**TOT**	**Respite**	**Home**	**Other**	**TOT**
1	n	56	193	72	321	56	193	72	321
	Weighting w/in Strata	1.9	0.6	1.5	1.0	107	106	107	321
	Prop'l Wtg w/in Strata	1.1	1.0	1.0	1.0	59	187	75	321
	Inv. Propensity Wtg*	2.0	0.6	1.3	1.0	112	114	95	321
2	n	55	95	66	216	55	95	66	216
	Weighting w/in Strata	1.3	0.8	1.1	1.0	72	72	72	216
	Prop'l Wtg w/in Strata	0.7	1.3	0.8	1.0	40	126	51	216
	Inv. Propensity Wtg*	1.3	0.7	1.2	1.0	69	70	80	218
3	n	20	89	32	141	20	89	32	141
	Weighting w/in Strata	2.4	0.5	1.5	1.0	47	47	47	141
	Prop'l Wtg w/in Strata	1.3	0.9	1.0	1.0	26	82	33	141
	Inv. Propensity Wtg*	2.6	0.5	1.9	1.1	53	43	60	155
4	n	5	56	4	65	5	56	4	65
	Weighting w/in Strata	4.3	0.4	5.4	1.0	22	22	22	65
	Prop'l Wtg w/in Strata	2.4	0.7	3.8	1.0	12	38	15	65
	Inv. Propensity Wtg*	2.2	0.4	3.6	0.8	11	24	14	49
TOTAL	n	136	433	174	743	136	433	174	743
	Weighting w/in Strata	1.8	0.6	1.4	1.0	248	247	247	743

	Prop'l Wtg w/in Strata	1.0	1.0	1.0	1.0	136	433	174	743
	Inv. Propensity Wtg*	1.8	0.6	1.4	1.0	245	251	249	743

* Value for inverse propensity weighting are means, since each observation can have a different weight. For example, in the first quintile, the weights range from 1.5 to 3.6.

TABLE 7. COMPARISON OF DATA DISTRIBUTION AND RESULTS FROM NAÏVE ANALYSIS AND FOUR METHODS OF PROPENSITY SCORE ANALYSIS – POLYCHOTOMOUS OUTCOME

	Pre-			Random Selection Within Strata			Weighting Within Strata			Proportional Weighting Within Strata			Inverse Propensity Weighting		
	Resp	Home	Other	Resp	Home	Other	Resp	Home	Other	Resp	Home	Other	Resp	Home	Other
n	136	433	174	135	135	135	247.7	247.7	247.7	136	433	174	245.0	249.1	248.8
Age															
Young	19%	31%	21%	19%	19%	20%	26%	50%	42%	26%	27%	28%	24%	27%	27%
Mid	55%	51%	52%	56%	59%	56%	53%	39%	42%	53%	52%	48%	54%	52%	50%
Old	26%	19%	27%	26%	21%	24%	21%	11%	16%	21%	21%	23%	22%	22%	22%
p-value		0.01			0.94			0.86			0.91			0.93	
Race															
White	56%	39%	48%	56%	54%	50%	50%	44%	42%	50%	44%	42%	45%	44%	44%
Black	35%	44%	40%	35%	33%	37%	39%	41%	42%	39%	41%	42%	42%	42%	41%
HispOth	10%	17%	13%	10%	13%	13%	11%	15%	16%	11%	15%	16%	14%	15%	15%
p-value		0.006			0.86			0.24			0.50			0.996	
AA	34%	31%	34%	34%	31%	35%	32%	32%	34%	32%	32%	34%	32%	32%	32%
p-value		0.66			0.79			0.86			0.88			0.99	
DA	8%	19%	16%	8%	15%	14%	9%	17%	19%	9%	17%	19%	15%	16%	17%
p-value		0.009			0.19			0.003			0.03			0.77	
DCG															
Low	10%	17%	5%	10%	5%	4%	17%	13%	10%	17%	13%	10%	11%	13%	12%
Mid	69%	65%	72%	70%	74%	72%	65%	68%	69%	65%	68%	69%	70%	67%	67%
High	21%	18%	23%	21%	21%	24%	17%	19%	21%	17%	19%	21%	19%	20%	20%
p-value		0.002			0.43			0.20			0.44			0.94	

OR H	1.36 (0.82, 2.26)	1.20 (0.65, 2.20)	1.41 (0.89, 2.22)	1.41 (0.85, 2.35)	1.50 (0.95, 2.37)
OR O	1.35 (0.77, 2.37)	1.06 (0.57, 1.96)	1.31 (0.83, 2.08)	1.31 (0.73, 2.35)	1.43 (0.90, 2.26)

SUMMARY

With the addition of weighting within strata and proportional weighting within strata, there are now eight methods for sample selection in propensity score analysis. The two new methods share the desirable properties of inverse probability weighting in that they use all the data and do not require randomization techniques which results in non-replicability of study results. Weighting within strata also adds a nice conceptual understanding, similar to random weighting within strata, as well as greatly reducing the problem of very high weights on observations unlike all others in the same treatment group, which can be a problem in weighting by inverse propensity scores. Proportional weighting provides standard errors are not artificially inflated by assuming equal sample sizes. These conclusions are reinforced by examination of the raw results from the respite dataset.

From the simulations, I demonstrated that propensity score regression, random selection within strata, and weighting within strata produce results that are the least biased, based on their coverage probabilities. The greedy algorithm might not be a good estimate in some cases compared to the more optimal nearest neighbor matching.

The eight methods can be summarized by the following table (8). The following notation is used:

p(Z) is the propensity score based on a matrix of data Z (which may or may not include some of X),

Q(p) is the quantile of the propensity score,

T is the treatment group (1 for treatment and 0 for control),

$n_{T(Q)}$ and $n_{C(Q)}$, are the number of treated and control observations in the Qth quantile, respectively

n_T is the overall number of treated (the sum of $n_{T(Q)}$ over all Q),

n_C is the overall number of control observations,

w_i is the weight for observation i,

$Y_{T(i)}$ is the i^{th} treated observation, for i=1 to n_T, and

$Y_{C(j)}$ is the j^{th} control observation, for $j=1$ to n_{C}.

TABLE 8: SUMMARY OF METHODS OF SAMPLE SELECTION METHODS

Method	Weighting Scheme	Randomization Necessary?	Comments
Random Selection w/in Strata	1 if T=1 and $n_{T(Q)}<n_{C(Q)}$ or T=0 and $n_{T(Q)}>n_{C(Q)}$ 0 otherwise	Y	
Greedy Algorithm Matching	$Y_{T(i)}$=1 and $Y_{C(j)}$=1 if $Y_{T(i)} - Y_{C(j)} \leq Y_{T(i)} - Y_{C(k)}$ for all $k \neq j$ for i=1,…., n_T 0 otherwise	N (but order matters)	Sub-optimal but easy to program
Nearest Neighbor Matching	1 for all i,j such that $\sum_i \min_j [Y_{T(i)} - Y_{C(j)}]$ is the minimum, and each j is used only once 0 otherwise	N	
Nearest Neighbor Matching w/in Caliper	1 for all i,j such that $\sum_i \min_j [Y_{T(i)} - Y_{C(j)}]$ such that $Y_{T(i)} - Y_{C(j)} < \varepsilon$, where ε is the caliper, is the	N	Suggested method by Rosenbaum and Rubin (1985)

	minimum, and each j is used only once 0 otherwise		
Regression Adjustment	None (use p(Z) in regression equation on outcome)	N	
Inverse Propensity Weighting	1/p(Z) if T=1 1/(1-p(Z)) if T=0	N	Imbens (2000)
Weighting w/in Strata	$(n_{T(Q)}+n_{C(Q)})$ / 2 / $n_{T(Q)}$ if T=1 $(n_{T(Q)}+n_{C(Q)})$ / 2 / $n_{C(Q)}$ if T=0	N	New method. Easily generalizable to k-group polychotomous situation (use $\Sigma n_{i(Q)}$ / (k * $n_{i(Q)}$) instead, where n_i is the sample size in quantile Q for treatment group T=i
Proportional Weighting w/in Strata	Multiply weight w/in strata weight by $\Sigma n_{(Q)}$ / $\Sigma n_{i(Q)}$ where $\Sigma n_{(Q)}$ is the overall sample size (n) and $\Sigma n_{i(Q)}$ is the sample size in treatment group T=I	N	New method. Easily generalizable to a k-group polychotomous situation, as above.