Published in final edited form as:

Nurs Res. 2016 ; 65(6): 465–474. doi:10.1097/NNR.000000000000187.

Which Propensity Score Method Best Reduces Confounder Imbalance? An Example From a Retrospective Evaluation of a Childhood Obesity Intervention

Krista Schroeder, PhD, RN, CCRN [Postdoctoral Fellow],

University of Pennsylvania, School of Nursing, Philadelphia, PA

Haomiao Jia, PhD [Associate Professor of Biostatistics], and

Columbia University, School of Nursing, New York, NY

Arlene Smaldone, PhD, CPNP, CDE [Associate Professor of Nursing and Dental Behavioral Sciences]

Columbia University, School of Nursing, New York, NY

Abstract

Background—Propensity score (PS) methods are increasingly being employed by researchers to reduce bias arising from confounder imbalance when using observational data to examine intervention effects.

Objective—The purpose of this study was to examine PS theory and methodology and compare application of three PS methods (matching, stratification, weighting) to determine which best improves confounder balance.

Methods—Baseline characteristics of a sample of 20,518 school-age children with severe obesity (of whom 1,054 received an obesity intervention) were assessed prior to PS application. Three PS methods were then applied to the data to determine which demonstrated the greatest improvement in confounder balance between the intervention and control group. The effect of each PS method on the outcome variable—body mass index (BMI) percentile change at one year—was also examined. SAS 9.4 and Comprehensive Meta-Analysis statistical software were used for analyses.

Results—Prior to PS adjustment, the intervention and control groups differed significantly on seven of 11 potential confounders. PS matching removed all differences. PS stratification and weighting both removed one difference but created two new differences. Sensitivity analyses did not change these results. BMI percentile at one year decreased in both groups. The size of the

Krista Schroeder, PhD, RN, CCRN, was a Pre-doctoral Fellow, Columbia University, School of Nursing, New York, NY.

Haomiao Jia, PhD, is Associate Professor of Biostatistics, Columbia University, School of Nursing, New York, NY.

Arlene Smaldone, PhD, CPNP, CDE, is Associate Professor of Nursing and Dental Behavioral Sciences, and Assistant Dean of Scholarship and Research, Columbia University, School of Nursing, New York, NY.

Institutional Review Board approval was obtained from Columbia University Medical Center, the New York City Department of Health and Mental Hygiene, and the New York City Department of Education.

The authors have no conflicts of interest to report.

Supplemental Digital Content 1. Sample SAS code used for implementing each PS method.

Supplemental Digital Content 2. Sample characteristics after PS matching are presented.

Supplemental Digital Content 3. Sample characteristics after PS stratification are presented.

Supplemental Digital Content 4. Sample characteristics after PS weighting are presented.

decrease was smaller in the intervention group, and the estimate of the decrease varied by PS method.

Discussion—Selection of a PS method should be guided by insight from statistical theory and simulation experiments, in addition to observed improvement in confounder balance. For this dataset, PS matching worked best to correct confounder imbalance. Because each method varied in correcting confounder imbalance, we recommend that multiple PS methods be compared for ability to improve confounder balance before implementation in evaluating treatment effects in observational data.

Keywords

children; nonrandomized controlled trial; obesity; propensity score; quasi-experimental study; school nursing

Childhood obesity is one of the nation's greatest child health threats, with 17% of children meeting criteria for obesity (body mass index [BMI] 95th percentile for age and sex) (Ogden, Carroll, Kit, & Flegal, 2014). Childhood obesity is associated with many health comorbidities during childhood, negative psychosocial and academic outcomes, and adult obesity (Daniels, 2006; Freedman et al., 2005; Rappaport, Daskalakis, & Andrel, 2011). Schools are an ideal setting in which to implement childhood obesity interventions (Institute of Medicine, 2012; Waters et al., 2011), and school-based obesity interventions can be effective in reducing children's BMI percentile (Katz, O'Connell, Njike, Yeh, & Nawaz, 2008; Schroeder, Travers, & Smaldone, 2016; Sobol-Goldberg, Rabinowitz, & Gross, 2013; Waters et al., 2011). However, many school-based obesity interventions are implemented in a voluntary manner, with only students who are interested in the intervention actually receiving the intervention. Students who do not want to participate (or whose parents do not want their child to participate) do not receive the intervention. As a result of lack of randomization, key differences may exist between intervention participants and nonparticipants (Shadish, Cook, & Campbell, 2002). Differences may include factors that are known to influence childhood obesity (e.g., gender, race/ethnicity, socioeconomic status) (Davison & Birch, 2001); these differences can confound the relationship between the intervention and outcome of interest, in this case BMI percentile change. Such confounder imbalance introduces bias (Rubin & Zell, 2016). Therefore, statistical methods that reduce bias in observational studies are required.

Various statistical methods can be used by the researcher to control for bias when assignment to treatment or control conditions is not at random (in quasi-experimental or observational studies; Eckardt & Rindskopf, 2016). One important approach is application of propensity scores (Rosenbaum & Rubin, 1983). A propensity score (PS), developed by Rosenbaum and Rubin (1983), is the probability of an individual being in the intervention group, given his or her baseline characteristics. The PS is calculated using logistic regression, in which the individual's characteristics (potential confounders) are the predictors and probability of being in the intervention group is the outcome. Propensity scores can be applied in observational studies to reduce bias in understanding the relationship between an intervention and outcome. Based on the potential outcomes framework, the unconfounded effect of an intervention can be ascertained when the PS

balances all confounders between the nonrandomized intervention and control groups (Rubin & Zell, 2016). Various methods of propensity score application exist (Rosenbaum & Rubin, 1983). Propensity score methods may be superior to other methods of controlling for confounding in nursing research, such as sample stratification and matching (which can only account for a limited number of known confounders) or regression analysis (which may result in residual bias if the intervention and control group are heavily imbalanced on baseline characteristics) (D'Agostino & Kwan, 1995; Qin, Titler, Shever, & Kim, 2008; Rosenbaum & Rubin, 1983).

The scholarly discussion about and use of propensity scores in both the general biomedical and nursing literature has markedly increased during the past 10 years (Figure 1), necessitating that nurse scientists who work with observational data have a working knowledge of this analytic technique (Eckardt, 2012; Qin et al., 2008; Shadish & Steiner, 2010). Need for PS use will likely increase with the explosion of observational data available to nurse scientists via electronic medical records (EMRs) (Clarke & Cossette, 2000; Lin, Jiao, Biskupiak, & McAdam-Marx, 2013). Currently, many papers in the nursing literature that discuss propensity score methods do so by applying propensity scores prior to their study analyses (e.g., Moser et al., 2014; Stimpfel, Rosen, & McHugh, 2014). This paper is one of the few within the nursing literature to focus on PS methodology (Eckardt, 2012; Qin et al., 2008; Shadish & Steiner, 2010) and, to our knowledge, the first in the nursing literature to examine effects of different PS methods on improvement in confounder balance.

Purpose

The purpose of our study was to apply three propensity score methods to an observational dataset in order to determine which method best improved confounder balance. We measured confounder balance by examining the number of confounders with significant differences between the intervention and control groups. We chose this approach to measuring confounder balance because it is easily observable, quantifiable, and applicable across various PS methods.

Propensity Scores

Definition and Creation

A propensity score is the conditional probability of assignment to an intervention given a vector of observed covariates (Rosenbaum & Rubin, 1983). Applied to nonrandomized intervention studies, a participant's PS is the probability of receiving the treatment or intervention in question given characteristics such as gender, race, or neighborhood poverty level. Because a PS is a probability, values range from 0 to 1. Two participants with identical propensity scores can be considered to have the same probability of receiving the intervention. In this manner, the PS can help create a proxy intervention and control group within observational cohort data (D'Agostino, 1998; D'Agostino & Kwan, 1995). For example, a student with a PS of 0.72 who received the intervention can be considered comparable on key confounders to a student with a PS of 0.72 who did not receive the intervention. Because of their identical propensity scores, these two participants can be

considered comparable, as the PS has adjusted for their differences. It is important to note that propensity scores are created using measured covariates only. Therefore, propensity scores methods cannot mimic a randomized controlled trial (D'Agostino, 1998; D'Agostino & Kwan, 1995; Wagner, 2012). However, in large datasets, many variables (i.e., potential confounders) are often available to the researcher for creation of a PS, reducing the risk of excluding important confounders (Brookhart et al., 2006; D'Agostino, 1998).

Logistic regression equations are used to compute propensity scores for each person. The outcome is the log odds of a binary variable (receiving the treatment or being in the control group). The predictors are the k potential confounders of treatment effect. The equation is

$$\ln[p/(1-p)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

where p is the probability of being in the treatment group, $(1-p)$ is the probability of being in the control group, and the coefficients give the difference in log odds of being in the treatment group for each unit difference in the confounder adjusting for all other confounders. Once the coefficients have been estimated, the propensity score is obtained as the estimated probability of membership in the treatment group.

The PS can then be used in a number of ways to balance confounders in the estimate of treatment effect on the experimental outcome. When determining which predictors to include in the PS model, Rosenbaum (2002) cautions against using only predictors which significantly differ between groups because (a) this does not consider the relationship between predictor and outcome; (b) just because the difference between groups on a predictor is not statistically significant, it doesn't mean that it can be ignored; and (c) this process considers predictors only one at a time, whereas the logistic model considers the predictors as a group. As such, it is suggested that all available predictors potentially related to the outcome are included in a PS model unless there is a theoretical reason not to do so (Brookhart et al., 2006).

After creation of the PS, data should be further examined before proceeding with analysis. The distribution of propensity scores by group should be examined via graphical display using a histogram or boxplot. Presence of a large overlapping area of propensity scores, or "common support," indicates that the use of propensity scores will help to balance the intervention and control groups on key confounders (Wagner, 2012). Though there is no theoretical guidance on what exactly merits common support, it can be considered the range of propensity scores with at least five observations per propensity score in both the intervention and control groups (Li, Kleinman, & Gillman, 2014). Limiting analyses to only participants whose propensity scores fall under the common support is recommended, because characteristics of individuals outside the common support may be too different to compare without introducing significant bias in the estimate of treatment effect (Li, Morgan, & Zaslavsky, 2014; Wagner, 2012). After the common support is examined, the PS can be applied to the data using multiple methods. We compared three methods—PS matching, PS stratification, and PS weighting—to determine which method best improved confounder balance in one dataset.

Propensity Score Methods

Matching—Using this method, participants who received the intervention are matched by PS to participants who did not receive the intervention (Rosenbaum & Rubin, 1985a), creating two groups that are comparable on potential confounders. After creation of the groups, the treatment outcomes for each group can be directly compared to estimate the intervention effect (Austin, 2011; D’Agostino, 1998; Hill & Reiter, 2006; Rosenbaum & Rubin, 1985a).

Several methodological decisions are required when implementing PS matching: (a) ratio of control to intervention group, (b) replacement versus nonreplacement, (c) greedy versus optimal, and (d) nearest neighbor versus caliper matching. The ratio refers to how many control subjects are matched to an intervention subject; usually, a 1:1 ratio is selected, but many-to-one matching may be employed (Gu & Rosenbaum, 1993). Matching can be performed either with or without replacement (Austin, 2011). When matching with replacement, control participants who are matched to intervention participants are returned to the pool for potential matching with another intervention participant. Matching with replacement requires a special variance estimator to consider the fact that one control participant may be matched to multiple intervention participants (or vice versa) (Hill & Reiter, 2006). Greedy versus optimal matching involves using proximity to match intervention and control participants. Greedy methods match the intervention participant with the closest control participant, regardless of whether the control participant would be a closer match for another intervention participant. Optimal matching is based on minimizing the total within-pair difference on propensity scores (Austin, 2011). Optimal matching has been found to be no better than greedy matching in reducing estimator bias (Gu & Rosenbaum, 1993). Lastly, PS matching may be performed using either nearest neighbor or caliper methods (or a combination of both methods). In nearest neighbor methods, an untreated participant is matched to a treated participant with the closest PS, but no defined PS distance is considered too great for matching. Alternately, in caliper matching, a predetermined distance between propensity scores, defined by the investigator, is considered the maximum allowable distance. Using the caliper matching approach, a participant lacking a match within that caliper is excluded from analysis. Recommended optimal caliper distances range between 0.05 (Kurth et al., 2006) and 0.2 *SD* of the logit of the propensity score (Austin, 2011). The nearest neighbor method is the easiest from a computational standpoint and may be superior to caliper matching methods in minimizing bias in the treatment effect estimate (Austin, 2010), though nearest neighbor matching within calipers defined by the PS may be superior in balancing covariates between groups (Rosenbaum & Rubin, 1985b). Of note, PS matching may lead to some participants being excluded from the final analyzed sample because they do not have a match within the specified caliper (e.g., closest match is within 0.07, but specified caliper is 0.05) or exceed the matching ratio (e.g., dataset includes more control than intervention participants, but matching ratio is 1:1). When a propensity score matching procedure is used, it is recommended that outcome analyses be conducted using a paired *t*-test or a Wilcoxon signed rank test because the matching process introduces a correlated structure to the data (Lanehart et al., 2012).

Stratification—In PS stratification, participants are first ranked by PS and then divided into strata (Austin, 2011). Though there is no predetermined number of strata that should be used in all studies, five are typically recommended because five strata can reduce up to 90% of bias (Cochran, 1968; Rosenbaum & Rubin, 1984); every participant is included in one of the five strata. Within each stratum, propensity scores of participants are similar—more similar than they are across the sample as a whole. After participants are divided into strata, the effect of treatment on outcome can be estimated for the intervention versus control group within each stratum. Outcome analyses can be conducted by pooling the weighted outcome for each stratum, with each stratum weighted by the percentage of participants within that stratum (Austin, 2011). For example, if participants are divided into five strata, then the outcome in each stratum will be weighted by 1/5, or 0.20, when estimating the effect of treatment on the outcome. This can be conceptualized as a “meta-analysis of a set of quasi-RCTs” (Austin, 2011).

Weighting—In PS weighting, participants are weighted by the inverse of their PS. This weighted sample can be used to determine an unconfounded estimate of the outcome (F. Li et al., 2014). Outcome scores of intervention participants are weighted by a factor of $1/PS$; scores of participants who did not receive the intervention are weighted by $1/(1-PS)$. PS weighting uses exact PS values, which avoids the possibility of residual confounding that exists when inexact matching and stratification are used (F. Li et al., 2014). One concern with PS weighting is that standard errors of the treatment effect may be larger than with other PS methods due to extreme weighting of participants with propensity scores near 0 or near 1. One potential solution to this issue is to trim weights (Potter, 1993) by excluding individuals with extreme weights (e.g., <1st or >99th percentile) from the analysis (Cole & Hernan, 2008; Lee, Lessler, & Stuart, 2011). Doing so may increase accuracy of the treatment effect (Kurth et al., 2006). Another method to minimize extreme weighting is to use bounded overlap weights. A discussion of bounded overlap weights method is beyond the scope of this paper, but interested readers are referred to L. Li et al. (2014). Once weighting has been applied to the data, outcome analyses can be conducted using weighted least squares regression.

Sample size—Final analytic sample sizes differ for each PS method. For PS 1:1 matching, the sample size is twice the number of intervention participants. For PS stratification, the sample size includes all intervention and control participants. For PS weighting, the sample size includes all intervention and control participants. Sample SAS code used for implementing each PS method is available (see Supplemental Digital Content 1).

Data Source

We designed a retrospective observational study using electronic medical record data to examine the effect of a school nurse-led obesity program on BMI percentile change at one year of kindergarten to 5th grade students who met the criteria for severe obesity (BMI for age and sex at the 99th percentile or 120% of the 95th percentile; Flegal et al., 2009). The program—the Healthy Options and Physical Activity Program (HOP)—was implemented in New York City (NYC) public schools in 2010. The goal of HOP is to help children improve health behaviors and maintain or decrease BMI percentile. During HOP sessions, school

nurses provide education and counseling, and assist students with goal setting around five health behaviors: fruit/vegetable intake, sedentary media use, physical activity, sugar-sweetened beverage intake, and portion size. School nurses document participation in the program in the students' electronic medical records. During the 2012/2013 school year, 1,054 children participated in HOP (intervention group) and 19,464 were eligible for, but did not participate in, HOP (control group). Because HOP participation is voluntary and not randomly assigned, there are likely inherent differences between children with severe obesity who received HOP and children with severe obesity who were eligible for but did not receive HOP.

Data used in the creation of the PS came from three sources: the EMR in the public schools (primary data source), the NYC Department of Health and Mental Hygiene Office of School Health Reports, and the NYC Center for Economic Opportunity. Variables were organized using the socioecological model, which acknowledges that individual, family, institution, and community-level factors influence a child's body weight (Davison & Birch, 2001). Although the purpose of the HOP study was to examine multiple aspects of HOP implementation and efficacy, in this paper we limit discussion to improvement in confounder balance resulting from use of propensity score methods.

PS Creation

Eleven potential confounders (community poverty level, school poverty level, school nurse workload, household food insecurity, baseline BMI, baseline BMI percentile, age, race/ethnicity, grade, gender, and diagnosis of at least one chronic illness) were available for incorporation into a PS. Logistic regression based on participation in HOP (1 = *yes*, 0 = *no*) was used to compute the PS. Following creation of the PS, a histogram was examined to assess common support and each PS method was applied.

Using PS matching, each intervention child was matched to a control child (1:1 matching) with the most similar PS (caliper of .05). Greedy matching without replacement was employed. Once a pair was matched, the control child was not returned to the pool for further matching. Using PS stratification, intervention children and control children were divided into five equal strata based on propensity scores. Using PS weighting, each intervention child was assigned a weight of $1/PS$, and each control child was assigned a weight of $1/(1-PS)$. Weights were then normalized by dividing each weight by the mean weight.

Improvement in Confounder Balance

Distribution of key confounders between the intervention and control groups was compared before application of propensity scores and after use of each PS method. For each confounder of interest, the difference between the intervention and control groups was tested using either Wilcoxon rank sum tests for continuous variables or χ^2 tests for dichotomous or categorical variables. For PS matching only, differences were tested using methods that account for the correlated structure of the matched data (Wilcoxon signed rank tests for continuous variables, McNemar tests for binary variables, and Friedman tests for nonbinary variables). Following application of each PS method, the total number of variables

demonstrating significant differences between the intervention and control groups was tallied. For PS stratification, differences between the intervention and control group were tested within each of the five strata; if a statistically significant difference existed in at least one strata, the confounder was counted as a significant difference even if it did not differ within the other four strata. All analyses were performed using SAS 9.4.

Impact on Outcome Variable

An exploratory analysis was conducted to examine the impact of each PS method on estimation of the treatment effect on the outcome variable—BMI percentile change at one year. Prior to application of propensity scores, BMI percentile change at one year was calculated for the intervention and control groups. Group differences were then compared using a Wilcoxon rank sum test. Following PS application, differences between groups were analyzed using the Wilcoxon signed rank test (matching), a fixed effect meta-analysis model (stratification), and weighted ordinary least squares regression (weighting) using SAS 9.4 (Cary, NC) and Comprehensive Meta-analysis statistical software (Comprehensive Meta-analysis, 2015).

Institutional Review Board approval was obtained from Columbia University Medical Center, the New York City Department of Health and Mental Hygiene, and the New York City Department of Education.

Results

Of 20,518 children with severe obesity, 1,054 (5.1%) received the HOP intervention. Baseline characteristics of the sample prior to application of propensity scores are listed in Table 1. Groups differed on seven of 11 potential key confounders. On average, intervention children went to a school where fewer students received free/reduced lunch and the nurse workload was lower. Intervention children had higher baseline body measures, were older, and were more likely to have at least one chronic condition.

PS Method and Confounding

Data from one or more variables required for PS creation were missing from 75 participants (3.7%), leaving a total sample of 1,049 children in the intervention group and 19,394 children in the control group for PS analysis. Propensity scores ranged between .002 and .621. The mean propensity score in the intervention group was .085 ($SD = 0.063$; range: .005–.491); the control group mean was .050 ($SD = .040$; range: .002–.621). The area of common support is shown in the histograms in Figure 2; visual inspection of a histogram demonstrated that there appeared to be good common support. Table 2 summarizes confounder distribution before and after application of each PS method.

Matching—Two equal groups of 1,049 participants were created. It was possible to match each intervention child with a control child within a .05 caliper. Using this method, all significant differences between the intervention and control groups were removed (see Table 2; Supplemental Digital Content 2). Sensitivity analysis was conducted by repeating the process with a 1:4 matching ratio; this did not change the results.

Stratification—Five strata were constructed. Each stratum included the same number of children; however, the number of intervention children differed among strata because intervention children had higher propensity scores than control children. Stratum 1 had 59 intervention children (1.4% of stratum); stratum 2 had 90 intervention children (2.2% of stratum); stratum 3 had 145 intervention children (3.6% of stratum); stratum 4 had 277 intervention children (6.8% of stratum); and stratum 5 had 478 intervention children (11.7% of stratum). As shown in Table 2, one significant difference (grade in school) was removed from all five strata, but new significant differences were created for two confounders (household food insecurity and race/ethnicity); these differences occurred in only one and two of the five strata, respectively. Following stratification, differences between the intervention and control groups remained for six variables: school poverty level (2 of 5 strata); school nurse workload (3 of 5 strata); baseline BMI percentile (1 of 5 strata); baseline BMI (3 of 5 strata); age (1 of 5 strata); and diagnosis of at least one chronic illness (1 of 5 strata). Results using this method are presented in greater detail in Supplemental Digital Content 3. Following use of PS stratification, differences between the intervention and control groups remained for eight of the 11 confounders.

Weighting—PS weighting applied a normalized weight to each child in the sample. Weights differed between groups. For the control group, normalized weights ranged between 0.50 and 1.32 ($M = 0.53$; $Mdn = 0.52$). For the intervention group, normalized weights ranged from 1.02 to 99.39 ($M = 9.75$; $Mdn = 7.24$). Because of these differences, we explored limiting the sample to children with normalized weights between the 5th and 95th percentile (Kurth et al., 2006; Potter, 1993). This action resulted in exclusion of 98% of children from the intervention group and, therefore, was not employed. Application of normalized weights to the full sample, one difference between groups was removed (grade) and two new differences (household food insecurity, race/ethnicity) were created (Table 2). Additional details may be found in Supplemental Digital Content 4. To further explore this result, two analyses were performed using the normalized log of the weight and the normalized square root of the weight. Results using the square root of the weight were similar (one difference removed, two differences created), as were the results using the log of the weight (no difference removed or created). Therefore, following PS weighting, differences between groups remained for eight of 11 confounders.

PS Method and Estimation of Treatment Effect

Table 3 presents the impact of the PS method on estimating the effect of the treatment on the outcome variable (BMI percentile change). Without propensity score adjustment, the intervention group had a 0.12 smaller decrease in BMI percentile compared to the control group ($p < .001$). With PS matching, the intervention group had 0.05 smaller BMI percentile decrease compared to the control group ($p = .01$). In the PS stratification analysis, the intervention group had a 0.14 smaller BMI percentile decrease ($p < .001$). With PS weighting, the intervention group demonstrated a 0.02 smaller BMI percentile decrease ($p = .001$).

Discussion

We compared confounder balance following use of three PS methods by applying them to one dataset with information about children with severe obesity—some of whom took part in the HOP intervention. Use of propensity scores was effective in reducing significant differences between groups. Prior to propensity score application, groups differed on seven of 11 potential confounders. After PS matching, no confounders differed between groups. After PS stratification, eight potential confounders differed between groups, though only two confounders differed in more than half of the five strata. After PS weighting, eight potential confounders differed between groups. For this dataset, PS matching was most effective in improving confounder balance. Assessment of the impact of PS method on estimation of the treatment effect on the outcome demonstrated that change in BMI percentile at one year varied by PS method. Estimated treatment effect on BMI percentile change was most similar to the uncontrolled analysis with PS stratification (0.02 greater than uncontrolled analysis) and most different with PS weighting (0.10 less than uncontrolled analysis) and PS matching (0.07 less than uncontrolled analysis).

We computed propensity scores using three approaches and then illustrated that choice of PS method may influence the estimate of treatment effect on the outcome. In practice, outcome analyses should not be conducted until *after* a PS method is selected.

Use of PS weighting and stratification methods increased confounder imbalance between groups. Stratification likely increased differences between groups due to our strict criterion for improvement in confounder balance. We considered balance to be improved only if differences were removed in all five strata. Researchers who impose less stringent criteria may find that stratification is more effective. The weighting procedure likely increased differences between groups because in this dataset, receipt of the intervention was a rare event (approximately 5%; 1,059 of 20,518 students). Rare events result in very small propensity scores. As a result, weights for the intervention group ($1/PS$) were much larger than the weights for the control group ($1/(1-PS)$). Further testing using the log of the normalized weights and the square root of the normalized weights did not improve the results. Thus, despite the reported strengths of the weighting method (Kurth et al., 2006; F. Li et al., 2014), we concluded that PS weighting may not be well suited for datasets where receipt of the intervention of interest is rare.

PS methods have inherent limitations. First, the PS can only adjust for confounders used in the PS logistic model. It cannot adjust for unknown confounders (Austin, 2011; Rosenbaum & Rubin, 1983). Large datasets with many available confounders are ideal for use with propensity score methods, yet no dataset will include all possible confounders. Because randomization theoretically balances groups on both known and unknown confounders, it remains the superior approach for minimizing bias. Second, use of propensity scores reduces bias in observational data, allowing approximation to a randomized design. However, propensity score application only reduces and does not remove the limitations inherent to observational studies—findings are limited to association and causality may not be inferred (Austin, 2011; Rosenbaum & Rubin, 1983). Therefore, a randomized controlled trial

remains the gold standard for evaluating an intervention and inferring causality (Shadish et al., 2002, p.13).

To our knowledge, our study is the first within the nursing literature to compare multiple PS methods. Examples from other fields have demonstrated conflicting results. In one study, researchers demonstrated that both stratification and matching removed confounder differences between groups—though matching did so more effectively (Austin & Mamdani, 2006). The researchers noted one weakness of matching to be the loss of participants due to inability to find a match; this was not a concern in our study because of the large pool of untreated participants. Another study (Kurth et al., 2006) demonstrated that different PS methods yielded widely differing results when extreme PS values (< 5th percentile) were included. When the data were clipped to remove such values, the methods become comparable. However, the authors noted that neither method was necessarily superior in reducing bias in estimating treatment effect; the best method depends on the sample of interest, the clinical question, and the data (Kurth et al., 2006).

Many researchers implement a PS method without first examining the improvement in confounder balance following its implementation or evaluating more than one PS method (Austin & Mamdani, 2006). The findings of this study highlight the importance of assessing multiple PS methods, as they may produce different results. It is particularly important to note that propensity score methods must always be selected a priori and evaluated in absence of outcome data, in order to avoid the potential selection of the propensity score method that produces the most desirable outcome analysis. For this reason, it is important for investigators to describe the propensity score method selection and evaluation process they used when reporting intervention studies that utilize propensity scores.

Limitations

Our study has several limitations. We created the propensity scores for this dataset using one logistic model that was determined a priori. We also did not limit our analyses to only propensity scores that fell under their common support. Additional PS model development and limitation of propensity scores to the range of scores common support may have led to different findings. Secondly, our results reflect PS application to only one dataset. Our dataset also may not be typical in that only a small number of subjects (5%) received the intervention, which led to differences in sample size between PS matching ($n = 2,058$) and PS stratification and PS weighting ($n = 20,443$). Thus, our comparison of the three methods was biased towards finding fewer significant differences between groups in the matching analysis because the sample size was smaller. In addition, we applied only the three most common PS methods; additional methods exist (Austin, 2011; Rosenbaum & Rubin, 1983). Also, our measure of confounder balance considered only whether or not a significant confounder differed between the intervention and control groups; it did not consider magnitude of differences between the intervention and control groups, nor did it consider significance of differences in confounder balance improvement between PS methods (Gelman & Stern, 2006). Lastly, additional potential confounders that would be ideally included in our PS—such as parental attitudes towards HOP—were not available in our dataset.

Conclusion

Our analysis of applying propensity scores to a dataset of a nurse-led obesity intervention demonstrated that propensity scores can be effectively applied to improve confounder balance in observational studies (studies of treatment effect not assigned at random). Of the three methods compared, PS matching removed all significant differences in confounders between groups, whereas PS stratification and weighting removed some pre-existing differences but created new differences. These results are likely influenced by the fact that only a small proportion of participants received the intervention. Because of these differences, we recommend that nurse scientists test multiple PS methods before selecting a method for use in their analyses. By adding PS methods to their toolbox, nurse scientists can harness the increasingly available large datasets to conduct studies with reduced bias to create a stronger body of knowledge to improve health.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge that this publication was supported by the National Institute of Nursing Research through grant numbers T32 NR014205 and T32 NR007100, and the National Center for Advancing Translational Sciences through Grant Number UL1 TR000040. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would also like to acknowledge the New York City Department of Health and Mental Hygiene (CT, TA, KK, JH) for their assistance with this research project and manuscript development.

References

- Austin PC. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology*. 2010; 172:1092–1097. [PubMed: 20802241]
- Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*. 2011; 46:399–424. [PubMed: 21818162]
- Austin PC, Mamdani MM. A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine*. 2006; 25:2084–2106. [PubMed: 16220490]
- Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *American Journal of Epidemiology*. 2006; 163:1149–1156. [PubMed: 16624967]
- Clarke SP, Cossette S. Secondary analysis: Theoretical, methodological, and practical considerations. *Canadian Journal of Nursing Research*. 2000; 32:109–129. [PubMed: 11928128]
- Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 1968:295–313. [PubMed: 5683871]
- Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*. 2008; 168:656–664. [PubMed: 18682488]
- Comprehensive Meta-Analysis. Comprehensive meta-analysis. 2015 Retrieved from <http://www.meta-analysis.com/index.php>.
- D'Agostino RB. Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*. 1998; 17:2265–2228. <http://bacbuc.hd.free.fr/WebDAV/data/DOM/StatMeths/DAGostino-SM1998.pdf>. [PubMed: 9802183]

- D'Agostino RB, Kwan H. Measuring effectiveness: What to expect without a randomized control group. *Medical Care*. 1995; 33:AS95–AS105. [PubMed: 7723466]
- Daniels SR. The consequences of childhood overweight and obesity. *Future of Children*. 2006; 16:47–67. [PubMed: 16532658]
- Davison KK, Birch LL. Childhood overweight: A contextual model and recommendations for future research. *Obesity Reviews*. 2001; 2:159–171. [PubMed: 12120101]
- Eckardt P. Propensity score estimates in multilevel models for causal inference. *Nursing Research*. 2012; 61:213–223. [PubMed: 22551996]
- Eckardt, P.; Rindskopf, D. Quasi-experimental design in nursing research. In: Henly, SJ., editor. *The Routledge international handbook of advanced quantitative methods in nursing research*. Abingdon, OX, UK: Routledge/Taylor & Francis; 2016. p. 224-237.
- Flegal KM, Wei R, Ogden CL, Freedman DS, Johnson CL, Curtin LR. Characterizing extreme values of body mass index-for-age by using the 2000 Centers for Disease Control and Prevention growth charts. *American Journal of Clinical Nutrition*. 2009; 90:1314–1320. [PubMed: 19776142]
- Freedman DS, Khan LK, Serdula MK, Dietz WH, Srinivasan SR, Berenson GS. The relation of childhood BMI to adult adiposity: The Bogalusa Heart Study. *Pediatrics*. 2005; 115:22–27. [PubMed: 15629977]
- Gelman A, Stern H. The difference between “significant” and “not significant” is not itself statistically significant. *American Statistician*. 2006; 60:328–331.
- Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*. 1993; 2:405–420.
- Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*. 2006; 25:2230–2256. [PubMed: 16220488]
- Institute of Medicine. Accelerating progress in obesity prevention: Solving the weight of the nation. 2012 Retrieved from.
- Katz DL, O'Connell M, Njike VY, Yeh M-C, Nawaz H. Strategies for the prevention and control of obesity in the school setting: Systematic review and meta-analysis. *International Journal of Obesity*. 2008; 32:1780–1789. [PubMed: 19079319]
- Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*. 2006; 163:262–270. [PubMed: 16371515]
- Lanehart, RE.; de Gil, PR.; Kim, ES.; Bellara, AP.; Kromrey, JD.; Lee, RS. Propensity score analysis and assessment of propensity score approaches using SAS® procedures; Paper presented at the SAS Global Forum: Statistics and Data Analysis; Orlando, FL. 2012 Apr. Retrieved from <http://support.sas.com/resources/papers/proceedings12/314-2012.pdf>
- Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLOS ONE*. 2011; 6(3):e18174. [PubMed: 21483818]
- Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. 2014 Retrived from *arXiv preprint arXiv:1404.1785*.
- Li L, Kleinman K, Gillman MW. A comparison of confounding adjustment methods with an application to early life determinants of childhood obesity. *Journal of Developmental Origins of Health and Disease*. 2014; 5:435–447. [PubMed: 25171142]
- Lin J, Jiao T, Biskupiak JE, McAdam-Marx C. Application of electronic medical record data for health outcomes research: A review of recent literature. *Expert Review of Pharmacoeconomics & Outcomes Research*. 2013; 13:191–200. [PubMed: 23570430]
- Moser DK, Lee KS, Wu J-R, Mud-Martin G, Jaarsma T, Huang T-Y, Riegel B. Identification of symptom clusters among patients with heart failure: An international observational study. *International Journal of Nursing Studies*. 2014; 51:1366–1372. [PubMed: 24636665]
- Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of childhood and adult obesity in the United States, 2011–2012. *JAMA*. 2014; 311:806–814. [PubMed: 24570244]
- Potter, FJ. Proceedings of the American Statistical Association, Section on Survey Research Methods. Washington, DC: American Statistical Association; 1993. The effect of weight trimming on

nonlinear survey estimates; p. 758-763. Retrieved from http://www.amstat.org/sections/srms/Proceedings/papers/1993_127.pdf

- Qin R, Titler MG, Shever LL, Kim T. Estimating effects of nursing intervention via propensity score analysis. *Nursing Research*. 2008; 57:444–452. [PubMed: 19018219]
- Rappaport EB, Daskalakis C, Andrel J. Obesity and other predictors of absenteeism in Philadelphia school children. *Journal of School Health*. 2011; 81:341–344. [PubMed: 21592129]
- Rosenbaum, PR. *Observational studies*. New York, NY: Springer; 2002. Overt bias in observational studies; p. 71-104.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
- Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*. 1984; 79:516–524.
- Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*. 1985a; 39:33–38.
- Rosenbaum PR, Rubin DB. The bias due to incomplete matching. *Biometrics*. 1985b; 41:103–116. [PubMed: 4005368]
- Rubin, DB.; Zell, ER. Causality in experiments and observational studies. In: Henly, SJ., editor. *The Routledge international handbook of advanced quantitative methods in nursing research*. Abingdon, OX, UK: Routledge/Taylor & Francis; 2016. p. 213-223.
- Schroeder K, Travers J, Smaldone A. Are school nurses an overlooked resource in reducing childhood obesity? A systematic review and meta-analysis. *Journal of School Health*. 2016; 86:309–321. [PubMed: 27040469]
- Shadish, WR.; Cook, TD.; Campbell, DT. *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Cengage Learning; 2002.
- Shadish WR, Steiner PM. A primer on propensity score analysis. *Newborn and Infant Nursing Reviews*. 2010; 10:19–26.
- Sobol-Goldberg S, Rabinowitz J, Gross R. School-based obesity prevention programs: A meta-analysis of randomized controlled trials. *Obesity*. 2013; 21:2422–2428. [PubMed: 23794226]
- Stimpfel AW, Rosen JE, McHugh MD. Understanding the role of the professional practice environment on quality of care in Magnet® and non-Magnet hospitals. *Journal of Nursing Administration*. 2014; 44:10–16. [PubMed: 24316613]
- Wagner, T. Propensity scores. United States Department of Veterans Affairs Health Services Research and Development Cyberseminar Series; 2012 Apr. Retrieved from http://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/video_archive.cfm?SessionID=530
- Waters E, de Silva-Sanigorski A, Hall BJ, Brown T, Campbell KJ, Gao Y, Summerbell CD. Interventions for preventing obesity in children. *Cochrane Database of Systematic Reviews*. 2011; 12:CD001871.

FIGURE 1.
Publications retrieved from PubMed and Cumulative Index to Nursing and Allied Health Literature (CINAHL) from 2004–2013 using the key term *propensity score**.

FIGURE 2.

Propensity score distributions in the non-HOP group (Panel A) and HOP group (Panel B). The area of common support is demarcated by the boxed area. HOP = Healthy Options and Physical Activity Program.

TABLE 1

Sample Characteristics: Treated and Untreated Groups

Level/characteristic	Treated (<i>n</i> = 1,054)		Untreated (<i>n</i> = 19,464)		Difference	<i>p</i>
	<i>M</i>	(<i>SD</i>)	<i>M</i>	(<i>SD</i>)		
Community						
Poverty in community (%)	23.4	(6.3)	23.8	(6.5)	0.4	.06
Institutional						
Poverty in school (%)	71.0	(20.1)	74.1	(18.0)	3.1	<.001
School nurse workload ^a	13.2	(6.6)	14.6	(6.4)	1.4	<.001
Interpersonal or individual						
Age in months	99.5	(19.8)	91.0	(21.5)	8.5	<.001
BMI	29.8	(4.9)	27.1	(4.4)	2.7	<.001
BMI percentile	99.5	(0.3)	99.4	(0.3)	0.1	<.001
	<i>n</i>	(%)	<i>n</i>	(%)		
Food insecurity (yes)	871	(82.6)	15805	(81.2)	(1.4)	.26
Gender						.07
Male	620	(58.8)	11,990	(61.6)	(2.8)	
Female	434	(41.2)	7474	(38.4)	(2.8)	
Grade						<.001
Kindergarten	73	(6.9)	4000	(20.6)	(13.7)	
1 st	172	(16.3)	3939	(20.2)	(3.9)	
2 nd	228	(21.6)	3684	(18.9)	(2.7)	
3 rd	198	(18.8)	3042	(15.6)	(3.2)	
4 th	202	(19.2)	2555	(13.1)	(6.1)	
5 th	181	(17.2)	2244	(11.5)	(5.7)	
Race/ethnicity						.10
Non-Hispanic White	113	(10.7)	1927	(9.9)	(0.8)	
Non-Hispanic Black	227	(21.5)	4924	(25.3)	(3.8)	
Hispanic	620	(58.8)	10973	(56.4)	(2.4)	

Level/characteristic	Treated (<i>n</i> = 1,054)		Untreated (<i>n</i> = 19,464)		Difference	<i>p</i>
	<i>M</i>	(<i>SD</i>)	<i>M</i>	(<i>SD</i>)		
Asian ^b	83	(7.9)	1426	(7.3)	(0.6)	
AI/AN	6	(0.6)	158	(0.8)	(0.2)	
Multiracial	4	(0.5)	64	(0.3)	(0.2)	
>1 chronic illness	485	(46.0)	5936	(30.5)	(15.5)	<.001

Note. Treated children received the HOP intervention; untreated children didn't receive the intervention. AI/AN = American Indian/Alaska Native; BMI = body mass index; HOP = Healthy Options and Physical Activity Program; *SD* = standard deviation.

^aSchool nurse workload is a composite measure of workload intensity based on number of children with diabetes, medication administrations, and walk-in visits at each school over one year. Range from 1.6 to 36.5 (*M* = 13.3) for this sample.

^bAsian, Native Hawaiian, or Pacific Islander.

TABLE 2

Confounder Imbalance in Treated and Untreated Groups Following Application of Three Propensity Score Methods

Confounder	Before	Matching (n = 2,098)	Stratification (n = 20,443)	Weighting (n = 20,443)
Community poverty	No	No	No	No
School poverty	Yes	No	Yes	Yes
School nurse workload	Yes	No	Yes	Yes
Household food insecurity	No	No	Yes	Yes
BMI	Yes	No	Yes	Yes
BMI percentile	Yes	No	Yes	Yes
Gender	No	No	No	No
Age	Yes	No	Yes	Yes
Grade	Yes	No	No	No
Race/ethnicity	No	No	Yes	Yes
1 chronic illness	Yes	No	Yes	Yes
Changes in balance				
Differences removed		7	1	1
Differences created		0	2	2
Total differences	7 of 11	0 of 11	8 of 11	8 of 11

Note. BMI = body mass index.

TABLE 3

BMI Percentile Change Before and After Application of Propensity Score Methods

Method	Treated	Untreated	Difference ^a	p ^b
No PS (<i>n</i> = 20,518)	−0.17	−0.29	0.12	<.001
Matching (<i>n</i> = 2,098)	−0.17	−0.22	0.05	.01
Stratification (<i>n</i> = 20,443)			0.14	<.001
Strata 0 (<i>n</i> = 4,088)	−0.51	−0.42	−0.09	.15
Strata 1 (<i>n</i> = 4,089)	−0.24	−0.32	0.08	.20
Strata 2 (<i>n</i> = 4,089)	−0.12	−0.25	0.13	.001
Strata 3 (<i>n</i> = 4,089)	−0.17	−0.23	0.06	.09
Strata 4 (<i>n</i> = 4,088)	−0.13	−0.19	0.06	.01
Weighting (<i>n</i> = 20,443)	−0.26	−0.28	0.02	.001

Note. PS = propensity score.

^aDifference in BMI percentile change in treated group (intervention) minus untreated group (control).

^bSignificance of the difference.