Received: 14 December 2016 | Accepted: 15 December 2016

DOI 10.1111/jep.12701

WILEY Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

ORIGINAL ARTICLE

A comparison of approaches for stratifying on the propensity score to reduce bias

Ariel Linden DrPH, [1,2]

[1] President, Linden Consulting Group, LLC, Ann Arbor, Michigan, USA

[2] Research Scientist, Division of General Medicine, Medical School, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Ariel Linden, Linden Consulting Group, LLC, 1301 North Bay Drive, Ann Arbor, Michigan 48103, USA.
Email: alinden@lindenconsulting.org

Abstract

Rationale, aims, and objectives Stratification is a popular propensity score (PS) adjustment technique. It has been shown that stratifying the PS into 5 quantiles can remove over 90% of the bias due to the covariates used to generate the PS. Because of this finding, many investigators partition their data into 5 quantiles of the PS without examining whether a more robust solution (one that increases covariate balance while potentially reducing bias in the outcome analysis) can be found for their data. Two approaches (referred to herein as PSCORE and PSTRATA) obtain the optimal stratification solution by repeatedly dividing the data into strata until balance is achieved between treatment and control groups on the PS. These algorithms differ in how they partition the data, and it is not known which is better, or if either is better than a 5-quantile default approach, for reducing bias in treatment effect estimates.

Method Monte Carlo simulations and empirical data are used to assess whether PS strata defined by PSCORE, PSTRATA, or 5 quantiles is best at reducing bias in treatment effect estimates, when used within a marginal mean weighting framework (MMWS). These estimates are further compared to results derived using inverse probability of treatment weights (IPTW).

Results PSTRATA was slightly better than PSCORE in balancing covariates and reducing bias, while both approaches outperformed the 5-quantile approach. Overall MMWS using any stratification method outperformed IPTW.

Conclusions Investigators should routinely use stratification approaches that obtain the optimal stratification solution, rather than simply partitioning the data into 5 quantiles of the PS. Moreover, MMWS (in conjunction with an optimal stratification approach) should be considered as an alternative to IPTW in studies that use PS weights.

KEYWORDS

covariate balance, inverse probability of treatment, marginal mean weighting through stratification, propensity score, stratification, treatment effects

1 | INTRODUCTION

When conducting a randomized controlled trial is not feasible, investigators typically use observational data and rely on statistical methods to adjust for confounding. Although conventional regression modeling remains the most common adjustment approach, methods that explicitly model the treatment assignment—such as those using instrumental variables[1,2] or the propensity score[3]—are now being used more widely. In health research in particular, propensity scoring techniques have become increasingly popular as a means of controlling for confounding when estimating treatment effects.[4–7]

The propensity score is defined as the probability of assignment to the treatment group given the observed characteristics.[3] It has been demonstrated that in large samples, when treatment and control groups have similar distributions of the propensity score, they generally have similar distributions of the underlying covariates used to create the propensity score. This implies that observed preintervention covariates can be considered independent of treatment assignment (as if they were randomized) and therefore will not bias treatment effect estimates.[3]

Stratification (also known as subclassification[8,9]) is a principal propensity score adjustment technique that is straightforward to

implement and more interpretable than most other adjustment approaches.[10] Stratification may be considered a coarser version of matching, where treated and nontreated individuals within each stratum are expected to be comparable on pretreatment characteristics. It has been shown that stratifying the propensity score into 5 quantiles can remove over 90% of the initial bias due to the covariates used to generate the propensity score.[8,9] Given this finding, many investigators simply partition their data into 5 quantiles of the propensity score without examining whether a more robust solution (one that increases covariate balance while potentially reducing bias in the outcome analysis) can be found for their data.

Fortunately, 2 alternative approaches exist for obtaining the optimal stratification solution. The first approach initially splits the data into 5 quantiles of the propensity score, tests whether the treated and control groups are balanced on the propensity score within each quantile, and splits the quantile in half if balance is not achieved. The process of splitting quantiles into smaller strata is repeated until balance on the propensity score is achieved within each and every stratum (we will refer to this approach as PSCORE, following the name of the user-written program for Stata).[11] The second approach initially splits the data into 5 quantiles of the propensity score and tests whether the treated and control groups are balanced on the propensity score within each quantile. The overall number of quantiles is incrementally increased by 1 until balance on the propensity score is achieved between treatment and control groups, within all quantiles (we will refer to this approach as PSTRATA, following the name of the user-written program for Stata).[12] While the difference between approaches may appear subtle, in effect, the number of strata—and thus the number of observations within strata—may differ substantially. More specifically, in the PSCORE approach, it is likely that the various strata will differ in their sample sizes, whereas in the PSTRATA approach, all strata will have equal sample sizes. However, it has yet to be established if these 2 approaches differ in their effect on bias in the outcome analysis, or if either (or both) carries any advantage over simply stratifying the data into 5 propensity score quantiles. Therefore, the purpose of this paper is to investigate, using both Monte Carlo simulation and empirical data, whether either of these 2 approaches is superior to 5 propensity score quantiles for improving covariate balance and reducing bias.

This paper is organized as follows: Section 2 details the construction and results of the Monte Carlo simulation. Section 3 describes the empirical study and reports the results, and Section 4 provides discussion and conclusions.

2 | MONTE CARLO SIMULATION STUDY

In this simulation study, we examine if either of the 2 propensity score stratification approaches, PSCORE and PSTRATA, is superior to the standard application of 5 propensity score quantiles for reducing bias in the outcome model. The basic simulation design generally follows that described by Hong,[13] in which the estimated propensity score is misspecified to varying degrees (4 scenarios) and the effect on bias is assessed across 4 different outcome distributions (normal linear, normal nonlinear, Poisson, and Bernoulli)—for a total of 16 separate scenarios. In each scenario, 10 000 replications are drawn from the data-generating process described below and repeated for sample sizes of 500 and 2000. For each replication, the treatment effect estimate and standard error (SE) for each model are recorded. Bias (the difference between the simulated effect and the true effect of 1.0) and the root mean squared error (RMSE), which is a measure that magnifies and severely penalizes large errors, are then calculated across all samples. Lower values for all measures indicate better bias reduction.

2.1 | Data generating process for the treatment model

As in Hong[13] (Simulation 2), the true propensity score assigns treatment according to a polynomial function of X:

$$\Pr = \alpha_0 + \alpha_1 X + \alpha_2 X^2,$$

where X is drawn from a standard normal distribution with a mean of 0 and a standard deviation of 1, and α_0, α_1, and α_2 are manipulated to induce varying degrees of nonlinearity as follows:

Model 1: $\alpha_0 = 1$, $\alpha_1 = .2$, $\alpha_2 = -.2$;

Model 2: $\alpha_0 = 1$, $\alpha_1 = .6$, $\alpha_2 = -.2$;

Model 3: $\alpha_0 = 1$, $\alpha_1 = .2$, $\alpha_2 = -.6$;

Model 4: $\alpha_0 = 1$, $\alpha_1 = .6$, $\alpha_2 = -.6$.

The treatment assignment indicator Z is a Bernoulli random variable with the parameter of its distribution equal to the inverse logit of the true propensity score. A misspecified propensity score, which excludes the quadratic term X^2, is used in all simulation models.

2.2 | Data generating process for the outcome model

As in Hong,[13] 4 linear and nonlinear models for potential outcomes were generated for each set of simulations. The first model generated 2 normally distributed potential outcomes $Y(1)$ and $Y(0)$ corresponding to the experimental condition $Z = 1$ and the control condition $Z = 0$. Both $Y(1)$ and $Y(0)$ are linear functions of a standard normal covariate X:

$Y(1) = 6 + 0.7X + \epsilon(1)$;

$Y(0) = 5 + 0.7X + \epsilon(0)$;

$\epsilon(1), \epsilon(0) \sim N(0, 0.25)$.

In the second outcome model, $Y(1)$ and $Y(0)$ are polynomial functions of a standard normal covariate X:

$Y(1) = 6 + 0.5X + 0.25X^2 - 0.125X^3 + \epsilon(1)$;

$Y(0) = 5 + 0.5X + 0.25X^2 \square 0.125X^3 + \epsilon(0)$;

$\epsilon(1), \epsilon(0) \sim N(0, 0.25)$.

In the third outcome model, $Y(1)$ and $Y(0)$ follow Poisson distributions in which the parameters are each a nonlinear function of a standard normal covariate X:

$Y(1) \sim \text{Poisson}(\mu(1))$; $\mu(1) = \exp(3 + 0.7X)$;

$Y(0) \sim \text{Poisson}(\mu(0))$; $\mu(0) = \exp(2 + 0.7X)$.

In the fourth outcome model, $Y(1)$ and $Y(0)$ are each a Bernoulli random variate. Their parameters are each nonlinear function of a standard normal covariate X:

$Y(1) \sim \text{Bernoulli}(\mu(1))$; $\mu(1) = [1 + \exp(-0.5-0.7X)]^{-1}$;

$Y(0) \sim \text{Bernoulli}(\mu(0))$; $\mu(0) = [1 + \exp(0.5-0.7X)]^{-1}$.

In all models, the true treatment effect is equal to 1.

2.3 | Data preprocessing

For each replication in the simulation study, the propensity score is stratified in 3 ways: (1) dividing the propensity score into 5 quantiles (that are equally sized), (2) using the PSCORE algorithm in Stata to determine the optimal number of strata (that may or may not be equally sized),[11] and (3) using the PSTRATA algorithm in Stata to determine the optimal number of strata (that are equally sized).[12]

Stratification is only an intermediate step in the overall process of estimating treatment effects. Typically, once strata are defined, treatment effects are calculated within each stratum and are then pooled to obtain an overall weighted treatment effect estimate.[14] Here, we use an alternate approach called marginal mean weighting through stratification (MMWS)[13,15,16] to generate weights for each individual based on their corresponding stratum and treatment assignment. The marginal mean weights are computed based on the following formula by Hong[13]:

$$\frac{n_s \times \Pr(Z = z)}{n_{z=z,s}},$$

where n_s is the total number of individuals in a given stratum s, $\Pr(Z = z)$ is the probability of assignment to treatment group z, and $n_{z=z,s}$ is the total number of individuals in stratum s that were actually assigned to treatment z. The weights are then used as sampling weights in the outcome model. Thus, for each replication, 3 different MMWS weights are computed, corresponding to each of the stratification approaches.

As an additional comparator in the study, we compute inverse probability of treatment weights (IPTW),[17,18] where participants have a weight equal to the inverse of the estimated propensity score (1/propensity score), and nonparticipants have a weight equal to the inverse of 1 minus the estimated propensity score (1/1−propensity score). The IPTW is a popular propensity score weighting technique often used in health research for point-treatment, longitudinal, and survival studies,[18–22] among others. Hong[13] found that MMWS (using 6 strata) outperformed IPTW under a variety of simulation scenarios. We replicate those simulations here to see how the 3 stratification approaches compare to this widely used weighting approach.

For the current analyses, all weights were generated with the user-written program for Stata called MMWS,[16] specifying the common support option (ie, to drop observations when there is no individual in the opposing study group with a similar propensity score to serve as the counterfactual), with weights computed to represent the average treatment effect in the population.

2.4 | Model estimation

Linear regression models are used for the first 2 outcome scenarios, Poisson is used for the third scenario, and logistic regression is used in the fourth scenario. For each model, the corresponding outcome (Y) is regressed on the treatment assignment variable (Z) and the respective MMWS or IPTW weights are specified as probability weights. A naïve treatment effect is estimated to illustrate the impact of not adjusting for selection bias. This model is estimated as described above but excludes an adjustment weight. Robust (sandwich) SEs are used in all models. All simulations and analyses were conducted using Stata version 14.2 (College Station, Texas).

2.5 | Monte Carlo simulation results

Tables 1 and 2 present the results of all simulations for sample sizes of 2000 and 500, respectively. PSCORE and PSTRATA approaches performed similarly across all scenarios, with all estimates close to the actual treatment effect of 1 and achieving similar RMSE (Table 1). Both PSCORE and PSTRATA approaches outperformed the standard 5-quantile approach in the nonlinear, and Poisson models but performed similarly in the linear and logistic regression models. All 3 stratification approaches using MMWS had consistently closer estimates to the true treatment effect and lower RMSE than IPTW.

As shown in Table 2, with the smaller sample size, no consistent pattern emerges to suggest a superior stratification approach, with perhaps the exception of lower RMSE values using the PSTRATA approach. Interestingly, in the Bernoulli (logistic) outcome models, the standard 5-quantile approach overestimated the treatment effect by roughly the same amount that PSTRATA underestimated the treatment effect. However, PSTATA did reliably produce lower RMSE overall. As in the larger sample, all 3 stratification approaches using MMWS generally outperformed IPTW.

3 | EMPRICAL EXAMPLE

3.1 | Data

The empirical example uses data from a prior evaluation of a primary care-based medical home pilot program that invited patients to enroll if they had a chronic illness or were predicted to have high costs in the following year. The goal of the program was to lower health care costs for program participants by providing intensified primary care (see the study of Linden[23] for a more comprehensive description). The retrospectively collected data consist of observations for 374 program participants and 1628 nonparticipants. Eleven preintervention characteristics were available; these included demographic variables (age and gender), health services usage in year prior to enrollment (primary care visits, other outpatient visits, laboratory tests, radiology tests, prescriptions filled, hospitalizations, emergency department visits, and home-health visits) and total medical costs (the amount paid for all those health services utilized in the prior

TABLE 1 Treatment effect estimates and root mean squared errors for 10 000 Monte Carlo simulations with N = 2000

	Parameters			Naive	IPTW		MMWS-5		MMWS-PSCORE		MMWS-PSTRATA	
Outcomes	α_0	α_1	α_2	Mean (SE)	Mean (SE)	RMSE	Mean (SE)	RMSE	Mean (SE)	RMSE	Mean (SE)	RMSE
Normal, linear	1	0.2	−0.2	1.15 (0.04)	0.98 (0.02)	0.03	1.01 (0.02)	0.02	1.02 (0.03)	0.04	1.00 (0.01)	0.01
	1	0.6	−0.2	1.41 (0.04)	0.93 (0.03)	0.08	1.03 (0.02)	0.03	1.02 (0.02)	0.02	1.00 (0.01)	0.02
	1	0.2	−0.6	1.11 (0.04)	0.98 (0.03)	0.04	1.00 (0.02)	0.02	1.00 (0.02)	0.02	1.00 (0.01)	0.01
	1	0.6	−0.6	1.32 (0.04)	0.92 (0.04)	0.09	1.01 (0.02)	0.02	1.00 (0.02)	0.02	1.00 (0.01)	0.01
Normal, nonlinear	1	0.2	−0.2	0.92 (0.03)	0.91 (0.02)	0.09	0.96 (0.02)	0.05	0.98 (0.02)	0.03	0.99 (0.01)	0.02
	1	0.6	−0.2	0.98 (0.03)	0.92 (0.02)	0.08	0.96 (0.02)	0.04	1.00 (0.02)	0.02	0.99 (0.01)	0.02
	1	0.2	−0.6	0.80 (0.03)	0.84 (0.02)	0.16	0.94 (0.02)	0.06	0.99 (0.02)	0.02	0.99 (0.02)	0.02
	1	0.6	−0.6	0.89 (0.03)	0.85 (0.02)	0.15	0.95 (0.02)	0.05	0.99 (0.02)	0.02	1.00 (0.02)	0.02
Poisson	1	0.2	−0.2	1.05 (0.04)	0.89 (0.03)	0.11	0.97 (0.02)	0.04	0.99 (0.03)	0.03	0.99 (0.02)	0.02
	1	0.6	−0.2	1.33 (0.04)	0.84 (0.05)	0.17	1.00 (0.03)	0.03	1.01 (0.03)	0.03	1.00 (0.02)	0.02
	1	0.2	−0.6	0.88 (0.04)	0.82 (0.04)	0.18	0.94 (0.02)	0.07	0.99 (0.02)	0.03	0.99 (0.02)	0.02
	1	0.6	−0.6	1.10 (0.04)	0.75 (0.05)	0.26	0.94 (0.02)	0.07	0.98 (0.02)	0.03	0.99 (0.02)	0.02
Bernoulli	1	0.2	−0.2	1.17 (0.12)	1.00 (0.10)	0.10	1.02 (0.10)	0.10	1.03 (0.10)	0.11	1.01 (0.10)	0.10
	1	0.6	−0.2	1.48 (0.15)	0.96 (0.10)	0.11	1.03 (0.10)	0.11	1.02 (0.10)	0.11	1.01 (0.10)	0.10
	1	0.2	−0.6	1.16 (0.11)	1.02 (0.10)	0.10	1.02 (0.10)	0.10	1.01 (0.10)	0.10	1.01 (0.10)	0.10
	1	0.6	−0.6	1.40 (0.13)	0.98 (0.10)	0.10	1.02 (0.10)	0.10	1.01 (0.10)	0.10	1.00 (0.09)	0.09

The true treatment effect is 1.0, the true propensity score is equal to $\alpha_0 + \alpha_1 X + \alpha_2 X^2$, and the misspecified propensity score is equal to $\alpha_0 + \alpha_1 X$.

Abbreviations: IPTW, inverse probability of treatment weights; MMWS-5, marginal mean weighting with 5 strata; MMWS-PSCORE, marginal mean weights using the PSCORE approach for stratifying the sample; MMWS-PSTRATA, marginal mean weights using the PSTRATA approach for stratifying the sample; RMSE, root mean squared error; SE, standard error.

TABLE 2 Treatment effect estimates and root mean squared errors for 10 000 Monte Carlo simulations with N = 500

	Parameters			Naive	IPTW		MMWS-5		MMWS-PSCORE		MMWS-PSTRATA	
Outcomes	α0	α1	α2	Mean (SE)	Mean (SE)	RMSE	Mean (SE)	RMSE	Mean (SE)	RMSE	Mean (SE)	RMSE
Normal, linear	1	0.2	−0.2	1.15 (0.08)	0.97 (0.05)	0.05	1.00 (0.03)	0.03	1.07 (0.05)	0.08	1.00 (0.01)	0.01
	1	0.6	−0.2	1.41 (0.07)	0.92 (0.06)	0.10	1.02 (0.03)	0.04	1.05 (0.04)	0.06	1.01 (0.02)	0.02
	1	0.2	−0.6	1.11 (0.07)	0.98 (0.06)	0.07	1.00 (0.03)	0.03	0.99 (0.05)	0.05	1.00 (0.01)	0.01
	1	0.6	−0.6	1.32 (0.07)	0.93 (0.07)	0.10	1.00 (0.03)	0.03	0.99 (0.04)	0.04	0.99 (0.02)	0.02
Normal, nonlinear	1	0.2	−0.2	0.92 (0.06)	0.93 (0.04)	0.08	0.97 (0.04)	0.05	0.99 (0.05)	0.05	0.97 (0.02)	0.02
	1	0.6	−0.2	0.98 (0.05)	0.93 (0.05)	0.10	0.98 (0.03)	0.04	1.01 (0.03)	0.03	0.99 (0.02)	0.02
	1	0.2	−0.6	0.80 (0.05)	0.87 (0.04)	0.07	0.96 (0.03)	0.05	0.96 (0.05)	0.06	0.99 (0.01)	0.02
	1	0.6	−0.6	0.89 (0.05)	0.87 (0.04)	0.10	0.97 (0.03)	0.04	0.98 (0.03)	0.04	1.00 (0.01)	0.01
Poisson	1	0.2	−0.2	1.05 (0.09)	0.91 (0.07)	0.12	0.98 (0.05)	0.05	1.03 (0.06)	0.07	0.98 (0.01)	0.02
	1	0.6	−0.2	1.33 (0.09)	0.84 (0.10)	0.18	0.99 (0.05)	0.05	1.04 (0.05)	0.07	0.99 (0.02)	0.02
	1	0.2	−0.6	0.88 (0.08)	0.86 (0.07)	0.16	0.96 (0.04)	0.06	0.94 (0.06)	0.08	0.98 (0.01)	0.02
	1	0.6	−0.6	1.10 (0.09)	0.80 (0.09)	0.22	0.96 (0.04)	0.06	0.97 (0.05)	0.06	0.99 (0.02)	0.02
Bernoulli	1	0.2	−0.2	1.19 (0.24)	1.01 (0.20)	0.20	1.03 (0.20)	0.21	1.09 (0.22)	0.24	0.94 (0.08)	0.10
	1	0.6	−0.2	1.50 (0.30)	0.96 (0.20)	0.21	1.04 (0.21)	0.22	1.06 (0.22)	0.23	0.98 (0.04)	0.05
	1	0.2	−0.6	1.17 (0.22)	1.03 (0.20)	0.20	1.03 (0.20)	0.20	1.02 (0.20)	0.20	1.01 (0.08)	0.08
	1	0.6	−0.6	1.42 (0.27)	0.99 (0.20)	0.20	1.03 (0.21)	0.21	1.02 (0.21)	0.21	0.97 (0.08)	0.08

The true treatment effect is 1.0, the true propensity score is equal to $\alpha_0 + \alpha_1 X + \alpha_2 X^2$, and the misspecified propensity score is equal to $\alpha_0 + \alpha_1 X$.

Abbreviations: IPTW, inverse probability of treatment weights; MMWS-5 , marginal mean weighting with 5 strata; MMWS-PSCORE, marginal mean weights using the PSCORE approach for stratifying the sample; MMWS-PSTRATA, marginal mean weights using the PSTRATA approach for stratifying the sample; RMSE, root mean squared error.

year). The outcome was total medical costs paid in the year in which individuals received the intervention.

3.2 | Analytic approach

The first step in the analytic process to estimate treatment effects involves estimating the propensity score using logistic regression to predict program participation. Here, the treatment indicator is regressed on the 11 preintervention covariates described above, all entered as main effects.

In the second step, the propensity score is stratified using all 3 stratification approaches described earlier (5 quantiles, PSCORE, and PSTRATA).

In the third step, separate MMWS weights are computed for each individual—for each of the 3 stratification approaches. The weights generated here represent the average treatment effect on the treated, and all observations with no common support are dropped from the analysis. In this analysis, we do not include a comparison using IPTW.

Fourth, covariate balance is assessed between study groups and stratification approaches using the absolute standardized difference

in means as implemented in the user-written program for Stata COVBAL.[24] In observational data, neither the true propensity score nor the true treatment effect is known. Therefore, measures of bias, such as those used in a simulated experiment (where we do know these true values), cannot be used to assess how far the estimated treatment effect deviates from the true effect. Instead, investigators assess covariate balance of observed characteristics between study groups as an indicator of how well confounding is controlled for. In theory, better balance on covariates (indicated by smaller standardized differences) should result in less bias in the outcomes analysis (due to better control of confounding).

Finally, median (quantile) regression is used to estimate treatment effects, where the outcome (total program year costs) is regressed on the treatment variable, the respective MMWS weights are specified as probability weights, and SEs for the treatment effect are computed using a bootstrap with 1000 replications.[25] In addition to the 3 MMWS-weighted estimates, a naïve effect is estimated (where no control for confounding is conducted) to illustrate a completely biased treatment effect. Quantile regression is used in this example to better handle the skewed distribution of costs in this sample.[23]

3.3 | Empirical example results

Table 3 presents the baseline characteristics of the treatment group compared to controls derived via the 3 stratification approaches and unadjusted. The left panel presents the means of the various covariates using each of the stratification approaches, and the right panel presents the absolute standardized differences. The PSCORE approach stratified the sample into 8 strata, whereas the PSTRATA approach stratified the sample into 12 strata. As shown, the adjusted means are comparable between the treated group and controls, using all of the stratification approaches. Similarly, the standardized differences between treatment and control groups across all 3 stratification approaches are all under 0.10 (which is generally considered a threshold for covariate balance, even though values closest to 0 are desirable).[26] That said, the PSTRATA approach achieves the lowest average absolute standardized difference across all covariates (0.020), slightly outperforming the PSCORE approach (0.028) and substantially outperforming the 5-quantile approach (0.052). Taken as a whole, all 3 stratification approaches achieved covariate balance.

Table 4 presents the treatment effect estimates for the 3 stratification adjusted models and the naïve estimate. Given that the weights were adjusted to represent the average treatment effect on the treated, all individuals in the treatment group receive a weight of 1, and thus, the median program year cost estimates for the treated group is identical for all models ($4765). The median program year costs for the control group differs by stratification approach, but within a narrow range of $186 between the lowest estimate, derived using the 5-quantile approach ($4098), and the highest estimate, derived using the PSTRATA approach ($4284). That said, the 5-quantile approach elicited a statistically significant difference in costs between the treatment and control group (P = .021), whereas the other 2 stratification approaches produced estimates that were not statistically different. In summary, the 3 stratification approaches did a comparable job in adjusting the observed data for confounding; however, the treatment effect estimates were sufficiently sensitive to minor differences that 1 of the 3 approaches produced a statistically significant result whereas the other 2 did not.

TABLE 3 Comparison of baseline characteristics of the treatment and control groups—unadjusted and marginal mean weighted using various stratification approaches

	Means					Absolute Standardized Differences			
Variable	Treated	Controls Unadjusted	Controls MMWS-5	Controls PSCORE	Controls PSTRATA	Unadjusted	MMWS-5	PSCORE	PSTRATA
Propensity score	0.47	0.12	0.45	0.47	0.48	1.234	0.091	0.028	0.018
Age	54.86	43.44	54.85	54.59	54.94	0.496	0.002	0.041	0.012
Female	0.57	0.50	0.55	0.59	0.58	0.122	0.026	0.047	0.026
Hospitalizations	0.23	0.07	0.18	0.22	0.22	0.333	0.094	0.019	0.015
Hospital days	0.71	0.20	0.57	0.65	0.69	0.220	0.070	0.031	0.010
ED visits	0.36	0.16	0.36	0.39	0.32	0.225	0.004	0.033	0.043
Office visits	10.89	4.63	10.42	11.02	11.24	0.889	0.070	0.020	0.050
Out-patient visits	17.55	7.25	16.67	16.85	17.59	0.591	0.054	0.042	0.002
Laboratory	5.88	2.38	5.50	5.56	5.93	0.657	0.064	0.056	0.008
Radiology	3.14	1.31	2.93	3.15	3.30	0.379	0.048	0.002	0.035
Home health visits	0.09	0.02	0.07	0.11	0.11	0.090	0.022	0.016	0.021
Prescriptions	39.10	11.95	37.80	39.04	39.73	0.917	0.045	0.002	0.021
Total costs ($)	8059	3047	7249	7769	8044	0.477	0.085	0.030	0.002
Average std-diff						0.510	0.052	0.028	0.020

Sample sizes are 367 and 949, for treatment and control groups, respectively.

Abbreviations: ED, emergency department; Std-diff, standardized difference; MMWS-5, marginal mean weights with 5 strata; PSCORE, marginal mean weights using the PSCORE approach, stratifying the sample into 8 strata; PSTRATA, marginal mean weights using the PSTRATA approach, stratifying the sample into 12 strata.

TABLE 4 Treatment effect estimates using quantile (median) regression

Estimator	Treated	Control	Difference	*P* value	95% CI
Naïve (unadjusted)	4765	2269	2496	<.001	2116, 2876
MMWS-5	4765	4098	667	.021	101, 1233
MMWS-PSCORE	4765	4244	521	.080	−63, 1105
MMWS-PSTRATA	4765	4284	481	.124	−133, 1095

Abbreviations: MMWS-5, marginal mean weights with 5 strata; MMWS-PSCORE, marginal mean weights using the PSCORE approach for stratifying the sample; MMWS-PSTRATA, marginal mean weights using the PSTRATA approach for stratifying the sample.

4 | DISCUSSION

The results of the Monte Carlo simulation study indicate that the PSTRATA approach produces slightly lower bias than the PSCORE approach, which in turn produces lower bias than simply partitioning the data into 5 quantiles. Similarly, the results of the empirical study indicate that the PSTRATA approach produces slightly better covariate balance than the PSCORE approach, which in turn, produces better covariate balance than the 5-quantile approach. Taken together, these findings suggest that PSTRATA is a marginally superior stratification approach than the PSCORE approach, but that either approach outperforms stratifying the data in 5 quantiles of the propensity score. The results of the empirical study also indicate that treatment effect estimates are sensitive to the number of strata produced by the 3 approaches, given that PSTRATA found 12 quantiles to be a better solution for ensuring balance on the propensity score than the standard 5, and PSCORE partitioned the data into 8 strata to achieve an optimal solution. As a consequence, while the estimated treatment effects were only marginally different between the 3 approaches, the 5-quantile approach generated a statistically significant effect while the other 2 did not. This may suggest that, in these data, an increased number of strata results in less bias. This may further explain why PSTRATA slightly outperformed PSCORE (that is, PSTRATA consistently partitioned the data into more strata than PSCORE).

Taken together, these results support searching for the optimal number of propensity score strata rather than simply partitioning the data into 5 quantiles. Furthermore, given that PSTRATA and PSCORE approach the stratification problem somewhat differently, investigators may be best served by testing both algorithms as a sensitivity analysis.[27] If both approaches achieve similar levels of covariate balance and derive comparable treatment effects, the investigator will have greater confidence that the study results are unbiased. Tangentially, in the simulation study when used within an MMWS framework, all 3 stratification techniques outperformed IPTW in reducing bias in the treatment effect estimates. This result concurs with that of Hong,[13] who found that MMWS (with 6 strata) is less sensitive to misspecification of the propensity score than IPTW, consequently reducing bias in the treatment effect estimate. Thus, investigators may want to consider MMWS as an alternative to IPTW in all studies, which are designed to implement a propensity score weighting approach.[19,20,28–30]

The primary limitation of the simulation study is that the performance of alternative propensity score stratification approaches on subsequent treatment effects was considered in the context of a specific data generating process. It is unclear how estimator performance may vary across different data generating processes, especially those using additional variable types and distributions and violations to assumptions in the causal model.

In summary, investigators should routinely use stratification approaches that obtain the optimal stratification solution, rather than simply partitioning the data into 5 quantiles of the propensity score. These methods improve covariate balance and reduce bias in treatment effect estimates when compared to the 5-quantile approach. Moreover, when used within the MMWS framework, stratification appears to outperform the more widely used IPTW under a variety of conditions. Thus, investigators should consider using MMWS (in conjunction with an optimal stratification approach) as an alternative to IPTW in studies that use propensity score weights.

ACKNOWLEDGEMENT

I wish to thank Julia Adler-Milstein for reviewing the manuscript and providing many helpful comments.

REFERENCES

1. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc*. 1996;91:444–455.
2. Linden A, Adams J. Evaluating disease management program effectiveness: an introduction to instrumental variables. *J Eval Clin Pract*. 2006;12:148–154.
3. Rosenbaum PR, Rubin DB. The central role of propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41–55.
4. Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss SA. Review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*. 2006;59:437–447.
5. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1–21.
6. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *J Thorac Cardiovasc Surg*. 2007;134(5):128–1135.
7. Austin PCA. Critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med*. 2008;27(12):2037–2049.
8. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 1968;24:205–213.
9. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984;79:516–524.
10. Linden A, Adams JL. Improving participant selection in disease management programs: insights gained from propensity score stratification. *J Eval Clin Pract*. 2008;14:914–918.
11. Becker SO, Ichino A. Estimation of average treatment effects based on propensity scores. *Stata J*. 2002;2:358–377.

12. Linden A. PSTRATA: Stata module for implementing optimal propensity score stratification. Statistical Software Components s458232, Boston College Department of Economics, 2016. Downloadable from http://ideas.repec.org/c/boc/bocode/s458232.html [Accessed on 29 November 2016].
13. Hong G. Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. *J Educ Behav Stat*. 2010;35:499–531.
14. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23:2937–2960.
15. Linden A. Combining propensity score-based stratification and weighting to improve causal inference in the evaluation of health care interventions. *J Eval Clin Pract*. 2014;20:1065–1071.
16. Linden A. MMWS: Stata module for implementing mean marginal weighting through stratification. Statistical Software Components s457886, Boston College Department of Economics, 2014. Downloadable from http://ideas.repec.org/c/boc/bocode/s457886.html [Accessed on December 2 2016].
17. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc*. 1987;82:387–394.
18. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550–560.
19. Linden A, Adams JL. Using propensity score-based weighting in the evaluation of health management programme effectiveness. *J Eval Clin Pract*. 2010;16:175–179.
20. Linden A, Adams JL. Evaluating health management programmes over time. Application of propensity score-based weighting to longitudinal data. *J Eval Clin Pract*. 2010;16:180–185.
21. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: an introduction to survival analysis. *Dis Manag*. 2004;7(3):180–190.
22. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med*. 2013;32:2837–2849.
23. Linden A. Identifying spin in health management evaluations. *J Eval Clin Pract*. 2011;17:1223–1230.
24. Linden A. COVBAL: Stata module for generating covariate balance statistics. Statistical Software Components s458188, Boston College Department of Economics, 2016. Downloadable from http://ideas.repec.org/c/boc/bocode/s458188.html [Accessed on December 2 2016].
25. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: an introduction to the bootstrap technique. *Dis Manag Health Out*. 2005;13(3):159–167.
26. Linden A, Samuels SJ. Using balance statistics to determine the optimal number of controls in matching studies. *J Eval Clin Pract*. 2013;19:968–975.
27. Linden A, Adams J, Roberts N. Strengthening the case for disease management effectiveness: unhiding the hidden bias. *J Eval Clin Pract*. 2006;12:140–147.
28. Linden A, Adams JL. Applying a propensity-score based weighting model to interrupted time series data: improving causal inference in program evaluation. *J Eval Clin Pract*. 2011;17:1231–1238.
29. Linden A, Adams JL. Combining the regression-discontinuity design and propensity-score based weighting to improve causal inference in program evaluation. *J Eval Clin Pract*. 2012;18:317–325.
30. Linden A, Uysal SD, Ryan A, Adams JL. Estimating causal effects for multivalued treatments: a comparison of approaches. *Stat Med*. 2016;35(4):534–552.

How to cite this article: Linden A. A comparison of approaches for stratifying on the propensity score to reduce bias. *J Eval Clin Pract*. 2017. doi: 10.1111/jep.12701