HSR Health Services Research

DOI: 10.1111/1475-6773.12182
METHODS CORNER

Methods for Constructing and Assessing Propensity Scores

Melissa M. Garrido, Amy S. Kelley, Julia Paris, Katherine Roza, Diane E. Meier, R. Sean Morrison, and Melissa D. Aldridge

Objectives. To model the steps involved in preparing for and carrying out propensity score analyses by providing step-by-step guidance and Stata code applied to an empirical dataset.

Study Design. Guidance, Stata code, and empirical examples are given to illustrate (1) the process of choosing variables to include in the propensity score; (2) balance of propensity score across treatment and comparison groups; (3) balance of covariates across treatment and comparison groups within blocks of the propensity score; (4) choice of matching and weighting strategies; (5) balance of covariates after matching or weighting the sample; and (6) interpretation of treatment effect estimates.

Empirical Application. We use data from the Palliative Care for Cancer Patients (PC4C) study, a multisite observational study of the effect of inpatient palliative care on patient health outcomes and health services use, to illustrate the development and use of a propensity score.

Conclusions. Propensity scores are one useful tool for accounting for observed differences between treated and comparison groups. Careful testing of propensity scores is required before using them to estimate treatment effects.

Key Words. Observational data/quasi-experiments, administrative data uses, patient outcomes/function

Recent national initiatives for comparative effectiveness research recommend harnessing the power of existing data to evaluate health-related treatment effects (Patient-Centered Outcomes Research Institute 2012). A difficulty in using observational data is that patient and provider characteristics may be associated with both treatment selection and outcome, leading to different distributions of covariates within treatment and comparison groups. Propensity score analysis is a useful tool to account for imbalance in covariates between treated and comparison groups. A propensity score is a single score that represents the probability of receiving a treatment, conditional on a set of observed covariates. The goal of creating a propensity score is to balance covariates between individuals who did and did not receive a treatment, making it easier to isolate the effect of a treatment.

While the advantages and disadvantages of using propensity scores are well known (e.g., Stuart 2010; Brooks and Ohsfeldt 2013), it is difficult to find specific guidance with accompanying statistical code for the steps involved in creating and assessing propensity scores. Other useful Stata references gloss over propensity score assessment (treatment effects manual, StataCorp. 2013a; Stata YouTube channel, www.youtube.com/user/statacorp) or provide disjointed information (www.stata.com/statalist). Here, we synthesize information on creation and assessment of propensity scores within one article. In the following sections, we introduce situations in which propensity scores might be used in health services research and provide step-by-step instructions and Stata 13 code and output to illustrate (1) choice of variables to include in the propensity score; (2) balance of propensity score across treatment and comparison groups; (3) balance of covariates across treatment and comparison groups within blocks of the propensity score; (4) choice of matching and weighting strategies; (5) balance of covariates after matching or weighting the sample by a propensity score; and (6) interpretation of treatment effect estimates.

When to Consider Propensity Scores

Propensity scores are useful when estimating a treatment's effect on an outcome using observational data and when selection bias due to nonrandom treatment assignment is likely. The classic experimental design for estimating treatment effects is a randomized controlled trial (RCT), where random assignment to treatment balances individuals' observed and unobserved

Address correspondence to Melissa M. Garrido, Ph.D., GRECC, James J Peters VA Medical Center, Bronx, NY; Brookdale Department of Geriatrics and Palliative Medicine, Icahn School of Medicine at Mount Sinai, New York, NY; 130 W. Kingsbridge Road, Room 4A-17, Bronx, NY 10468; e-mail: melissa.garrido@mssm.edu. Amy S. Kelley, M.D., M.S.H.S., and Melissa D. Aldridge, Ph.D., M.B.A., are also with the Brookdale Department of Geriatrics and Palliative Medicine, Icahn School of Medicine at Mount Sinai, New York, NY; GRECC, James J Peters VA Medical Center, Bronx, NY. Julia Paris, B.A., and Katherine Roza, B.A., are with the Brookdale Department of Geriatrics and Palliative Medicine, Icahn School of Medicine at Mount Sinai, New York, NY; Diane E. Meier, M.D., is with the Center to Advance Palliative Care; Brookdale Department of Geriatrics and Palliative Medicine, Icahn School of Medicine at Mount Sinai, New York, NY. R. Sean Morrison, M.D., is with the National Palliative Care Research Center, Hertzberg Palliative Care Institute; Brookdale Department of Geriatrics and Palliative Medicine, Icahn School of Medicine at Mount Sinai, New York, NY; GRECC, James J Peters VA Medical Center, Bronx, NY.

characteristics across treatment and control groups. Because only one treatment state can be observed at a time for each individual, control individuals that are similar to treated individuals in everything but treatment receipt are used as proxies for the counterfactual. In observational data, however, treatment assignment is not random. This leads to selection bias, where measured and unmeasured characteristics of individuals are associated with likelihood of receiving treatment and with the outcome. Propensity scores provide a way to balance *measured* covariates across treatment and comparison groups and better approximate the counterfactual for treated individuals.

Propensity scores can be thought of as an advanced matching technique. For instance, if one were concerned that age might affect both treatment selection and outcome, one strategy would be to compare individuals of similar age in both treatment and comparison groups. As variables are added to the matching process, however, it becomes more and more difficult to find exact matches for individuals (i.e., it is unlikely to find individuals in both the treatment and comparison groups with identical gender, age, race, comorbidity level, and insurance status). Propensity scores solve this dimensionality problem by compressing the relevant factors into a single score. Individuals with similar propensity scores are then compared across treatment and comparison groups.

Within health services research, propensity scores are useful when randomization of treatments is impossible (Medicare demonstration projects) or unethical (end-of-life care). In addition, health services researchers are often interested in a treatment's effect on multiple outcomes (such as cost and quality), and a single propensity score can be used to evaluate multiple outcomes (Wyss et al. 2013). Recently, health services researchers have used propensity scores to reduce confounding due to selection bias in evaluations of the effects of physical health events on mental health service use (Yoon and Bernell 2013), assertive community treatment on medical costs (Slade et al. 2013), and pay-for-performance on Medicare costs (Kruse et al. 2012).

The theory and principles behind propensity scores are described elsewhere (Rubin 1980; Rosenbaum and Rubin 1984, 1985; Imbens 2004; Ho et al. 2007; Stuart 2010; Brooks and Ohsfeldt 2013). This article is an introductory "how-to" guide and focuses on the steps to create and assess propensity scores for a dichotomous treatment. More advanced readers may wish to use propensity scores with survey-weighted data (DuGoff, Schuler, and Stuart 2014) or with multilevel categorical (Imbens 2000; Huang et al. 2005) or continuous treatments (Jiang and Foster 2013). We use data from the Palliative Care for Cancer Patients (PC4C) study, an observational study of inpatient palliative care's effect on multiple health outcomes for individuals with cancer. We used propensity scores to account for the fact that patients' baseline health affects both probability of receiving palliative care and experiencing adverse health outcomes.

Data

PC4C patients were hospitalized in five facilities with established palliative care programs in New York, Ohio, Pennsylvania, Virginia, and Wisconsin. IRB approval was obtained from each study site. Eligible patients were 18 years of age or older, had an advanced cancer diagnosis, and spoke English. Nonverbal patients, patients with dementia, and those who had previously received palliative care were admitted for chemotherapy or had lengths of stay less than 48 hours were excluded. Of the 3,227 eligible patients who consented to participate, 1,537 (47.6 percent) had complete interview and medical record data. Most patients with incomplete data were too medically ill to continue study participation.

Our treatment variable was receipt of inpatient palliative care from an interdisciplinary dedicated consultation team. Care consisted of symptom assessment and treatment, goals of care discussions, and care transition planning. Data for the propensity score come from medical record review completed by trained project staff and patient baseline interviews and daily symptom inventories.

Stata Code and Output

Stata code fragments to accompany the steps listed below are detailed in the technical appendix. We present code integrated within Stata 13 (-teffects-; StataCorp. 2013b) as well as user-written commands that one downloads: -pscore- (st0026), -psmatch2-, -pstest- (within the -psmatch2- package), and -pbalchk- (Becker and Ichino 2002; Leuven and Sianesi 2003; Lunt 2013).

Although the -teffects- package constructs a propensity score and calculates a treatment effect with a one-line command (described in Step 6), it does not check whether the propensity score adequately balances covariates across treatment and comparison groups (described in Steps 3 and 5). Therefore, we recommend carrying out the following steps with user-written commands to construct and assess propensity scores before calculating treatment effects.

Steps Involved in Constructing and Assessing Propensity Scores

Step One: Choice of Variables to Include in the Propensity Score

Propensity scores are used to reduce confounding and thus include variables thought to be related to both treatment and outcome. To create a propensity score, a common first step is to use a logit or probit regression with treatment as the outcome variable and the potential confounders as explanatory variables. Covariate selection is guided by tradeoffs between variables' effects on bias (distance of estimated treatment effect from true effect) and efficiency (precision of estimated treatment effect).

If a variable is thought to be related to the outcome but not the treatment, including it in the propensity score should reduce bias (Brookhart et al. 2006; Austin 2011a). This is because there is a chance that a variable related to the outcome is also related to treatment. If it is not accounted for in the propensity score, it is an unmeasured confounder and will bias the treatment effect (Brookhart et al. 2006). With sufficiently large datasets, it is beneficial to include all variables that are potentially related to the outcome. In some cases, propensity scores can include hundreds of covariates. In smaller datasets, however, potentially irrelevant covariates may introduce too much "noise" into treatment effect estimates and obscure any reduction in bias that is achieved by their inclusion (Imbens 2004; Brookhart et al. 2006; Ho et al. 2007). In this case, consider excluding variables that may be only weakly associated with the outcome.

Controlling for variables that are hypothesized to be associated with treatment but *not* outcome, however, can decrease precision (by adding more "noise" to the estimate) and will not improve bias because they do not address confounding and are irrelevant for the purposes of the propensity score (Brookhart et al. 2006; Brooks and Ohsfeldt 2013).

Example. From our data, we chose variables from categories hypothesized to be associated with multiple outcomes (including readmission rates and symptom burden): medications, sociodemographics, advance care plans, help at home and place of residence before hospitalization, functional status, comorbidities, symptom burden, cancer site, and delirium. Some variables, such as help at home, were not hypothesized to be associated with palliative care receipt but are commonly associated with health outcomes and were included in case they were confounders. Others, such as attending physician identity, were not included, because they were hypothesized to be associated only with treatment likelihood and not with outcomes (Garrido et al. 2012).

Caution. Exclude from consideration covariates that might be affected by the treatment (Imbens 2004; Ho et al. 2007). A propensity score that includes covariates affected by the treatment (e.g., postconsult analgesic prescriptions in our dataset) obscures part of the treatment effect that one is trying to estimate. Exclude any covariates that predict treatment status perfectly, as distributions of covariates need to overlap between treatment and comparison groups (see Step 2). Finally, the propensity score should be created without knowledge of the outcome. Creation, balancing, and matching steps are akin to the preparatory steps of an RCT: treatment assignment occurs prior to provision of treatment and measurement of outcome.

Step Two: Balance of Propensity Score across Treatment and Comparison Groups

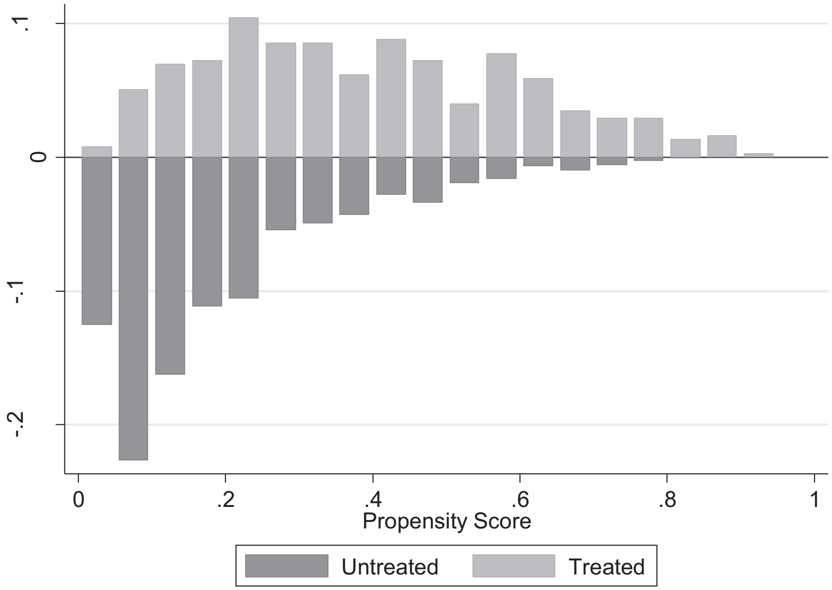
Once a propensity score has been calculated for each observation, one must ensure that there is overlap in the range of propensity scores across treatment and comparison groups (called "common support"). No inferences about treatment effects can be made for a treated individual for whom there is not a comparison individual with a similar propensity score. Common support is subjectively assessed by examining a graph of propensity scores across treatment and comparison groups (Figure 1).

Besides overlapping, the propensity score should have a similar distribution ("balance") in the treated and comparison groups. A rough estimate of the propensity score's distribution can be obtained by splitting the sample by quintiles of the propensity score. A starting test of balance is to ensure that the mean propensity score is equivalent in the treatment and comparison groups within each of the five quintiles (Imbens 2004). If it is not equivalent, one or more of the quintiles can be split into smaller blocks. If balance within smaller blocks cannot be achieved, the covariates or functional forms of covariates included in the score can be modified.

Example. The overlap of the distribution of the propensity scores across treatment and comparison groups is displayed in Figure 1. We found the extent of overlap to be satisfactory. In our final propensity score specification, balance

Figure 1: Distribution of Propensity Score across Treatment and Comparison Groups

psgraph, treated (*treatment*) **pscore**(*mypscore*)



was achieved across the treatment and comparison groups within all quintiles except Block 1. Block 1 was split into two blocks and balance was reevaluated. In this case, one split was sufficient to balance the propensity score within each block, leaving us with a total of six blocks (Data S1, eFigure 1).

Caution. Propensity scores only balance measured covariates, and balance in measured covariates does not necessarily indicate balance in unmeasured covariates. If unmeasured covariates are confounders, they can bias treatment effect estimates. This bias may increase as the relationship between measured and unmeasured covariates becomes stronger (Brooks and Ohsfeldt 2013).

Step Three: Balance of Covariates across Treatment and Comparison Groups within Blocks of the Propensity Score

After the propensity score is balanced within blocks across the treatment and comparison groups, a check for balance of individual covariates across treatment and comparison groups within blocks of the propensity score should be performed. This ensures that the propensity score's distribution is similar across groups within each block and that the propensity score is properly specified (Imbens 2004). There is no agreed-upon best method of balancing the propensity score. Imbalance in the mean indicates the propensity score needs to be respecified, but balance in the mean does not indicate balance in higher order moments (Basu, Polsky, and Manning 2008). Instead, one can compute standardized differences (which take into account both means and variances) (Rosenbaum and Rubin 1985; see Austin 2009a for equations).

There is no rule regarding how much imbalance is acceptable in a propensity score. Proposed maximum standardized differences for specific covariates range from 10 to 25 percent (Austin 2009a; Stuart, Lee, and Leacy 2013). Imbalance in some covariates is expected; even in RCTs, exact balance is a large-sample property (Austin 2009a). Balance in theoretically important covariates is more crucial than balance in covariates that are less likely to impact the outcome. More imbalance is expected at the tails of the propensity score's distribution, which include individuals who may be outside the range of common support. More detailed balance diagnostics are performed after the sample has been matched or weighted on the propensity score (Step 5).

The initial specification will likely not be balanced. In this case, possible solutions include dropping variables that are less theoretically important, recategorizing variables (e.g., making a continuous variable categorical or dichotomous), including interactions between variables, or including higher order terms or splines of variables. A transformed variable may have a slightly different distribution across treatment and comparison groups, enabling balance across groups to be achieved.

Example. We performed numerous iterations of Step 2 with changes in our list of potential confounders. Although in nearly every specification we achieved balance across groups within blocks (Step 2), it took over 100 iterations before we achieved balance in most of the specific covariates across groups within blocks. We achieved balance in all but one covariate in one block of the propensity score using *t*-tests (symptom count at reference day was unbalanced in block 2; Figure S2). We then evaluated the standardized differences of covariates across blocks of the propensity score. Of the variables tested, 89.7 percent had standardized differences ≤25 percent, with most larger standardized differences in the tails of the propensity score distribution (data not shown). Because this was not the final balancing step, we deemed these differences acceptable.

Dropping variables (including ones with little variability, such as delirium incidence) and recategorizing others (e.g., changing age from a continuous to a categorical variable) led us to a sufficiently balanced propensity score. Interacting variables (such as comorbidities and age) did not improve balance. We dropped 23 palliative care and 18 usual care patients from our sample with missing values for variables included in the propensity score. Other strategies for dealing with missing data in the context of propensity scores are described elsewhere (D'Agostino et al. 2001; Qu and Lipkovich 2009).

Caution. Do not use c-statistics or the area under the curve (AUC) to measure propensity score performance. The use of these measures is questionable, as propensity scores are intended for reducing confounding and not for predictive modeling (Stuart 2010). Moreover, simulation experiments have shown AUC to be unable to distinguish between correctly specified and misspecified propensity scores (Brookhart et al. 2006; Austin 2009a).

In addition, be cautious if using *t*-tests to check balance of covariates. Because the goal of matching is to ensure balance within a sample, the larger population from which the sample was drawn is not of concern. Moreover, *t*-tests are affected by sample size and might not be statistically significant even in the presence of covariate imbalance (Ho et al. 2007; Austin 2009a).

Step Four: Choice of Matching and Weighting Strategies

After creating a balanced propensity score, the next step is choosing how to use the propensity score to compare treatment and comparison groups. This choice involves evaluating tradeoffs between bias and efficiency. Matching and weighting strategies are discussed here, as they are among the most popular comparison strategies (Austin 2009b, 2011a).

Within matching strategies, a treated individual can be matched to the comparison individual with the most similar propensity score, no matter how poor the match (nearest neighbor) or within a certain caliper (.2 of the standard deviation of the logit of the propensity score[1] is optimal

[Austin 2011b]). One can match each treated individual to one or many comparison group individuals. When matching at the individual level, the first match is always best and will lead to the least biased estimates, but the decrease in bias from fewer matches needs to be weighed against the lower efficiency of the estimate that will occur with fewer observations. A broader one-to-many match will increase sample size and efficiency but can also result in greater bias from matches that are not as close as the initial match.

Rather than discarding unmatched individuals from the comparison group and reducing the sample size, a kernel weight can be used to estimate the counterfactual. While lesser known among health services researchers, kernel matching (also known as kernel weighting) is a potentially useful technique for researchers using survey data with sampling weights or continuous or multilevel categorical treatments, where other matching strategies are not always viable options (Imbens 2000; DuGoff, Schuler, and Stuart 2014). In kernel matching, each treated individual is given a weight of one. A weighted composite of comparison observations is used to create a match for each treated individual, where comparison individuals are weighted by their distance in propensity score from treated individuals within a range, or bandwidth, of the propensity score. Only observations outside the range of common support are discarded. Kernel matching maximizes precision (by retaining sample size) without worsening bias (by giving greater weight to better matches).

The bandwidths used in kernel functions are equivalent to half the width of bins in a histogram (DiNardo and Tobias 2001). Unlike bins in a histogram, bandwidths in a kernel function overlap. In addition, rather than assigning a single weight to each observation in a bin, as occurs in a histogram, a kernel function assigns higher weights to untreated individuals who have closer propensity scores to the treated individuals (see DiNardo and Tobias 2001 for formulas). The choice of bandwidth is more important than the specific kernel function (Caliendo and Kopeinig 2008). A bandwidth of 0.06 (propensity score −0.06 to propensity score +0.06) may optimize the tradeoff between variance and bias (Heckman, Ichimura, and Todd 1997), though others suggest using a bandwidth that increases with lower density of untreated individuals (Galdo, Smith, and Black 2008).

Kernel weights lend themselves to calculation of the average treatment effect on the treated. If an investigator is more interested in the average treatment effect on the entire sample, however, inverse-probability treatment weights (IPTW) may be chosen (Imbens 2004; Stuart 2010).

(More detail on treatment effects is presented in Step 6.) In IPTWs, each treated person receives a weight equal to the inverse of the propensity score, and each comparison individual receives a weight equal to the inverse of one minus the propensity score. IPTWs should be normalized to one (Imbens 2004).

More detailed discussions of the advantages and disadvantages of specific matching and weighting strategies are available elsewhere (Caliendo and Kopeinig 2008; Busso, DiNardo, and McCrary 2009; Stuart 2010; Huber, Lechner, and Wunsch 2013). There is not a clearly superior method of matching or weighting data by propensity scores; others recommend testing several methods and choosing the strategy that best balances the sample (Ho et al. 2007; Luo, Gardiner, and Bradley 2010) and fits the analytic goal (Stuart 2010). Stata code for some of the more popular strategies is listed in the technical appendix.

Example. With our dataset, we tried several matching and weighting strategies. Output from caliper matching, kernel matching, and IPTW are presented in conjunction with Step 5 (evaluating covariate balance after matching or weighting on the propensity score).

Step Five: Balance of Covariates after Matching or Weighting the Sample by a Propensity Score

After choosing a matching or weighting strategy, it is important to evaluate how well the treatment and comparison groups are balanced in the matched or weighted samples. If the treatment and comparison groups are poorly balanced, the propensity score needs to be respecified (Ho et al. 2007; Austin 2009a). As with the balancing steps outlined earlier, a common first test is comparing standardized differences. Smaller differences in means and higher order moments are better (Ho et al. 2007), especially in confounders hypothesized to be strongly related to the outcome.

Other balance diagnostics include graphs and variance ratios. With unweighted data, the distribution of a continuous covariate in the treated group can be plotted against its distribution in the comparison group in a quantile-quantile plot. If both distributions lie along a 45-degree line, the covariate is balanced (Stuart 2010). With weighted data, density functions of continuous covariates in treated and comparison groups can be graphed together and compared subjectively (Austin 2009a). In addition, the ratio of variances of the propensity score and covariates from the treatment and comparison groups should be near one if the treatment and comparison groups are balanced ("1/2 or 2 are far too extreme," p. 174, Rubin 2001). One can also compare the balance of interaction terms between treatment and comparison groups.

Because the outcome has not yet been examined, a range of balance diagnostics can be run for multiple matching and weighting strategies. If variables appear balanced within multiple checks, there is more evidence that the propensity score has been properly specified. The strategy that leads to the best balance can be chosen for outcome analyses.

Example. The mean standardized difference in covariates across treatment and comparison groups in the original sample was 24.6 percent (Table 1). Of the matching and weighting strategies, kernel matching and IPTW had the best reduction in mean standardized difference while retaining nearly all observations from the original sample. In IPTW, two conceptually important covariates (mean physical symptom severity score at baseline and type of cancer) had standardized differences >10 percent (data not shown). After kernel weighting, the means of every covariate were balanced across the treatment and comparison groups (standardized differences <10 percent for all covariates, and <5 percent for all covariates except for one pain measurement; Table 2). Kernel densities were plotted to examine distributions of continuous variables across matched treatment

Table 1: Sample Size, Mean, and Median Standardized Differences across All Covariates in Original and Matched and Weighted Samples

Sample Type	*Total Sample Size*	*Number of Treated Observations*	*Number of Comparison Observations*	*Mean Standardized Difference in Covariates (%)*	*Median Standardized Difference in Covariates (%)*
Original sample	1,537	374	1,163	24.6	23.9
Caliper 1:1 with replacement	614	374	240	5.4	4.8
Caliper 1:3 with replacement	885	374	511	3.4	2.2
Kernel matching	1,536	373	1,163	2.1	1.2
Inverse probability of treatment weighting	1,537	374	1,163	3.3	2.4

Table 2: Covariate Balance across Treatment and Comparison Groups before and after Matching or Weighting on the Propensity Score

	Original Sample			*Kernel Matched Sample*		
Variable	*Mean Treatment (n = 374)*	*Mean Comparison (n = 1,163)*	*Standardized Difference (%)*	*Mean Treatment (n = 373)*	*Mean Comparison (n = 1,163)*	*Standardized Difference (%)*
Sociodemographics (yes/no)						
Age 55–75	0.55	0.60	−10.0	0.55	0.55	1.1
Age > 75	0.11	0.10	3.2	0.12	0.12	−1.8
Female	0.56	0.57	−1.4	0.56	0.58	−3.8
Race – White	0.66	0.77	−24.1*	0.66	0.66	0.7
Race – Black	0.29	0.18	27.4*	0.29	0.30	−1.1
Education – College graduate or higher	0.40	0.53	−26.3*	0.40	0.40	0.8
Education – High school or some college	0.54	0.42	23.9*	0.54	0.54	−0.4
Medicare	0.34	0.32	4.4	0.35	0.36	−2.5
Medicaid	0.21	0.11	27.8*	0.20	0.22	−4.2
Medication in week before hospitalization						
Morphine equivalent dose (mg)	26.51	12.46	31.0*	25.31	23.29	4.5
Advance care plans (yes/no)						
Health care proxy	0.45	0.52	−15.1*	0.45	0.46	−1.7
Living will	0.39	0.45	−12.4*	0.39	0.39	0.0
Help at home in 2 weeks prior to the hospitalization						
Hours of home health aide help per week	0.74	0.97	−3.5	0.74	0.79	−0.6
Visiting nurse services (yes/no)	0.15	0.12	7.3	0.15	0.15	−0.5
Illness and symptom severity measures†						
Lymphoma/myeloma	0.05	0.08	−14.6*	0.05	0.05	−1.4
Hospital complications before reference day	0.03	0.08	−20.1*	0.03	0.04	−1.2

Continued

Table 2. *Continued*

Variable	*Original Sample* *Mean Treatment (n = 374)*	 *Mean Comparison (n = 1,163)*	 *Standardized Difference (%)*	*Kernel Matched Sample* *Mean Treatment (n = 373)*	 *Mean Comparison (n = 1,163)*	 *Standardized Difference (%)*
Number of Elixhauser[‡] comorbidities	4.06	3.20	43.7*	4.05	3.98	3.9
Needs complete assistance with 1+ ADL	0.13	0.07	21.6*	0.13	0.14	−2.8
Needs complete or partial assistance with bathing	0.34	0.18	38.2*	0.34	0.35	−1.9
Needs complete or partial assistance with transferring	0.35	0.14	49.3*	0.35	0.35	0.8
Symptom severity at baseline						
No. physical and psychological symptoms	8.88	6.97	56.0*	8.87	8.86	0.2
Mean severity of physical symptoms[§]	1.90	1.32	68.8*	1.90	1.91	−1.0
Mean severity of psychological symptoms[§]	1.59	1.45	11.2*	1.58	1.53	4.4
Symptom severity at reference day[¶]						
No. physical and psychological symptoms	7.85	6.66	39.4*	7.83	7.82	0.5
Mean severity of physical symptoms[§]	1.82	1.34	63.7*	1.82	1.79	3.6
Mean severity of psychological symptoms[§]	1.39	1.03	29.5*	1.39	1.45	−4.9
Pain: somewhat	0.11	0.14	−8.6	0.11	0.11	1.0
Pain: quite a bit	0.25	0.22	9.0	0.25	0.27	−4.5
Pain: very much	0.33	0.17	36.3*	0.33	0.30	7.3
Fatigue: a little, somewhat, or quite a bit	0.35	0.37	−4.9	0.35	0.36	−0.9
Fatigue: very much	0.31	0.18	29.3*	0.31	0.31	0.3

*Absolute value of mean standardized difference above 10%.

[†] Measures are dichotomous unless otherwise indicated.

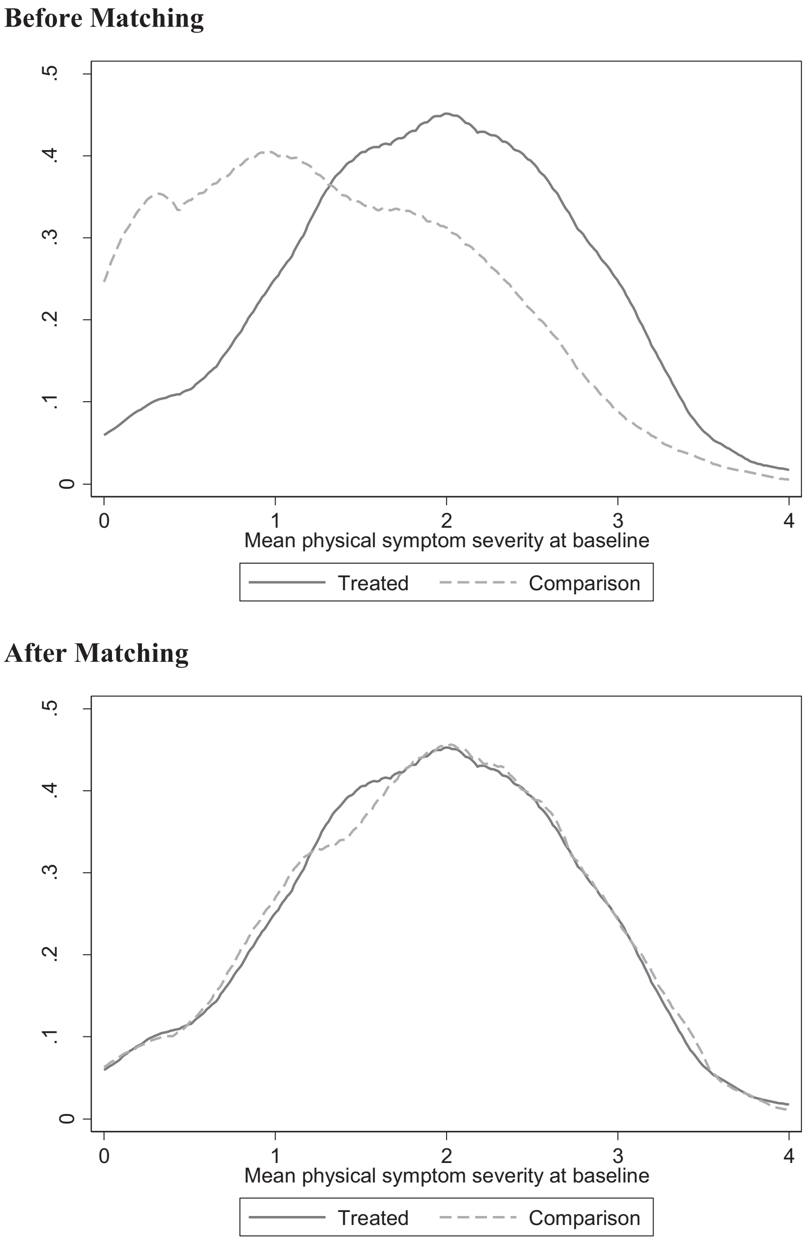
[‡] Elixhauser comorbidity scale (Elixhauser et al. 1998).

[§] Range 0–4; higher numbers indicate worse severity on Condensed Memorial Symptom Assessment Scale (Chang et al. 2004).

[¶] Reference day refers to day of consult for treated patients. The reference day for usual care patients is the day in which they were most similar to treated patients based on symptom severity and sociodemographic characteristics.

ADL, activities of daily living.

Figure 2: Example of Density Plots of Mean Physical Symptom Severity at Baseline before and after Kernel Matching on Empirical Dataset



Before Matching

After Matching

and comparison groups and were reasonably similar (Figure 2). The ratio of variances in the propensity score between the treated and comparison group changed from 1.73 in the unmatched sample to 1.01 in the matched sample. Because kernel weighting led to the best covariate-specific balance across treatment and comparison groups, we chose it as the way to adjust our sample for selection bias.

Estimation and Interpretation of Treatment Effects

Two common treatment effects include the average treatment effect on the treated (ATT) and the average treatment effect for the entire sample (ATE); the choice of treatment effect depends on the investigator's goals. The ATT is the estimated effect of the intervention among treated individuals. The ATE combines the ATT with the estimated treatment effect for untreated individuals.

Interpretation of ATEs and ATTs depends on standard errors. When the propensity score is estimated before the treatment effect, uncertainty from the estimation of the propensity score affects the standard error of the treatment effect estimate. Ignoring this uncertainty leads to conservative standard errors on ATEs, and to either conservative or overly generous standard errors for ATT estimates, depending on the data-generating process (Austin 2009c; Abadie and Imbens 2012). When a propensity score is estimated and the sample is weighted in a separate step by the propensity score, standard errors can be adjusted by bootstrap methods. For matched data, however, bootstrap methods provide unreliable estimates, and standard errors need to be calculated with the Abadie-Imbens (AI) method (Abadie and Imbens 2008, 2012; StataCorp. 2013a).

Example. In our dataset, the ATT is the estimated average effect of palliative care on outcomes for individuals who received palliative care. The ATE is the estimated average effect of palliative care on outcomes for those who did and did not receive palliative care.

Caution. Restricting the sample to the range of common support affects treatment effect estimates. Conclusions about a treatment's effect can only be made for individuals with propensity scores represented in both the treatment and comparison groups. Therefore, the ATE is only an average treatment effect for the sample within the range of common support, not the entire sample.

Conclusion

Propensity scores are one useful tool for health services researchers seeking to account for observed differences between treated and comparison groups in order to isolate the effect of a treatment on a health outcome. It is important to keep in mind that propensity scores cannot adjust for unobserved differences between groups. Researchers considering using propensity scores should carefully consider which variables are included in the propensity score and check for balance before and after matching or weighting.

Acknowledgments

Joint Acknowledgment/Disclosure Statement: This work was funded by NCI/NINR 5R01CA116227 (PI: Diane E. Meier) and partially supported by the National Palliative Care Research Center. Dr. Garrido is supported by Department of Veterans Affairs HSR&D CDA 11-201/CDP 12-255, Dr. Kelley is supported by National Institute on Aging (1K23AG040774-01A1) and the American Federation for Aging Research, Ms. Paris and Ms. Roza are supported by the Doris Duke Charitable Foundation Clinical Research Program, and Dr. Morrison is supported by a Mid-Career Investigator Award in Patient Oriented Research from the National Institute on Aging (K24 AG022345). Dr. Meier directs the Center to Advance Palliative Care, and Dr. Morrison directs the National Center for Palliative Care Research. The authors wish to thank Peter May and two anonymous reviewers for helpful comments on previous drafts of this paper. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

Disclosures: None.

Disclaimers: None.

Note

1. In Stata, **gen** logitpscore = **ln**(*mypscore*/(**1**-*mypscore*))

References

Abadie, A., and G. Imbens. 2008. "On the Failure of the Bootstrap for Matching Estimators." *Econometrica* 76 (6): 1537–58.

——— 2012. "Matching on the Estimated Propensity Score." National Bureau of Economic Research Working Paper No. 15301 [accessed on November 8, 2013]. Available at http://www.nber.org/papers/w15301

Austin, P. C. 2009a. "Balance Diagnostics for Comparing the Distribution of Baseline Covariates between Treatment Groups in Propensity-Score Matched Samples." *Statistics in Medicine* 28: 3083–107.

——— 2009b. "The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates between Treated and Untreated Subjects in Observational Studies." *Medical Decision Making* 29: 661–77.

——— 2009c. "Type I Error Rates, Coverage of Confidence Intervals, and Variance Estimation in Propensity-Score Matched Analyses." *International Journal of Biostatistics* 5 (1): 13.

——— 2011a. "A Tutorial and Case Study in Propensity Score Analysis: An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality." *Multivariate Behavioral Research* 46: 119–51.

——— 2011b. "Optimal Caliper Widths for Propensity-Score Matching When Estimating Differences in Means and Differences in Proportions in Observational Studies." *Pharmaceutical Statistics* 10: 150–61.

Basu, A., D. Polsky, and W. G. Manning. 2008. "Use of Propensity Scores in Non-Linear Response Models: The Case for Health Care Expenditures." National Bureau of Economic Research Working Paper Series. Working Paper 14086 [accessed on June 4, 2013]. Available at http://www.nber.org/papers/w14086

Becker, S., and A. Ichino. 2002. "Estimation of Average Treatment Effects Based on Propensity Scores." *The Stata Journal* 2 (4): 358–77.

Brookhart, M. A., S. Schneewiess, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer. 2006. "Variable Selection for Propensity Score Models." *American Journal of Epidemiology* 163 (12): 1149–56.

Brooks, J. M., and R. L. Ohsfeldt. 2013. "Squeezing the Balloon: Propensity Scores and Unmeasured Covariate Balance." *Health Services Research* 48 (4): 1487–507.

Busso, M., J. DiNardo, and J. McCrary. 2009. "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators [Discussion Paper]." Institute for the Study of Labor Discussion Papers No. 3998 [accessed on July 1, 2013]. Available at http://ftp.iza.org/dp3998.pdf

Caliendo, M., and S. Kopeinig. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching." *Journal of Economic Surveys* 22 (1): 31–72.

Chang, V. T., S. S. Hwang, B. Kasimis, and H. T. Thaler. 2004. "Shorter Symptom Assessment Instruments: The Condensed Memorial Symptom Assessment Scale (CMSAS)." *Cancer Investigation* 22 (4): 526–36.

D'Agostino, R., W. Lang, M. Walkup, T. Morgan, and A. Karter. 2001. "Examining the Impact of Missing Data on Propensity Score Estimation in Determining the

Effectiveness of Self-Monitoring of Blood Glucose (SMBG)." *Health Services & Outcomes Research Methodology* 2: 291–315.

DiNardo, J., and J. L. Tobias. 2001. "Nonparametric Density and Regression Estimation." *Journal of Economic Perspectives* 15 (4): 11–28.

DuGoff, E. H., M. Schuler, and E. Stuart. 2014. "Generalizing Observational Study Results: Applying Propensity Score Methods to Complex Surveys." *Health Services Research* 49 (1): 284–303.

Elixhauser, A., C. Steiner, D. R. Harris, and R. M. Coffey. 1998. "Comorbidity Measures for Use with Administrative Data." *Medical Care* 36 (1): 8–27.

Galdo, J. C., J. Smith, and D. Black. 2008. "Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data." *Annals of Economics and Statistics* 91/92: 189–216.

Garrido, M. M., P. Deb, J. F. Burgess, and J. D. Penrod. 2012. "Choosing Models for Cost Analyses: Issues of Nonlinearity and Endogeneity." *Health Services Research* 47 (6): 2377–97.

Heckman, J. J., H. Ichimura, and P. E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64: 605–54.

Ho, D. E., K. Imai, G. King, and E. A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15: 199–236.

Huang, I., C. Frangakis, F. Dominici, G. B. Diette, and A. W. Wu. 2005. "Application of a Propensity Score Approach for Risk Adjustment in Profiling Multiple Physician Groups on Asthma Care." *Health Services Research* 40 (1): 253–78.

Huber, M., M. Lechner, and C. Wunsch. 2013. "The Performance of Estimators Based on the Propensity Score." *Journal of Econometrics* 175: 1–21.

Imbens, G. W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrica* 87 (3): 706–10.

——— 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." *Review of Economics and Statistics* 86 (1): 4–29.

Jiang, M., and E. M. Foster. 2013. "Duration of Breastfeeding and Childhood Obesity: A Generalized Propensity Score Approach." *Health Services Research* 48 (2): 628–51.

Kruse, G. B., D. Polsky, E. A. Stuart, and R. M. Werner. 2012. "The Impact of Hospital Pay-for-Performance on Hospital and Medicare Costs." *Health Services Research* 47 (6): 2118–36.

Leuven, E., and B. Sianesi. 2003. "PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing, version 4.0.6" [accessed on June 4, 2013]. Available at http://ideas.repec.org/c/boc/bocode/s432001.html

Lunt, M. 2013. "PBALCHK: Checking Covariate Balance" [accessed on May 23, 2013]. Available at http://personalpages.manchester.ac.uk/staff/mark.lunt/propensity.html

Luo, Z., J. C. Gardiner, and C. J. Bradley. 2010. "Applying Propensity Score Methods in Medical Research: Pitfalls and Prospects." *Medical Care Research and Review* 67 (5): 528–54.

Patient-Centered Outcomes Research Institute. 2012. "National Priorities for Research and Research Agenda" [accessed on July 1, 2013]. Available at http://pcori.org

Qu, Y., and I. Lipkovich. 2009. "Propensity Score Estimation with Missing Values Using a Multiple Imputation Missingness Pattern (MIMP) Approach." *Statistics in Medicine* 28 (9): 1402–14.

Rosenbaum, P. R., and D. R. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79 (387): 516–24.

——— 1985. "Constructing a Control Group using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician* 39 (1): 33–8.

Rubin, D. R. 1980. "Bias Reduction Using Mahalanobis Matching." *Biometrica* 36 (2): 293–8.

——— 2001. "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation." *Health Services & Outcomes Research Methodology* 2: 169–88.

Slade, E. P., J. F. McCarthy, M. Valenstein, S. Visnic, and L. B. Dixon. 2013. "Cost Savings from Assertive Community Treatment Services in an Era of Declining Psychiatric Inpatient Use." *Health Services Research* 48 (1): 195–217.

StataCorp. 2013a. *Stata 13 Base Reference Manual.* College Station, TX: Stata Press.

StataCorp. 2013b. *Stata Statistical Software: Release 13.* College Station, TX: StataCorp LP.

Stuart, E. A. 2010. "Matching Methods for Causal Inference: A Review and Look Forward." *Statistical Science* 25 (1): 1–21.

Stuart, E. A., B. K. Lee, and F. P. Leacy. 2013. "Prognostic Score-Based Balance Measures Can Be a Useful Diagnostic for Propensity Scores in Comparative Effectiveness Research." *Journal of Clinical Epidemiology* 66: S84–90.

Wyss, R., C. J. Girman, R. J. LoCasale, M. A. Brookhart, and T. Stürmer. 2013. "Variable Selection for Propensity Score Models when Estimating Treatment Effects on Multiple Outcomes: A Simulation Study." *Pharmacoepidemiology and Drug Safety* 22: 77–85.

Yoon, J., and S. L. Bernell. 2013. "The Role of Adverse Physical Events on the Utilization of Mental Health Services." *Health Services Research* 48 (1): 175–94.

Supporting Information

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.
Data S1: Stata 13 Code to Create and Assess a Propensity Score.