Comparative Effectiveness of Matching Methods for Causal Inference*

Gary King[†] Richard Nielsen[‡] Carter Coberley[§] James E. Pope[¶]
Aaron Wells[‖]

December 9, 2011

Abstract

Matching methods for causal inference selectively prune observations from the data in order to reduce model dependence. They are successful when simultaneously maximizing balance (between the treated and control groups on the pre-treatment covariates) and the number of observations remaining in the data set. However, existing matching methods either fix the matched sample size ex ante and attempt to reduce imbalance as a result of the procedure (e.g., propensity score and Mahalanobis distance matching) or fix imbalance ex ante and attempt to lose as few observations as possible ex post (e.g., coarsened exact matching and calpier-based approaches). As an alternative, we offer a simple graphical approach that addresses both criteria simultaneously and lets the user choose a matching solution from the imbalance-sample size frontier. In the process of applying our approach, we also discover that propensity score matching (PSM) often approximates random matching, both in real applications and in data simulated by the processes that fit PSM theory. Moreover, contrary to conventional wisdom, random matching is not benign: it (and thus often PSM) can degrade inferences relative to not matching at all. Other methods we study do not have these or other problems we describe. However, with our easy-to-use graphical approach, users can focus on choosing a matching solution for a particular application rather than whatever method happened to be used to generate it.

*Our thanks to Stefano Iacus and Giuseppe Porro for always helpful insights, suggestions, and collaboration on related research; this paper could not have been written without them. Thanks also to Alberto Abadie, Seth Hill, Kosuke Imai, John Londregan, Adam Meirowitz, Brandon Stewart, Liz Stuart and participants in the Applied Statistics workshop at Harvard for helpful comments; Healthways and the Institute for Quantitative Social Science at Harvard for research support; and the National Science Foundation for a Graduate Research Fellowship to Richard Nielsen.

[†]Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://GKing.harvard.edu, king@harvard.edu, (617) 500-7570.

[‡]Ph.D. Candidate, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://people.fas.harvard.edu/∼rnielsen, rnielsen@fas.harvard.edu, (857) 998-8039.

[§]Director of Health Research and Outcomes, Healthways, Inc. 701 Cool Springs Blvd, Franklin TN 37067

[¶]Chief Science Officer, Healthways, Inc., 701 Cool Springs Blvd, Franklin TN 37067

[‖]Principal Investigator of Health Outcomes Research, Healthways, Inc., 701 Cool Springs Blvd, Franklin TN 37067

1 Introduction

As is widely recognized, matching methods for causal inference are best applied with an extensive, iterative, and typically manual search across different matching solutions, simultaneously seeking to maximize covariate balance between the treated and control groups and the matched sample size. A measure of the difficulty of this process may be the fact that most applied publications report running only a single matching solution. This is especially problematic because no existing matching method simultaneously optimizes both goals: For example, Mahalanobis Distance Matching (MDM) and Propensity Score Matching (PSM) choose a fixed number of observations ex ante (typically a multiple of the number of treated units) and hope for imbalance reduction as a result of the procdedure. In contrast, Coarsened Exact Matching (CEM) and caliper-based approaches choose a fixed level of imbalance ex ante and hope that the number of observations left as a result of the procedure is sufficiently large.

We attempt to close some of this gap between methodological best practices and the practical use of these methods. To do this, we describe a simple graphical tool to identify the frontier (and evaluate the trade off) of matching solutions that define maximal levels of balance for given matched sample sizes; users can then easily choose a solution that best suits their needs in real applications. We also use this approach to elucidate differences among individual matching methods. We study PSM and MDM (with and without calipers) and CEM. PSM and CEM each represent the most common member of one of the two known classes of matching methods (Rubin, 1976; Iacus, King and Porro, 2011). Other matching methods can easily be incorporated in our framework too.

In the process of developing this graphical tool for applied researchers, we are also led to several new methodological findings. For example, we discover a serious problem with PSM (and especially PSM with calipers) that causes it to approximate random matching in common situations. Moreover, and contrary to conventional wisdom, random matching is not benign; on average, it increases imbalance (and variance) compared to not matching. As we show, such results call into question the automatic use of PSM for matching without comparison with other methods; the common practice of calipering PSM matches worse

than 1/4 of standard deviation; using PSM to adjust experimental results; including all available covariates in PSM models; and some other commonly recommended practices. To understand this problem, we analyze data simulated to fit propensity score theoretical requirements, and we find the problem again. With these simulations, and a simple 12-observation example we develop, we illuminate precisely why the problem occurs and what to do about it. Even with these findings, PSM can be a useful procedure, but it needs to be used comparatively; our graphical procedure can help.

The comparative effectiveness approach to matching methodology discussed here offers an easy way to implement best practices, enabling applied researchers to choose matching solutions that improve inferences in their data rather than having to decide on a theoretical basis which matching methods may or may not work best in general.

2 Matching Methods

Here we summarize three matching methods, as they are commonly used in practice. We begin with notation, the quantities of interest and estimation, and then discuss how to choose among matching methods.

Notation For unit i ($i = 1, \ldots, n$), let T_i denote a treatment variable coded 1 for units from the treated group and 0 for units from the control group. For example, $T_i = 1$ may indicate that person i is given a particular medicine and $T_i = 0$ means that person i is given a different medicine (or a placebo). Let $Y_i(t)$ (for $t = 0, 1$) be the value the outcome variable would take if $T_i = t$, a so-called "potential outcome". By definition, for each i, either $Y_i(1)$ or $Y_i(0)$ is observed, but never both. This means we observe $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$. Finally, we denote a vector of available pre-treatment control variables X_i.

Quantities of Interest and Estimation Denote the treatment effect of T on Y for unit i as $\text{TE}_i = Y_i(1) - Y_i(0)$. Now consider an observation from the treated group, so $Y_i(1) = Y_i$ is observed, and so $Y_i(0)$ is unobserved. The simplest version of matching is to estimate TE_i by replacing the unobserved $Y_i(0)$ with an observed unit (or units) j from the control group such that it is matched to observation i either as $X_i = X_j$ (which is known as "exact matching") or, as is usually necessary, $X_i \approx X_j$. (Matching methods differ primarily by

how they define approximate matching.) Unmatched observations are pruned from the data set before further analyses.

The immediate goal then is to maximize both balance, the similarity between the multivariate distributions of the treated and control units, and the size of the matched data set. Any remaining imbalance must be dealt with by statistical modeling assumptions. The primary advantage of matching is that it greatly reduces the dependence of our conclusions on these assumptions (Ho et al., 2007).

In most studies, we do not seek to estimate TE_i for each i; we instead estimate averages of it over relevant subsets of units. One such quantity of interest is the *sample average treatment effect on the treated*, $\text{SATT} = \frac{1}{n_T} \sum_{i \in \{T=1\}} \text{TE}_i$, which is the treatment effect averaged over all (n_T) treated units. However, since most real data sets contain some observations without good matches, perhaps because common support between the treated and control groups do not fully coincide, best practice has been to compute a causal effect among only those treated observations for which good matches exist. We designate this as the *feasible sample average treatment effect on the treated* or FSATT.

This decision is somewhat unusual in that FSATT is defined as part of the estimation process, but in fact it follows the usual practice in observational data analysis of collecting data and making inferences only where it is possible to learn something. The difference is that the methodology makes a contribution to ascertaining which quantities of interest can be estimated (e.g., Crump et al., 2009; Iacus, King and Porro, 2011); one must only be careful to characterize the units being used to define the new estimand. As Rubin (2010, p.1993) puts it, "In many cases, this search for balance will reveal that there are members of each treatment arm who are so unlike any member of the other treatment arm that they cannot serve as points of comparison for the two treatments. This is often the rule rather than the exception, and then such units must be discarded.... Discarding such units is the correct choice: A general answer whose estimated precision is high, but whose validity rests on unwarranted and unstated assumptions, is worse than a less precise but plausible answer to a more restricted question." In the analyses below, we use FSATT, and so let the quantity of interest change, but everything described below can be restricted to a fixed

quantity of interest, decided ex ante, which will of course will sometimes be of interest.

Matching Procedures *Mahalanobis distance matching* (MDM) and *propensity score matching* (PSM) are built on specific notions of distance between observations of pre-treatment covariates. MDM measures the distance between the two observations X_i and X_j with the Mahalanobis distance, $\text{M}(X_i, X_j) = \sqrt{(X_i - X_j)'S^{-1}(X_i - X_j)}$, where S is the sample covariance matrix of X. In PSM, we first collapse the vectors to a scalar "propensity score," which is the probability that an observation receives treatment given the covariates, usually estimated by a logistic regression, $\pi_i \equiv \Pr(T_i = 1|X) = 1/(1 + e^{X_i\beta})$; then, the distance between observations with vectors X_i and X_j is the simple scalar difference between the two estimates $\hat{\pi}_i - \hat{\pi}_j$ (or sometimes $X_i\beta - X_j\beta$).

The most common implementation of each approach is to apply one-to-one nearest neighbor greedy matching without replacement (Austin, 2009, p.173). This procedure matches each treated unit in some arbitrary sequence to the nearest control unit, using that method's chosen distance metric. Some procedure is then applied to remove treated units that are unreasonably distant from (or outside the common support of) the control units to which they were matched. The most common such procedure is *calipers*, which are chosen cutoffs for the maximum distance allowed (Stuart and Rubin, 2007; Rosenbaum and Rubin, 1985).

Coarsened Exact Matching (CEM) works somewhat differently. First, each variable is temporarily coarsened as much as is reasonable. For example, years of education might be coarsened for some purposes into grade school, high school, college, and post-graduate. Second, units with the same values for all the coarsened variables are placed in a single stratum. And finally, for further analyses, control units within each stratum are weighted to equal the number of treated units in that stratum. Strata without at least one treated and one control unit are thereby weighted at zero, and thus pruned from the data set. (The weights for each treated unit is 1; the weights for each control unit equals the number of treated units in its stratum divided by the number of control units in the same stratum, normalized so that the sum of the weights equals the total matched sample size.) The unpruned units with the original uncoarsened values of their variables are passed on to

the analysis stage. Unreasonably bad matches are dropped as an integral part of the CEM procedure and so a second step, such as calipers, is not required. See Iacus, King and Porro (2011, 2012).

Many refinements of these and other matching procedures have been proposed, only a few of which are widely used in applied work (for reviews, see Ho et al., 2007; Stuart, 2010).

3 Evaluating Matching Methods

To evaluate a matching method, we confront the same bias-variance trade-off as exists with most statistical methods. However, two issues can prevent one from optimizing on this scale directly. First, matching can be used to define the quantity of interest, FSATT, in which case a particular point on the bias-variance frontier is not known but instead merely becomes possible to achieve. Whether any level of bias or variance is actually achieved depends on the particular analytic method chosen to apply to the matched data.

Second, best practice in matching involves avoiding selection bias by intentionally ignoring the outcome variable while matching (Rubin, 2008*b*), the consequence of which is that we give up the ability to control either bias or variance directly. Thus, instead of bias, we focus on reducing the closely related quantity, imbalance, the difference between the multivariate empirical densities of the treated and control units (for the specific mathematical relationship between the two, see Imai, King and Stuart, 2008). Similarly, the variance of the causal effect estimator can be reduced when heterogeneous observations are pruned by matching, but a too small matched sample size can inflate the variance. Thus, in matching, the bias-variance trade off is affected through the crucial trade off between the degree of imbalance and the size of the matched sample.

We repeat all analyses with each of three types of imbalance metrics. The most commonly used metric in empirical evaluations (and theoretical analyses) of PSM is the difference in means between the treated and control groups for variable j: $d(j) = \bar{X}^{(j)} - \tilde{X}^{(j)}$, where $\bar{X}^{(j)}$ is the mean of variable j in treated group and $\tilde{X}^{(j)}$ is the corresponding mean in the control group. We examine each variable separately and then (rescaling if necessary) report the average of the absolute values: $D = \frac{1}{J} \sum_{j=1}^{J} |d(j)|$.

Second is the Mahalanobis matching discrepancy, defined as the Mahalanobis distance between each unit i and the closest unit in opposite group, averaged over all units: $M = \frac{1}{N}\sum_{i=1}^{n} M(X_i, X_{j(i)})$, where $X_{j(i)} = \min_{j\in\{1-T_i\}} M(X_i, X_j)$, where $\{1 - T_i\}$ is the set of units in the group that does not contain i, and N is the size of the observed or matched data set being evaluated.

Our final imbalance metric is L_1, which directly measures the difference between the multivariate histogram of the treated group and the multivariate histogram of the control group (Iacus, King and Porro, 2011). With a chosen set of bin sizes H, we form the multivariate histogram of the treated units and separately the multivariate histogram of the control units. Let $f_{\ell_1 \cdots \ell_k}$ be the relative empirical frequency of treated units in bin with coordinates $\ell_1 \cdots \ell_k$, and similarly for $g_{\ell_1 \cdots \ell_k}$ among control units. Then let

$$L_1(H) = \frac{1}{2} \sum_{(\ell_1 \cdots \ell_k) \in H} |f_{\ell_1 \cdots \ell_k} - g_{\ell_1 \cdots \ell_k}|. \tag{1}$$

The final L_1 imbalance measure then arises when fixing the bin sizes H to that which produce the median value of $L_1(H)$ from (approximately) all possible bin sizes in the original unmatched data. The typically numerous empty cells of each of the histograms do not affect L_1, and so the summation in (1) has at most only n nonzero terms. The relative frequencies also control for what may be different sample sizes for the treated and control groups. Weights may be used when matching methods produce weights; otherwise all observations are weighted equally. A value of $L_1 = 0$ indicates identical treatment and control distributions, and $L_1 = 1$ indicates complete imbalance and no overlap between the densities.

We might think that we could choose one of these metrics, and optimize it directly to choose a matching method. However, existing methods do not optimize with respect imbalance and matched sample size simultaneously. To avoid the resulting difficulties, statisticians have recommended that researchers conduct an iterative search for a good matching solution by evaluating the matched sample size and more general imbalance criteria (e.g., Austin, 2008; Caliendo and Kopeinig, 2008; Imbens and Rubin, 2009; Rosenbaum, Ross and Silber, 2007; Stuart, 2008). For example, Rosenbaum and Rubin (1984) detail their "gradual refinement" of an initial model by including and excluding covari-

ates until they obtain a final model with 45 covariates, "including 7 interaction degrees of freedom and 1 quadratic term". And Ho et al. (2007, p.216) write "one should try as many matching solutions as possible and choose the one with the best balance." This advice is unfortunately not often followed in applied research. The graphical approach we offer in the next section is meant to be a convenient way of applying this advice by giving the researcher the big picture in terms of viewing and choosing from a range of possible matching solutions.

4 Comparing Matching Solutions

We now introduce the *space graph*, the goal of which is to compare matching solutions, without regard to the method that produced them, and to approximate the imbalance-matched sample size *frontier*. The frontier is the set of matching solutions for which no other solution has lower imbalance for a given sample size or larger sample size for given imbalance. Matching solutions not on the frontier should not be used in applications (at least not without further justification) since they are strictly dominated in terms of imbalance and number of observations by those on the frontier.

Among the solutions along the frontier, one can reasonably choose one based on the matching version of the bias-variance trade-off: For example, in a large data set, we can afford to prune a large number of observations in return for lower imbalance (since our confidence intervals will still be fairly narrow), but in smaller data sets or when an application requires especially narrow confidence intervals we might prefer a solution on the frontier with fewer pruned observations at the cost of having to build a model to cope with imbalance left after matching. If we choose to let matching help define FSATT, rather than fixing the quantity of interest as SATT, the choice of a solution on the frontier also helps us choose the estimand. Either way, the the choice of a matching solution from the imbalance-matched sample size frontier is analogous to, but of course different than, the process of choosing an estimator that reflects the bias-variance trade off.

The space graphs we show in this section attempt to identify the overall frontier, and also a frontier constrained within the solutions given by each specific matching method. Only the overall frontier is of importance for applications; the method-specific frontier is

a methodological tool for understanding the contribution of individual matching methods.

4.1 Space Graph Construction

We construct a space graph to parallel a bias-variance plot by displaying matching solutions with a measure of imbalance on the vertical axis and the matched sample size (starting from the full sample on the left and dropping to the right) on the horizontal axis. See, for example, the axes for the three space graphs in Figure 1 (about which more below).

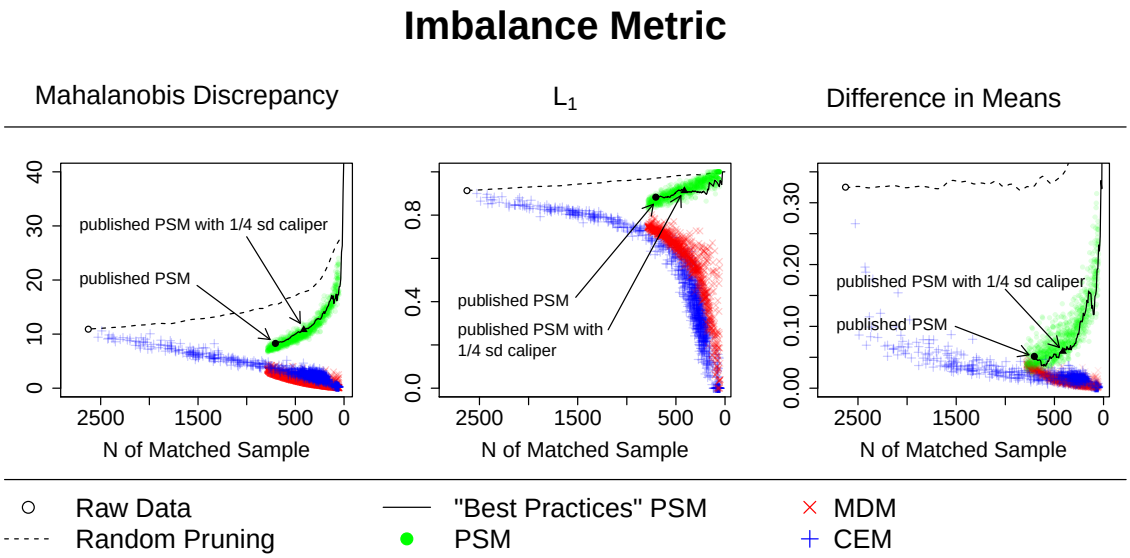

Figure 1: Space graphs based on data from Nielsen et al. (2011). Each point represents one matching solution plotted horizontally according to the matched sample size and vertically by one of three imbalance metrics. Solutions closer to the bottom left of each plot are better than others. The observed frontier for a method is the set of points such that no other point from that method appears to its lower left.

A point in a space graph represents one matching solution. Solutions appearing closer to the bottom left of the figure, that is with low imbalance and large sample size, are preferable. Of course, no uniquely best solution is likely to exist, and so instead, we attempt to identify the frontier, the set of points such that no other point appears to its lower left. The simplest, fastest, and most reliable algorithm we have found to identify the frontier is to begin with a large set of matching solutions generated by existing matching methods and to identify those for which no solution can be found to their lower left. (Of course, this observed frontier may be improved if even better matching solutions can be

found, or if a method can be developed to reliably find the optimal frontier.)

To fix ideas, we construct in Figure 1 space graphs for data analyzed in Nielsen et al. (2011) for the purpose of estimating the effect of a sudden decrease in international foreign aid on the probability of violent conflict in developing countries. The data set includes 18 covariates representing national levels of democracy, wealth, population, ethnic and religious fractionalization, and prior upheaval and violence. The open circle at the left of each graph represents the original data prior to pruning any observations through matching. Other points on the graph represent those we randomly generated from each matching method. For PSM and MDM, we randomly sample 1,000 specifications from the set of all variables, their squares, and two-variable interactions. Thus, we added to the 18 raw pre-treatment covariates, 18 squared terms, and $\binom{18}{2} = 153$ one-way interactions. We then sample subsets of these 189 variables. In a separate simulation, we restricted variable selection to enter main effects for all variables included in the interactions selected, and to always include all 18 main effects, but these had no substantive effect on the results. (And of course, eliminating solutions, even if done based on some version of expert judgment, can never improve the observed frontier.) We also tried adding three-way interactions; this also led to similar conclusions, but in this case many PSM models do not converge and MDM variance matrices were not invertable.

We then matched observations under MDM and PSM using 1-to-1, 1-to-2, and 1-to-3 treatment-to-control matching. For each match, we also choose a caliper c at random from a uniform distribution on the integers between 0 and the number of treated units. We then prune the c worst matches, where "worst" is defined by the matching method's distance metric. Most implementations of calipers instead set a single arbitrary threshold of imbalance, such as one-quarter of a standard deviation of the propensity score, above which all matches are excluded. Our analysis thus includes all possible thresholds, including this most common approach, which we also discuss below. Limiting matching solutions in the space graph to only some caliper thresholds can never improve the observed frontier. (In the following section, we separate out the effects of calipering, which in PSM and MDM effects n, from variable selection, which does not.)

For CEM, we create matching solutions by changing the coarsenings (and where a null coarsening is equivalent to excluding the covariate). To search over a large number of CEM solutions, we sample from the set of possible coarsenings equally spaced between minimum and maximum levels. Again, using only a subset of these that are "plausible" by some expert definition can never improve the observed frontier.

We choose to display all the matching solutions, and to identify the frontier by visual inspection, rather than withholding all but the frontier points. Solutions from different methods are represented with different symbols and colors: PSM (green disk), MDM (red cross), CEM (blue plus), and as a baseline the average for each given sample size of random matching solutions (black dashed line). (Space graphs were introduced by Iacus, King and Porro (2009), who also offer software to compute and plot them.)

4.2 Results

We describe the results in Figure 1 in several steps.

The Overall Frontier The conclusion in the literature that matching solutions should be compared prior to use is clearly confirmed by the great diversity in the quality of solutions represented here. It would be easy to inadvertently choose a solution far from the frontier by failing to compare different solutions.

In all three graphs, the overall frontier (the set of symbols such that none appears to the lower left of each) is usually defined by CEM and occasionally by MDM solutions. The CEM frontier (the subset of blue plus symbols such that no other blue plus appears to its lower left) is always either equal or close to the overall frontier. PSM solutions do not usually approach the overall frontier. Applied researchers will of course only need to identify the overall frontier, the method giving rise to it being unimportant for applications. In contrast, the methods-specific frontiers may be relevant for those developing specific matching methods.

Of course, the specific set of matching methods we consider, and the definition of each individual method, are not sacrosanct: if we added matching with replacement to MDM, replaced greedy with optimal matching under MDM or PSM, applied MDM within CEM strata, or adjusted in many other ways, we might well be able to generate a better overall

frontier than CEM. And, we still leave the open question as to whether it is possible to invent an optimization algorithm to identify the theoretical (i.e., best possible) frontier directly.

Method-Specific Frontiers The overall, CEM-specific, and MDM-specific frontiers in all three graphs in Figure 1 display the expected trade off between imbalance and sample size: As we prune more observations imbalance drops. For PSM, the result is very different: as PSM and its calipers prune more observations, imbalance increases (see the green dots moving up and to the right). We refer to the troubling behavior as the *propensity score paradox*, and seek to explain it in Section 5.

Consequences of Random Matching As a baseline of comparison, we offer in each graph in Figure 1 a dashed line representing the average for each sample size of a large number of matching solutions constructed by randomly choosing obserations to prune. That this line is increasing from left to right indicates that the more observations that are randomly deleted the larger the average imbalance becomes. This may seem counterintuitive, and to our knowledge has not before been noted in the literature (Imai, King and Stuart, 2008, p.495), but is in fact to be expected. Although no one would intentionally match randomly, we show below that, in many situations, PSM is at times doing something similar, and so understanding why random matching increases imbalance will prove important.

For intuition, consider four observations perfectly matched on sex with two males and two females: MM, and FF. Randomly dropping two of the four observations will leave us with one matched pair among MM, FF, MF, or FM, with equal probability. This means that with 1/2 probability the remaining data set will be balanced and with 1/2 probability it will be completely imbalanced; on average, then, randomly dropping an observation will increase imbalance from its starting point of $L_1 = 0$ to $L_1 = 1/2$. Similarly, although the expected difference in means does not change as observations are randomly pruned, the variance of the difference increases, as of course does the average absolute difference in means. The same intuition applies to randomly deleting observations in larger data sets where, even if the empirical distribution is unchanged, the actual units are more sparsely

arranged, and thus any two points are further apart on average (think of the consequence of removing every other point in a two-dimensional grid, or the reason density estimation becomes easier with a larger n).

What About Expert Specification Selection? To the these basic space graphs, we add the solution chosen by Nielsen et al. (2011). The authors followed best practices by selecting variables for substantive reasons, considering many individual PSM matching solutions, and then choosing the best solution based on imbalance comparisons (using the difference in means metric). This solution appears as a black dot in each graph (noted "published PSM"). We then calipered off the worst match (based on PSM's distance metric), the next worst match, etc., each time generating a different solution; the set of these solutions is traced out by the black line starting at the black dot.

From this analysis we draw three conclusions. First, the general upward trajectory of the black line in all three graphs demonstrates that the propensity score paradox even afflicts results from this expert-selected PSM solution.

Second, the figure also represents the matching solution that begins with a published PSM match and then calipers off all matches larger than 1/4 of a standard deviation, as widely suggested in the PSM literature (since Rosenbaum and Rubin, 1985). As can be seen, this advice is harmful here and is dominated by the original PSM solution (and numerous others). This result invalidates the 1/4 caliper advice in this data set, and the paradox questions whether this common practice is advisable in other applications as well.

Finally, the published PSM solution is not at the observed PSM-specific frontier, and thus is strictly dominated by other PSM solutions, according to any of the three imbalance metrics (including the difference in means metric used by Nielsen et al. 2011). Although best practice should be to select a matching solution from the frontier regardless of the matching method that created it, even researchers restricting their search to PSM solutions will likely find better matching solutions using a space graph than they could by hand.

What About Other Data Sets? Space graphs for four additional data sets (described in Appendix A) appear in Figure 2, using the same symbols and color schemes. Instead of expert-selected results for these data, we take advantage of a "benchmark" procedure pro-

posed by Imbens and Rubin (2009) to automate previous best practices in the use of PSM. The procedure runs PSM, checks imbalance, adjusts the specification, and automatically reruns until convergence. To each of the graphs in Figure 2, we add this solution (the leftmost point on the solid black line), and then, as before, caliper off the worst match, one observation at a time (traced out with the rest of the black line).

The conclusions from these four data sets generally confirm those in Figure 1. First, and most importantly, the space graph is a useful tool to identify the overall frontier from which applied researchers can choose matching solutions. Second, the overall frontier is usually defined by CEM and occasionally by MDM; the CEM-specific frontier is always near the overall frontier; and PSM solutions do not usually approach the overall frontier. Third, the "best practices" line is sometimes better than the mass of PSM points, sometimes worse, and sometimes about the same, but across four data sets and three imbalance metrics, the propensity score paradox is in clear evidence by the line often or always heading upwards (indicating more imbalance) as more observations are dropped. Fourth, for the first three data sets, the PSM-specific frontier heads upward, displaying the paradox.

Interestingly, the Lalonde data, which many methodology articles use for their sole data analysis example, exhibit the paradox the least among the four data sets across both figures. This may account for why the paradox was not previously identified. Neither CEM nor MDM reveal this paradoxical behavior for any of the four data sets. Random matching (dashed lines) is the only other approach that displays similar behavior to PSM.

How Widespread is the Paradox? To get a sense of how widespread these patterns are in other social science data sets, we advertised on line to help researchers doing matching analyses in return for being able to access their data before publication (and promising not to distribute the data or scoop them for their substantive results). For almost all of the 27 data sets we received, the patterns we found above were repeated, and so we can be reasonably confident that the results from the five data sets presented above are more general and apply to a large number of data sets now being analyzed by social scientists around the world.

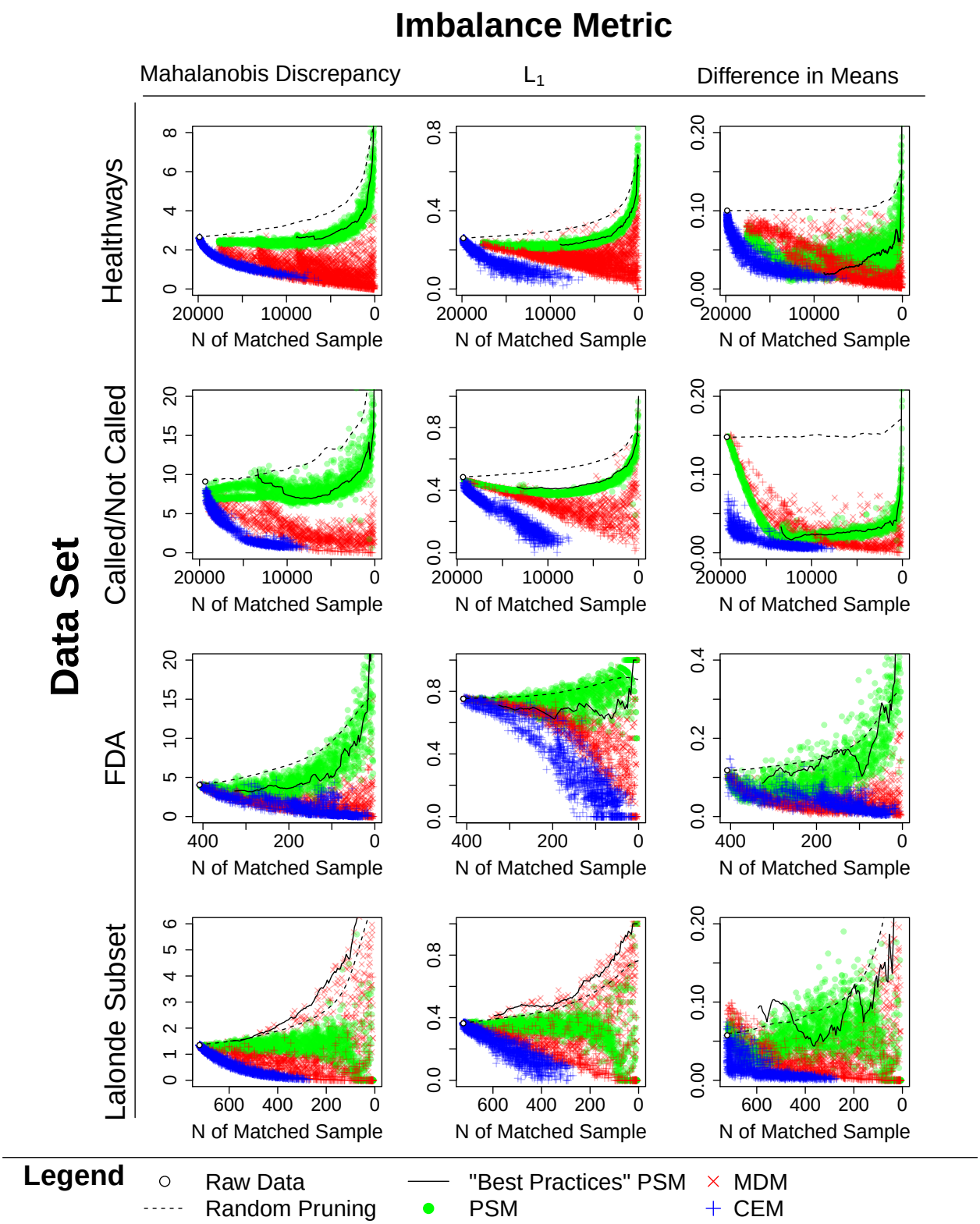

Figure 2: Space graphs for four data sets described in Appendix A.

5 Explaining the Propensity Score Paradox

As Section 4 shows, PSM evaluates some matched pairs as highly imbalanced, even though removing them increases overall imbalance. We now explain why this occurs. To begin, if the goal of PSM were to match only the means in the treated and control groups, then this pattern could possibly be innocuous: That is, a large positive imbalance in one matched pair might counterbalance a large negative imbalance in another, and so removing only one of the two would be unwise. However, the goal of matching in general, and PSM in particular, is to balance the entire distribution of the treated and control groups (Stuart, 2010, p.13), and so we must look elsewhere for an explanation.

We proceed by first ruling out the possibility that the problem is unique to the many data sets we analyze above. We do this by using two data generation processes, either one of which is required for PSM to have attractive theoretical properties (Rubin and Stuart, 2006).

Data Generation Processes We set aside the process of pruning from the original data set to a matched set (i.e., with observations equal to twice the number of treated units). Instead, we zero in on the qualities of the resulting matched set and PSM imbalance metric. We thus simulate in one of two ways: First, following the procedures of Gu and Rosenbaum (1993), for j covariates ($j = 1, 2, 3$), we repeat the following 50 times and average. We first draw a data set with 250 treated and 250 control units, which is equivalent for our purposes to generating a larger pool of controls and pruning down to 250. The data generation process is multivariate normal (which meets the requirements of the equal percent bias reducing class of methods, which includes PSM) with variances of 1 and covariances of 0.2. The control group always has a mean vector of (0,0,0). We create data sets with high, medium, and low levels of balance by setting the treated group means to (0.1,0.1,0.1), (1,1,1), and (2,2,2), respectively.

We estimate the propensity score (specified with all the main effects in a logit model, as usual) for each of the 50 randomly selected data sets. Then, for each, we remove the single worst match — i.e., for which the scalar difference in propensity scores is greatest — and compute an imbalance metric. (If the paradox found above does not appear in

these data, then the matched data set with the worst match removed would have lower imbalance.) We then repeat this for each of the 50 data sets, average, and put a point on the graph. We then iteratively remove the next worst match and so on, each time recalculating L_1, and adding a point to the graph. We also conduct the analogous procedures for CEM and MDM, using the identical randomly generated data sets as for PSM.

For CEM, we begin with the loosest possible coarsening, so that all data fall in a single stratum and no observations are pruned. We then randomly select a variable, add one cutpoint to its coarsening (always arranging the cutpoints so they divide the space between the minimum and maximum values into equal sized bins), and compute L_1. Additional cutpoints eventually lead to more observations being dropped; results are then averaged over the 50 randomly generated data sets.

We also conducted analyses with two other data generation processes. For one, we repeat the above simulation but using formal specification searches to define the PSM model. For the other, we used the same simulated covariates as above, but we defined the treatment assignment vector using a true propensity score equation. We then estimated a propensity score model using the correct specification (from the true propensity score) and iteratively apply calipers as above. As we find virtually identical results to those for the first data generation process, we only present results for it.

Results Figure 3 gives the results for the methods in three columns (PSM, CEM, and MDM) and different numbers of covariates (1,2,3) in separate rows. Each of the nine graphs in the figure give results for data sets with low (solid line), medium (dashed line), and high (dotted line) levels of imbalance between the treated and control units. For graphical clarity, unlike Figure 2, individual matching solutions do not appear and instead we average over the simulations in each graph for a given matched sample size and level of imbalance. The figure presents results with L_1 (the other two imbalance metrics gives essentially the same results and so we do not present them). The PSM paradox is revealed whenever one of the lines increase (i.e., with imbalance increasing as the number of observations drops).

As the figure shows, CEM and MDM do not suffer from the paradox: all the lines in

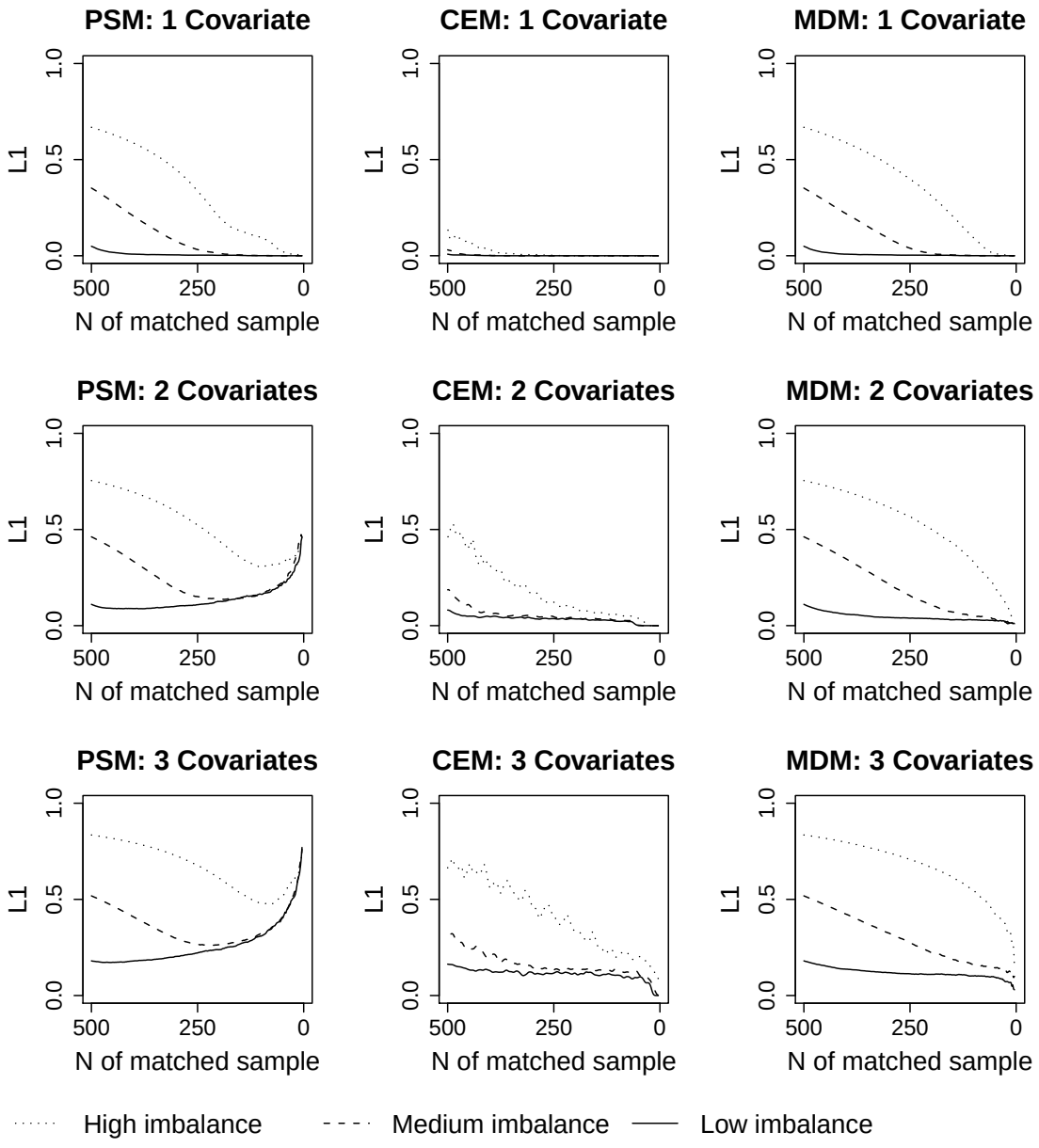

Figure 3: For PSM, CEM, and MDM in columns, and 1, 2, and 3 covariates in rows, these graphs give average values from a simulation with low (solid), medium (dashed), and high (dotted) levels of imbalance between the treated and control groups. The paradox is revealed for portions of lines that systematically increase. This can be seen for PSM but not for CEM and MDM. (The lines represent averages over 50 simulations from individual matching solutions in each graph for a given matched sample size and level of balance.)

all the graphs in the last two columns have imbalance dropping as more observations are pruned, just as we would want to be the case. There is also no evidence of the paradox with PSM applied to a single covariate, which makes sense since the distance metric is based on a simple monotonic (logit) transformation of the scalar X.

For PSM, two problematic patterns emerge. First, the paradox appears with PSM as soon as some dimension reduction begins (in the bottom two graphs in the left column). With two covariates, the lines increase as n drops, and the problem intensifies with three covariates (and the paradox continues to intensify with more covariates). PSM, but not MDM or CEM, fails due to the curse of dimensionality: the scalar propensity score $\hat{\pi}_i$ is an increasingly worse summary of a vector of covariate values, X_i, as the number of elements in the vector increase (see Brookhart et al., 2006).

Second, the PSM graphs for 2 and 3 covariates also reveal the fact that propensity score matching performs systematically worse in better balanced data. Data with better balance means that the propensity score logit discriminates less well between treateds and controls. At the extreme inability to predict, where every observation has a propensity score of 0.5 (as in a classic randomized experiment), the propensity score is equivalent to random matching.

Explanation What explains the fact that the PSM paradox occurs and intensifies in higher dimensions and with better balanced data? Why does it not occur with MDM or CEM? Our answer is that when the predictive capacity of the logit worsens, the scalar propensity score metric approximates random matching which, as we have shown, increases imbalance. Three factors lead to this conclusion.

First, under MDM and CEM, the chosen measure of match quality is computed *directly* from the multivariate data sets of treated and control units. In contrast, PSM's *two step* procedure first attempts to reduce the multivariate data set to a scalar (propensity score) measure, and only afterward computes the measure of matching quality. The problem is that a great deal of information can be lost in the (unnecessary) first step.

Second, the methods pose different estimation demands: whereas CEM does not require estimation, and MDM requires only the computation of the sample covariance ma-

trix, PSM involves a full scale (usually logistic) estimation step. When the logit's classification performance is poor, such as in well balanced data, PSM can approximate random matching.

Illustration To clarify this issue, we first show how in balanced data PSM, but not MDM and CEM, is highly sensitive to trivial changes in the covariates, often producing nonsensical results. In the left panel of Figure 4, we generate data with 12 observations and two covariates. The simulations are designed to be extreme so that the problem becomes obvious. The graph plots one covariate by the other. The data are well balanced between treated (black disks) and control (open circles) units, which means that each treated unit is relatively close (even if not identical) to at least one control unit in the covariate space. From these initial data, we generate 10 data sets, where we add to each observation a tiny amount of random error drawn from a normal distribution with mean zero and variance 0.001. This error is so small relative to the scale of the covariates that the new points are visually indistinguishable from the original points (in fact, the graph plots all 10 sets of 12 points on top of one another, but it only appears that one set is there). Next, we run CEM (with automated coarsening) and MDM; in both cases, as we would expect, the treated units are matched to the nearest control in every one of the 10 data sets (as portrayed by the pair of points in each red dashed circle). (The light grid denotes the CEM coarsenings, and so treateds and controls within the same rectangles are matched; MDM matches in this example based on distances close to Euclidean.)

However, when we run PSM on each of the 10 data sets generated for Figure 4, the four treated units are each matched to different control units (as portrayed by the maze of green lines connecting the black disks to different open circles). PSM is approximating random matching in this situation because it is unable to distinguish treated and control units except for the tiny perturbations. From the perspective of PSM, this may *seem* to make sense because covariates that do not discriminate between treated and control status are not confounders and so which they match to is irrelevant. However, and crucially, matching on covariates that have no effect on treatment is also equivalent to random matching, which increases imbalance on average. Thus, PSM will approximate random

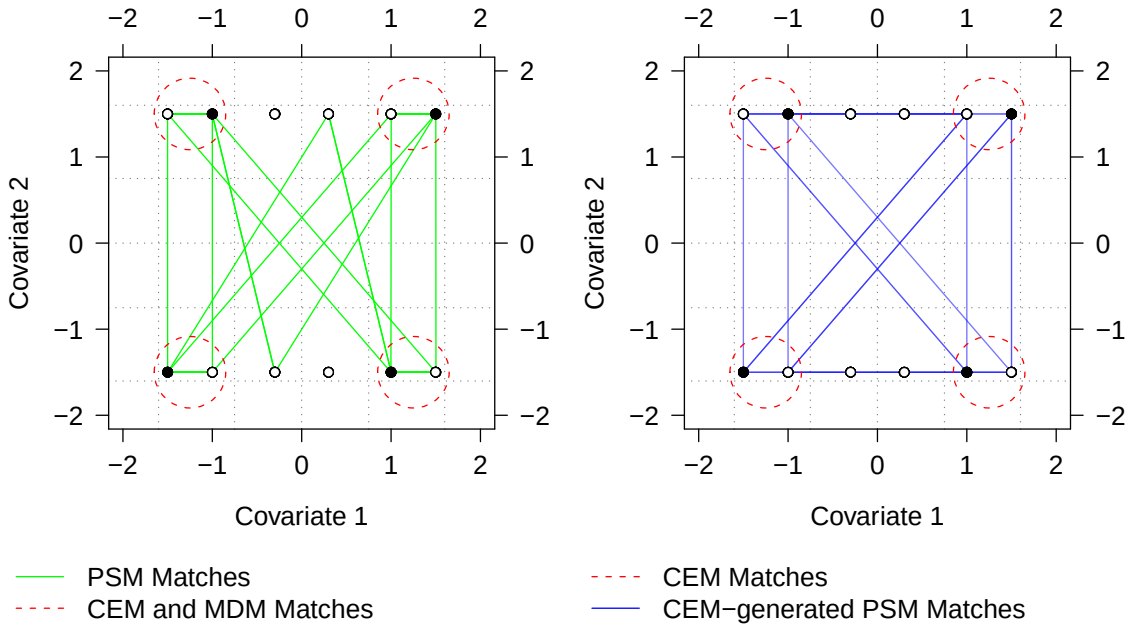

Figure 4: Ten data sets (that differ from each other by tiny amounts of random error, imperceptibly different from each other visually) with 4 treated units (black disks) and 8 control units (open circles). CEM and MDM match the closest control units to each treated (red dashed circles). The two-step procedures match different control units for each data set, as can be seen for PSM (green lines, left panel) and PS-CEM (blue lines, right panel). (The four open circles in the middle of the right panel are never matched; lines are passing through them on the way to show how other points are matched.)

matching and result in degraded inferences relative to not matching at all when the data are reasonably well balanced, and the problem is exacerbated when using PSM with large numbers of covariates.

Finally, we illustrate how the paradox results from PSM's two-step procedure. We do this by developing a (similarly unnecessary and ill-advised) two-step "propensity score CEM" (PS-CEM) algorithm: first we use CEM to compute a nonparametric estimate of the propensity score (i.e., the proportion of treated units within each coarsened stratum; see Iacus, King and Porro 2011) and, second, we match on this estimate. The right panel in Figure 4 is constructed the same way as the left panel except that instead of the green lines representing propensity score matches, we now plot blue lines for the PS-CEM matches. The result is almost as bad as PSM. The blue lines in the right panel show how in the different (but highly similar) data sets, the two-step PS-CEM procedure matches control

units (circles) close to and also distant from treated (closed disks) units. Thus, collapsing to a scalar and then matching based on a scalar distance is causing the PSM paradox.

6 Concluding Remarks

Our results suggest at least seven implications for practice. First, once the actual matched sample from any particular matching solution is known, the matching method that produced it is unimportant. Matching can be performed by any method, or even by hand, so long as a solution on the balance-sample size frontier is found. The key to the productive use of modern matching methods is that many matching solutions be compared, and a final solution chosen from the diversity of possibilities on the frontier. The space graph offers an easy method for doing this and thereby following recognized best practices.

Second, among the methods we studied, results from CEM, and occasionally MDM, usually define the frontier. This was the case in numerous real data sets and simulations. It is true with respect to all three imbalance metrics we evaluated. Of course, nothing guarantees this result will hold in future data sets or that CEM dominates other matching methods not examined in our analysis. Likely, one could improve on the results here by adjusting or combining features of the three methods used here or others. As such, users are well advised to use a space graph to choose a matching solution from a large number of comparisons.

Third, we discovered a serious problem that causes PSM, as commonly used, to act like random matching and degrade inferences, sometimes worse than the original data. CEM and MDM do not suffer from this problem. Our conclusion is not that researchers should necessarily discard PSM, but that it should not be used without comparing its results with other methods, such as via a space graph.

Fourth, the recommendation sometimes given in the literature that PSM should be used to adjust the data from (inadequately blocked) randomized experiments (Rubin, 2008*a*) is ill advised. In well balanced data, such as from experiments (with all propensity scores near 0.5), pruning observations by PSM will often degrade causal inferences relative to the raw unmatched data. Adjusting via some other matching method may still be advantageous in these situations. In addition, it may be possible in some instances

to use a space graph to chose a matching solution generated by PSM which ensures that balance is improved; the results here suggest that using other methods are more likely to be successful, but comparing more solutions to better identify the frontier is always worthwhile.

Fifth, the long-standing advice in the literature to routinely use a 1/4 standard deviation caliper with PSM to remove the worst matches should not be used without a space graph or some other comparative approach to evaluation. This procedure will often increase imbalance even while shedding observations; in some cases, the solution will be worse than not matching at all. This procedure may sometimes be advisable with MDM (it is probably unnecessary with CEM).

Sixth, notwithstanding Pearl's (2000) argument about how controlling for certain "irrelevant" variables can induce bias, most applied researchers claim to have "little or no reason to avoid adjustment for [every] true covariate" (Rosenbaum, 2002, p.76). However, even when controlling for all available pre-treatment variables is theoretically appropriate, our results indicate that it can lead to problems if PSM is the method of adjustment, since PSM destroys more information in higher dimensions and thus makes the PSM paradox more problemmatic. The advice is of course still appropriate for other matching methods. This problem with PSM is exacerbated by the fact that performance of most classifiers (including the propensity score regression) degrades fast with the number of covariates, because in high dimensions the points are massively spread out and so what would have been a somewhat easier interpolation problems in lower dimensions now require extrapolation (see Hastie, Tibshirani and Friedman 2001, pp.22–24 and Beyer et al. 1998).

Finally, some researchers reestimate the propensity score logit after using PSM to remove observations outside the common support of the treated and control groups. The justification is that the non-common support data may be thought of as outliers, but the use of PSM to perform this task is probably not advisable. That is, reestimation may pose a serious problem because the remaining data will be better balanced and thus make the PSM paradox more prominent. This problem with PSM and the others can be remedied by the user of a space graph or other method of comparison.

A Data

Figure 2 analyzes four data sets: (1) a *Healthways* data set ($n = 19,846$) designed to evaluate the firm's disease management programs. The treatment variable includes intervention via telephone, mailings, and the web, among other practices, to attempt to get patients with chronic illnesses to follow their doctor's advice. The outcome variable is the number of dollars spent in caring for a patient in the year in which treatment was administered. We match on 18 pre-treatment covariates (such as demographics, specific chronic illnesses, prior costs, etc.) that may affect health costs and that factor into the decision of who receives treatment. (2) a *Call/No Call* data set, designed to evaluate only the causal effect of a telephone call by Healthways for the same purpose and with the same outcome variable as in the first data set. This data set includes to 84,868 individuals in a large employer-based contract with Healthways. The 41 covariates include demographics, health conditions, prior spending, and prior interactions with the medical establishment. (3) An *FDA* (Food and Drug Administration) data set collected and analyzed by Carpenter (2002) and reanalyzed by Ho et al. (2007), designed to assess the effect of partisan control of the U.S. Senate on FDA drug approval times. The data includes 18 control variables to match on, including clinical and epidemiology factors and firm characteristics; it also includes 408 new drugs reviewed by the FDA, of which 262 were eventually approved. Finally, (4) a data set introduced by Lalonde (1986), and cleaned and subsetted by Dehejia and Wahba (2002) to evaluate a 12-18 month job training program, and used as an example throughout the methodological literature on matching. These data include 297 treated units, 2,490 controls, and 10 pre-treatment covariates including demographics, and unemployment and earnings history.

References

Austin, Peter C. 2008. "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003." *jasa* 72:2037–2049.

Austin, Peter C. 2009. "Some Methods of Propensity-Score Matching had Superior Performance to Others: Results of an Empirical Investigation and Monte Carlo simulations." *Biometrical Journal* 51(1, February):171–184.

Beyer, Kevin, Jonathan Goldstein, Raghu Ramakrishnan and Uri Shaf. 1998. When is 'Nearest Neighbor' Meaningful? In *ICDT'99, LNCS 1540*, ed. Catriel Beeri and Peter Buneman. Berlin: Springer-Verlag pp. 217–235.

Brookhart, M. Alan, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn and Til Sturmer. 2006. "Variable Selection for Propensity Score Models." *American Journal of Epidemiology* 163(April):1149–1156.

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

Carpenter, Daniel Paul. 2002. "Groups, the Media, Agency Waiting Costs, and FDA Drug Approval." *American Journal of Political Science* 46(2, July):490–505.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens and Oscar Mitnik. 2009. "Dealing with limited overlap in estimation of average treatment effects." *Biometrika* 96(1):187.

Dehejia, Rajeev H. and Sadek Wahba. 2002. "Propensity Score Matching Methods for Non-Experimental Causal Studies." *Review of Economics and Statistics* 84(1):151–161.

Gu, X.S. and Paul R. Rosenbaum. 1993. "Comparison of multivariate matching methods: structures, distances, and algorithms." *Journal of Computational and Graphical Statistics* 2:405–420.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Ho, Daniel, Kosuke Imai, Gary King and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15:199–236. http://gking.harvard.edu/files/abs/matchp-abs.shtml.

Iacus, Stefano M., Gary King and Giuseppe Porro. 2009. "CEM: Coarsened Exact Matching Software." *Journal of Statistical Software* 30. http://gking.harvard.edu/cem.

Iacus, Stefano M., Gary King and Giuseppe Porro. 2011. "Multivariate Matching Methods that are Monotonic Imbalance Bounding." *Journal of the American Statistical Association* 106:345–361. http://gking.harvard.edu/files/abs/cem-math-abs.shtml.

Iacus, Stefano M., Gary King and Giuseppe Porro. 2012. "Causal Inference Without Balance Checking: Coarsened Exact Matching." *Political Analysis* . http://gking.harvard.edu/files/abs/cem-plus-abs.shtml.

Imai, Kosuke, Gary King and Elizabeth Stuart. 2008. "Misunderstandings Among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* 171, part 2:481–502. http://gking.harvard.edu/files/abs/matchse-

abs.shtml.

Imbens, Guido W. and Donald B. Rubin. 2009. "Causal Inference." Book Manuscript.

Lalonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs." *American Economic Review* 76:604–620.

Nielsen, Richard A., Michael G. Findley, Zachary S. Davis, Tara Candland and Daniel L. Nielson. 2011. "Foreign Aid Shocks as a Cause of Violent Armed Conflict." *American Journal of Political Science* 55(2, April):219–232.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Rosenbaum, Paul R. 2002. *Observational Studies, 2nd Edition*. New York, NY: Springer Verlag.

Rosenbaum, Paul R. and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79:515–524.

Rosenbaum, Paul R. and Donald B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician* 39:33–38.

Rosenbaum, P.R., R.N. Ross and J.H. Silber. 2007. "Minimum Distance Matched Sampling With Fine Balance in an Observational Study of Treatment for Ovarian Cancer." *Journal of the American Statistical Association* 102(477):75–83.

Rubin, Donald. 1976. "Inference and Missing Data." *Biometrika* 63:581–592.

Rubin, Donald B. 2008*a*. "Comment: The Design and Analysis of Gold Standard Randomized Experiments." *Journal of the American Statistical Association* 103(484):1350–1353.

Rubin, Donald B. 2008*b*. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2(3):808–840.

Rubin, Donald B. 2010. "On the Limitations of Comparative Effectiveness Research." *Statistics in Medicine* 29(19, August):1991–1995.

Rubin, Donald B. and Elizabeth A. Stuart. 2006. "Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions." *Annals of Statistics* 34(4):1814–1826.

Stuart, Elizabeth A. 2008. "Developing practical recommendations for the use of propen-

sity scores: Discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003'." *Statistics in Medicine* 27(2062–2065).

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

Stuart, Elizabeth A. and Donald B. Rubin. 2007. Best practices in quasi-experimental designs: Matching methods for causal inference. In *Best Practices in Quantitative Methods*, ed. Jason Osborne. New York: Sage pp. 155–176.