Why Propensity Scores Should Not Be Used for Matching[*]

Gary King[†] Richard Nielsen[‡]

July 17, 2015

Abstract

Researchers use propensity score matching (PSM) as a data preprocessing step to selectively prune units prior to applying a model to estimate a causal effect. The goal of PSM is to reduce imbalance in the chosen pre-treatment covariates between the treated and control groups, thereby reducing the degree of model dependence and potential for bias. We show here that PSM often accomplishes the opposite of what is intended — increasing imbalance, inefficiency, model dependence, and bias. The weakness of PSM is that it attempts to approximate a completely randomized experiment, rather than, as with other matching methods, a more powerful fully blocked randomized experiment. PSM, unlike other matching methods, is thus blind to the often large portion of imbalance that could have been eliminated by approximating full blocking. Moreover, in data balanced enough to approximate complete randomization, either to begin with or after pruning some observations, PSM approximates random matching which turns out to increase imbalance. For other matching methods, the point where additional pruning increases imbalance occurs much later in the pruning process, when full blocking is approximated and there is no reason to prune, and so the danger is considerably less. We show that these problems with PSM occur even in data designed for PSM, with as few as two covariates, and in many real applications. Although these results suggest that researchers replace PSM with one of the other available methods when performing matching, propensity scores have many other productive uses.

[*] We thank Alberto Abadie, Justin Grimmer, Jens Hainmueller, Chad Hazlett, Seth Hill, Stefano Iacus, Kosuke Imai, Simon Jackman, John Londregan, Adam Meirowitz, Giuseppe Porro, Molly Roberts, Jamie Robins, Bradley Spahn, Brandon Stewart, Liz Stuart, Chris Winship, and Yiqing Xu for helpful suggestions, and Connor Jerzak, Chris Lucas, Jason Sclar for superb research assistance. We also appreciate the insights from our collaborators on a previous related project, Carter Coberley, James E. Pope, Aaron Wells.

[†] Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge MA 02138; GaryKing.org, king@harvard.edu, (617) 500-7570.

[‡] Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139; mit.edu/∼rnielsen, rnielsen@mit.edu, (857) 998-8039.

1 Introduction

Matching is an increasingly popular method for preprocessing data to improve causal inferences in observational data (Ho et al., 2007; Morgan and Winship, 2014). The goal of matching is to reduce imbalance in the empirical distribution of the pre-treatment confounders between the treated and control groups (Stuart, 2010, p.13). Lowering imbalance reduces, or reduces the bound on, the degree of model dependence in the statistical estimation of causal effects (Ho et al., 2007; Imai, King and Stuart, 2008; Iacus, King and Porro, 2011*b*), and, as a result, reduces inefficiency and bias. The resulting process amounts to a search for a data set that might have resulted from a randomized experiment hidden within an observational data set. If matching can reveal this "hidden experiment," many of the severe problems of observational data analysis vanish.

Propensity score matching (PSM) (Rosenbaum and Rubin, 1983) is the most common matching method, possibly even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010); it is used or referenced in over 48,200 scholarly articles.[1]

We show here that PSM, as it is most commonly used in practice or with some of the refinements that have been proposed, can and usually does increase imbalance, inefficiency, model dependence, and bias at some point in both real data and in data generated to meet the requirements of PSM theory. In fact, the more balanced the data, or the more balanced it becomes by pruning some observations through matching, the more likely PSM will degrade inferences — a problem we refer to as the *PSM paradox*. If one's data are so imbalanced that making valid causal inferences from it without heavy modeling assumptions is impossible, then the paradox we identify is avoidable and PSM will reduce imbalance but then the data are probably not very useful for causal inference by any method.

We trace the PSM paradox to the particular way propensity scores interact with matching. Thus, our results do not necessarily implicate the many other productive uses of propensity scores, such as regression adjustment (Vansteelandt and Daniel, 2014), inverse

[1] Count according to Google Scholar, accessed 1/4/2015, searching for: "propensity score" (matching OR matched OR match).

weighting (Robins, Hernan and Brumback, 2000), stratification (Rosenbaum and Rubin, 1984), and some uses of the propensity score within other methods (Diamond and Sekhon, 2012; Imai and Ratkovic, 2014), among others. In particular, the mathematical theorems in the literature used to justify propensity scores in general, such as in Rosenbaum and Rubin (1983), are of course correct and useful elsewhere. However, most of the theorems do not apply to finite samples, do not attempt to account for bias generated by model dependence, and are not specialized to matching — all of which are crucial for applications.

We give our notation and the goals of matching methods in Section 2. In Section 3, we show that PSM is blind to an important source of information in observational studies because it approximates a completely randomized rather than a more informative and powerful, fully blocked experiment (Rubin and Thomas, 2000). We then show, in Section 4, that PSM's inefficiencies are not merely a matter of ignoring potentially useful covariate information. When data are well balanced either to begin with or after pruning some observations, the fact that PSM is approximating the coin flips of a completely randomized experiment means that it will prune observations approximately randomly, which we show increases imbalance, model dependence, and bias. As a result, other matching methods are normally able to achieve lower levels of imbalance than PSM and do not generate a similar paradox until much later in the pruning process, when a fully blocked experiment is approximated and pruning is not needed.

Furthermore, we show that the imbalance and model dependence generated by PSM leads directly to biased estimates, since it leaves researchers making qualitative choices among larger sets of causal estimates. Similar to common warnings that researchers should match without looking at the outcome variable so as not to inadvertently induce selection bias (e.g., Rubin, 2008*b*), we must limit model dependence to avoid bias induced by the researcher (perhaps inadvertently) cherry-picking causal estimates from a wider than necessary set of possible choices. As we explain below, the social psychological (and statistical) evidence that qualitative choices like these lead to bias is overwhelming.

Fortunately, since other commonly used matching methods reduce imbalance, model dependence, and bias more effectively than PSM, and do not typically suffer from the same paradox, matching in general should remain a recommended method of causal inference. Section 5 offers advice to those who wish to use PSM despite the problems and to those using other methods.

2 The Goal of Matching Methods

For applied researchers, "the goal of matching is to create a setting within which treatment effects can be estimated without making heroic parametric assumptions," (Hill, 2008). Our notation, setup, and results apply more generally, but we focus on the simplest and most common use of PSM.

For unit i ($i = 1, \ldots, n$), denote $T_i \in \{0, 1\}$ as the treatment variable, where 0 denotes the "control group" and 1 the "treated" group. Let X_i be a vector of k pre-treatment covariates and Y_i a scalar outcome variable. Crucially, the process by which values of T are assigned is not necessarily random, controlled by the researcher, or known.

2.1 Quantity of Interest

Define $Y_i(1)$ and $Y_i(0)$ as the "potential outcomes," the values Y_i would take on if treatment or control were applied, respectively; only one of the potential outcomes is observed for each unit i: $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$. The treatment effect for unit i is the difference $\text{TE}_i = Y_i(1) - Y_i(0)$. Causal quantities of interest are then averages of TE_i over different subsets of units in the sample, or the population from which we can imagine the sample was drawn. For simplicity, we focus on the sample average treatment effect on the treated,

$$\text{SATT} = \frac{1}{\#\{T_i = 1\}} \sum_{i \in \{T_i = 1\}} \text{TE}_i, \tag{1}$$

and, if some treated units have insufficiently good matches and are thereby pruned as part of the matching procedure, then the feasible SATT (or FSATT).

Using FSATT is widely recommended in the literature, and widely used by applied researchers, so long as one is careful to characterize the resulting quantity of interest (see Crump et al., 2009; Rubin, 2010; Iacus, King and Porro, 2011*b*). Thus, the inferential goal in most cases in matching is a quantity that is well characterized by the distribution of the values of the covariates.

2.2 How Model Dependence Generates Bias

Identifying assumptions and, conditional on the choice of an outcome model applied to raw or matched data, optimal estimation procedures and the necessary associated assumptions are well known (Lechner, 2001; Imbens, 2000; Iacus, King and Porro, 2015). However, the nature of observational research means that the data generation process necessary to choose the correct model to condition on is often unknown, and the differences in causal estimates across these models can be large. Thus, from this large range of possible models and corresponding estimates, the researcher typically chooses one or, at best, 4–5 (often in different columns of a table) to publish. Crucially, the analyst uses the emirical estimates while selecting which to report. This results in the *model dependence* problem, leading Ho et al. (2007, p.199) to ask "How do readers know that publications are not merely demonstrations that it is *possible* to find a specification that fits the author's favorite hypothesis?" More precisely, model dependence is defined as the variation in causal estimates from two or more models that fit a data set approximately equally (King and Zeng, 2006; Iacus, King and Porro, 2011*b*).

At best, model dependence generates additional often unaccounted for uncertainty (Athey and Imbens, 2015; Efron, 2015; King and Zeng, 2007). However, a researcher's choice of estimates to report can turn a set of even unbiased estimates into a severely biased estimator. To see this, consider a set of J models, $M_1, \ldots, M_J$, which generate estimators $\hat{\tau}_1, \ldots, \hat{\tau}_J$ of the causal effect τ. Suppose we have model dependence, so that in any one data set the estimates vary: $\frac{1}{J}\sum_{j=1}^{J}(\hat{\tau}_j - \bar{\hat{\tau}})^2 > 0$, where $\bar{\hat{\tau}} = \frac{1}{J}\sum_{j=1}^{J}\hat{\tau}_j$. Assume that each estimator is unbiased conditional on its model (i.e., the average over repeated samples equals the true causal estimate): $E(\hat{\tau}_j|M_j) = \tau$ (for $j = 1, \ldots, J$).

Now consider a new estimator $\hat{\tau}_0$, in which a researcher chooses one of the existing J estimates to report, in part on the basis of the empirical estimates. Since the researcher would hypothetically choose a different estimate on each randomly drawn data set, we can no longer condition on the model. Define the estimator as $\hat{\tau}_0 = g(\hat{\tau}_1, \ldots, \hat{\tau}_J)$, for any $g(\cdot)$ other than a fixed weighted linear average. For example, one simple but realistic example has the researcher choosing the maximum among the estimates, $\hat{\tau}_0 = \max(\hat{\tau}_1, \ldots, \hat{\tau}_J)$.

The resulting estimator is biased; $E(\hat{\tau}_0) \neq \tau$. Thus, *a human making an unconstrained qualitative choice from among a set of unbiased but different estimates is a biased estimator*. This is the reason that scholars who study matching uniformly recommend that Y should not be consulted during the matching process to prevent inadvertent biases (Rubin, 2008*b*). The same logic also implies that controlling bias requires that we reduce model dependence.

The social psychological literature has unambiguously shown that biases are likely to affect qualitative choices like these even when researchers conscientiously try to avoid them (Banaji and Greenwald, 2013). The tendency to imperceptibly favor one's own hypotheses, or to be swayed in unanticipated directions even without strong priors, is unavoidable. People do not have easy access to their own mental processes and they have little self-evident information to use to avoid the problem (Wilson and Brekke, 1994). To make matters worse, subject matter experts overestimate their ability to control their personal biases more than nonexperts, and more prominent experts are the most overconfident (Tetlock, 2005). Moreover, training researchers to make better qualitative decisions based on empirical estimates when there exists little information to choose among them scientifically is unlikely to reduce bias even when people are taught these social-psychological results. As Kahneman (2011, p.170) explains, "teaching psychology is mostly a waste of time". Or, in conclusion, "Do not trust anyone — including yourself — to tell you how much you should trust their judgment" (Kahneman, 2011, p.240).

Scientists are no different from other human beings in this regard. Researchers have long shown that flexibility in reporting, presentation, and analytical choices routinely leads directly to biased decisions, consistent with the researcher's hypotheses (Ioannidis, 2005; Mahoney, 1977; Simmons, Nelson and Simonsohn, 2011). The literature makes clear that the way to avoid these biases is to remove the flexibility from the researcher as much as possible, rather than instituting training sessions or encouraging everyone to try harder to avoid bias (Wilson and Brekke, 1994, p.118).

Indeed, we can even conceptualize one of the central projects of statistical science to be automating an ever increasing portion of what was previously unaided, qualitative human decision making. In the present case, bias can be removed by taking unnecessary arbitrary decisions away from the researcher by reducing levels of model dependence and bias.

2.3 Matching to Reduce Model Dependence

The origin of model dependence can be understood in at least three related ways, all of which are functions of *the distance of counterfactuals from the data, measured in the space of the covariates*. First, model dependence increases when making counterfactual predictions farther from the data (King and Zeng, 2006, §2.1). Counterfactual predictions arise in causal inference when observed outcome variable values for control units $Y_i \equiv Y_i(0)$ are used to fill in for unobserved $Y_j(0)$ potential control outcomes in treated units. Matches that are closer in the space of X are less strained counterfactuals and generate less model dependence. Second, explicit measures of *imbalance*, the overall difference in the empirical distribution of X between the treated and control groups, translate directly into model dependence. A precise mathematical description and bias decomposition in these terms are given in Imai, King and Stuart (2008). Finally, a proof, that controlling imbalance directly controls and mathematically bounds the degree of model dependence, is given in Iacus, King and Porro (2011*b*, §2.4).

The statistical foundation of the claim, that model dependence is a function of the distance from the data to the counterfactuals measured in the space of X, is the assumption that observations similar on their covariate values are also similar with respect to the potential outcomes. This assumption, which is usually formalized by restricting the relationship to the space of Lipschitz functions, is at least implicitly accepted throughout the matching literature, including those that match based on calculations in other projected spaces such as PSM (see King and Zeng 2006; Iacus, King and Porro 2011*b*; Kallus 2015.

Successful applications of matching thus involve pruning units so that imbalance is reduced in the data set remaining, which in turn reduces the maximum level of model dependence. Reducing model dependence reduces the discretion of the researcher and can, as a result, greatly reduce bias.

2.4 Matching Methods

We briefly describe here PSM and two other matching methods representative of the large variety used in the literature. We first present the simplest and most widely used version of each of the three methods and then discuss more rarely used refinements of PSM. We also report on a content analysis we conducted of the prevalence of these refinements across the applied literatures.

Each method we define here represents one of the two existing classes of matching methods: Mahalanobis Distance Matching (MDM) is one of the longest standing matching methods that can fall within the Equal Percent Bias Reducing (EPBR) class (Rubin, 1976; Rubin and Stuart, 2006) and Coarsened Exact Matching (CEM) is the leading example within the Monotonic Imbalance Bounding (MIB) class (Iacus, King and Porro, 2011*b*). PSM can also be EPBR, if used with appropriate data.

We begin with MDM and PSM, which are based on distance metrics. Under MDM, the metric is the Mahalanobis distance between treated unit i and control unit j: $\mathrm{D}(X_i, X_j) = \sqrt{(X_i - X_j)S^{-1}(X_i - X_j)}$, where S is the sample covariance matrix of the original data matrix X. (MDM works well for methods articles, where the standardization makes the variables unitless; in applications, metrics such as Euclidean distance better enable a researcher to represent knowledge of the underlying variables and their relative importance.) In contrast, under PSM, the k-dimensional vector X_i is first reduced to a scalar propensity score, $\pi_i \equiv \Pr(T_i = 1|X_i)$, in practice almost always by assuming and estimating a logistic regression model $\hat{\pi}_i = (1+e^{-X_i\hat{\beta}})^{-1}$. Then the distance between treated unit i and control unit j is based on the absolute difference in propensity scores $\mathrm{D}(X_i, X_j) = |\hat{\pi}_i - \hat{\pi}_j|$ (or their log odds). A key point is that under PSM, information in X other than that projected in the propensity scores $\hat{\pi}$ is ignored.

Then, following the most common implementation of either PSM or MDM (see Austin 2009, p.173 and the content analysis below) we (greedily) match the first treated unit to the closest control; then, we match the second treated unit with the closest control unit that has not already been matched, and so on until all treated units have matches. Optimal matching is usually preferred to greedy (Hansen, 2004); similarly, j-to-k matching (with

j and k varying over units) is usually preferred to one-to-one matching. However, for our analyses below, no conclusions change when we reanalyze our results in these ways, and so for this paper we only present the simpler, and more commonly used, approach.

Finally, any treated-control matched pairs that have distances larger than a chosen caliper are also pruned (Rosenbaum and Rubin, 1985; Stuart and Rubin, 2007). Since a caliper is an arbitrary choice widely recommended and routinely made in practice, a matching method can be thought of as producing a *sequence of matched data sets*, the first with no caliper (and the number of observations equaling twice the number of treated units), and subsequent matched data sets created by continuing to caliper pairs of observations off until no data are left.

Under CEM, we first temporarily coarsen each of the pre-treatment covariates (Iacus, King and Porro, 2011*a*). Continuous covariates could be coarsened at "natural breakpoints," such as high school and college degrees in years of education, poverty level for income, etc. Discrete variables can be left as is or categories can be combined, such as when data analysts combine strong and weak Democrats into one category and strong and weak Republicans into another. (Variables can also be coarsened in groups of related variables, such as requiring the sum of three dichotomous variables to be equal.) Then units that do not match exactly on all the temporarily coarsened variables are pruned. Finally, the uncoarsened values of X are passed on to the estimation stage. Since coarsening is an adjustable choice, like calipers, CEM should also be thought of as producing a sequence of matched data sets that prune more units as data are less coarsened.

Numerous refinements of these methods, and many others, have appeared (e.g., Lunceford and Davidian, 2004; Imbens, 2004; Ho et al., 2007; Stuart, 2010), including preceeding PSM with exact matching on a few variables and several ways of iterating between PSM and balance calculations in the space of X (e.g., Rosenbaum and Rubin, 1984; Ho et al., 2007; Rosenbaum, Ross and Silber, 2007; Imbens and Rubin, 2015; Austin, 2008; Caliendo and Kopeinig, 2008; Stuart, 2010). As we now show, these alternative approaches are relatively rare in the applied literature and so we focus on the simplest and most widely used approaches described above. To measure how PSM is used, we conduct a content analysis of articles which use it. To do this, we downloaded from the JSTOR repostory 1,000 randomly selected English language articles, 1983–2015, which reference PSM. We then downloaded all 709 of those not behind a paywall, read each one, and narrowed the list to the 279 which used PSM and, of these, the 230 which applied PSM to real data (49 additional articles were primarily methodological).

We find that only 6% of the applied articles use the iterative procedure. The rest use the simple version of PSM described here (80%) or do so after exact matching on a few crude variables, such as school district in education or age group and sex in public health (14%). We reserve discussion of these two alternative versions of PSM until Section 5.

3 Information Ignored by Propensity Scores

Matching can be thought of as a technique for finding experimental data hidden within an observational data set. However, we show here that different matching methods try to approximate different experimental ideals, and PSM makes a choice with low standards, causing it to be blind to a valuable source of information available in observational data.

3.1 Different Experimental Ideals of Different Matching Methods

One way to think about MDM, CEM, and other methods aside from PSM is that they attempt to approximate a *fully blocked randomized experimental design*, such as a matched pair randomized experiment. In this design, treated and control groups are blocked at the start of the experiment exactly on the observed covariates X — so that all potential bias with respect to X is eliminated exactly in sample — and potential bias due to unobserved variables is not controlled in sample but is eliminated on average across experiments via randomization within the homogeneous blocks.

Similarly, in observational data, MDM and CEM seek to eliminate potential bias due to the observed X by replacing blocking ex ante with matching ex post. They also attempt to control for potential bias due to unobserved covariates by replacing physical randomization with the usual unconfoundedness and common support assumptions.

The theoretical justification for propensity scores comes from the result that, on average across samples, unconfoundedness conditional on the raw covariates, $Y_i(0) \perp T_i \mid X_i$, implies unconfoundedness conditional on the propensity score, $Y_i(0) \perp T_i \mid \pi_i$ (Rosenbaum and Rubin, 1983). Thus, we can think of PSM as attempting to approximate a *completely randomized experimental design*, where treatment assignment T_i depends only on the given probability of treatment π_i. The simplest version of this design assigns the same probability of treatment to all units (say 0.5), but a common alternative version assigns a constant probability of treatment to all units within each of several assigned strata, and allows variation in the probability across strata. (This could happen, for example, if the experimenter could afford the treatment for say half the patients for one month but then only 25% for the next month.) This could of course be described as a stratified design, so long as we recognize that the strata are *not* subsets of the covariate space, as they are in a fully blocked experiment. Regardless, in this design, imbalance in either the observed or unobserved variables is not controlled in sample; imbalance is instead controlled only on average across experiments. The crucial difference between the designs, then, is that in any one data set, the observed X variables may be imbalanced in a completely randomized experiment but are never imbalanced for a matched pair experiment (Iacus, King and Porro, 2011*b*, p.349).

Similarly, in observational data, PSM seeks to eliminate potential bias by replacing ex ante randomized treatment assignment with an attempt ex post to discover a set of strata (such as matched treated-control pairs) within which propensity scores are constant. To see the information left on the table by PSM, note that equality between any two estimated scalar propensity scores, $\hat{\pi}_i = \hat{\pi}_j$, *does not* imply that the two corresponding k-dimensional covariate vectors are matched exactly, $X_i = X_j$ — even though exact matching on the covariates $X_i = X_j$ *does* imply that the propensity scores are exactly matched $\hat{\pi}_i = \hat{\pi}_j$. (PSM attempts to control for potential bias due to unobserved covariates in the same way as MDM and CEM, by adding the usual unconfoundedness and common support assumptions.)

The difference between the experimental ideals is crucial since, compared to a completely randomized experimental design, a fully blocked randomized experimental design has more power, more efficiency, lower research costs, more robustness, less imbalance, and — most importantly from the perspective here — lower model dependence and thus less bias (Box, Hunger and Hunter, 1978; Greevy et al., 2004; Imai, King and Stuart, 2008; Imai, King and Nall, 2009). For example, Imai, King and Nall 2009 found that standard errors differed in their data between the two designs by as much as a factor of six. Compared to fully blocked randomized experiments "for gold standard answers, complete randomization may not be good enough, except for point estimation in very large experiments," (Rubin, 2008*a*). In fact, if the design is fully blocked on all X variables, all model dependence is eliminated with respect to these variables: That is, regardless of how a researcher controls for perfectly matched variables in the post-matching modeling step, the causal estimate is the same. In any one data set, completely randomized experimental designs usually have larger imbalance on X than if a fully blocked design were used. Of course, the discrepancy between the estimate and the truth in the one data set a researcher gets to analyze and thus publishes from is far more important to that researcher than what happens across hypothetical repeated samples from the same hypothetical population (cf. Gu and Rosenbaum, 1993). Put differently, an unbiased but inefficient estimator, conditional even on a randomly genenerated treatment assignment that in sample is to some degree imbalanced, is a biased estimator (Robins and Morgenstern, 1987).

3.2 Evidence for Propensity Score's Lower Standards

We now simulate 1,000 data sets, each of which mixes data from three separate sources: (1) a matched pair randomized experiment, (2) a completely randomized experiment, (3) observations that, when added to the first two components, make the entire collection an imbalanced observational data set. We then study whether MDM and PSM prune individual observations in the correct order — starting with those at the highest level of imbalance (data set 3) to the lowest (data set 1). For clarity, we use two covariates (using more covariates generates patterns like those we present here, only stronger).

To fix ideas, we display one of our 1,000 data sets in the left panel of Figure 1, which highlights its three parts in separate colors. In blue in the upper right is the matched pair experiment with 25 treated units drawn uniformly from the $[-2, 2] \times [-2, 2]$ square and 25 control units which are slightly jittered versions of each of these treated units. In red, at the bottom right is a completely randomized experiment, with 50 random observations drawn

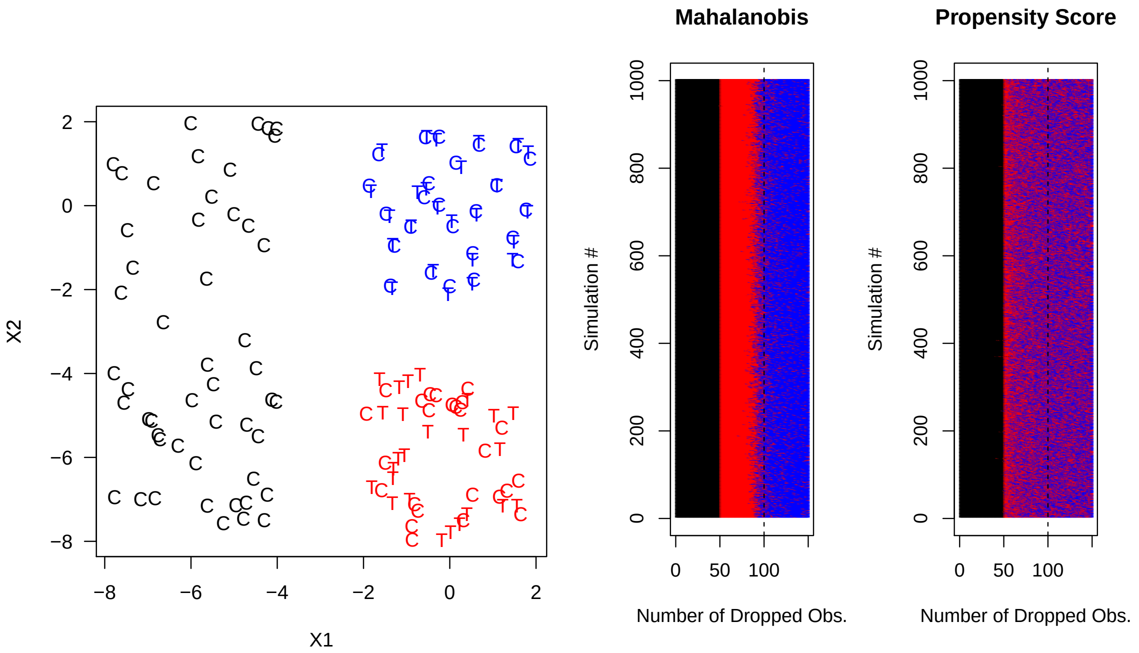

Figure 1: Finding experiments hidden in observational data, with PSM, but not MDM, blind to information from full blocking. Left panel: One (of 1,000) randomly generated data sets from a matched pair randomized experiment (in blue), a completely randomized experiment (in red), and control units from an imbalanced observational data set (in black). Right Panels: Each of the 1,000 simulations is represented by a separate row of pixels, color-coded by experiment type to indicate the order (from left to right) of which observations are pruned by MDM (left) and PSM (right).

uniformly from the $[-2, 2] \times [-8, -4]$ rectangle and with 25 of these randomly assigned to treatment and 25 to control. Finally, we simulate part of an imbalanced observational study by adding 50 control observations in black, drawn uniformly from the $[-6, -4] \times [-8, 2]$ square, and without corresponding randomly drawn (or otherwise overlapping) treated units.

We apply PSM and MDM, following the usual applied methodology described in Section 2.4, to each data set, and iteratively remove the (next) worst observation as defined by each matching method. In the two panels at the right of Figure 1, each row of pixels stands for one simulated data set, with each individual pixel in a row representing one pruned observation color-coded by data set type. The results show that both MDM and PSM do well at removing the 50 control units that lack common support with any treated units (black is separate in both). MDM is able to separate the fully randomized experiment from the matched pair experiment (red is clearly separated from blue) but PSM is unable to separate the more informative matched-pair experiment from the fully randomized experiment (red and blue are mixed).

In an application, a researcher may prefer to prune only the control units from the left part of the graph and no others. This would be best if SATT were the quantity of interest or, in some cases, to ensure that the variance is not increased too much by not pruning further. However, if the researcher chooses to prune more, and is willing to estimate FSATT, then using PSM would be a mistake. This simulation clearly shows that PSM cannot recover a matched pair experiment from these data reliably. At best, it can recover something that looks like a fully randomized experiment, meaning that the covariates can no longer predict treatment on average. This is useful, since it makes possible estimation that is unbiased before conditioning on the treatment assignment. However, by definition some model dependence and researcher discretion remains which, when combined can lead to bias. In observational work, the ideal should be a fully blocked experiment, which is approximated by exact matching, not merely overlapping data clouds. This is essential so that model dependence can be reduced as far as possible.

4 The Propensity Score Paradox

Given the differing goals of PSM and other methods, it is no surprise, after PSM's goal of complete randomization has been approximated, that other methods would be more effective at continuing to reduce imbalance on X than PSM. However, we find that pruning after this point with PSM does genuine damage — increasing imbalance, model dependence, and bias. That is, after this point, pruning the observations with the worst matched observations, according to the absolute propensity score distance in treated and control pairs, will increase imbalance, model dependence, and bias; this will also be true when pruning the pair with the next largest distance, and so on. We call this the PSM Paradox. The paradox is apparent in data designed for PSM to work well (Section 4.2) and often worse in real applications (Section 4.3).

The reason for the PSM Paradox is because, after PSM achieves its goals of finding a subset of the data that approximates complete randomization, with approximately constant propensity scores within strata, any further pruning is being done essentially at random, exactly as a completely randomized experiment. And, as we show in Section 4.1, random pruning, contrary to conventional wisdom, increases imbalance.

Another way to describe the reason for the paradox is as the consequence of PSM's unusual two-step procedure, which collapses the multivariate covariate vector to a scalar propensity score outside the space of the original covariates and then matches on the score. PSM is sometimes described as solving the curse of dimensionality problem, since researchers only need to condition on a scalar rather than k covariates (Dehejia, 2004). In fact, we show in Section 4.4 that PSM's two-step procedure is an increasingly worse summary of X_i as the number of elements k in the vector increase beyond one (see Brookhart et al., 2006). Although the curse of dimensionality affects every matching method — and in high enough dimensions no matching method will be very effective — the problem with PSM starts earlier, often with only two covariates.

Some issues with PSM are well known. For example, estimating the propensity score regression with misspecification can bias estimates (Drake, 1993; Smith and Todd, 2005; Kang and Schafer, 2007; Zhao, 2008; Diamond and Sekhon, 2012). For another, if X_i contains continuous variables, π_i will be continuous and so is no easier to match exactly than X. The PSM Paradox is in addition to these important points; even when the propensity score logit is correctly specified, and even if matches can be found, PSM discards considerable information, because it approximates a completely randomized rather than fully blocked experimental design, and this unconditional inefficiency is equivalent to conditional bias.

4.1 Random Matching

Obviously, one should only delete observations from a data set if some benefit results. As such, no sensible researcher would randomly prune observations from a data set (except perhaps for computational reasons with very large data sets). We show here that random pruning not only cuts down the information available; it actually *increases* the level of imbalance in the resulting data set. This is may seem counterintuitive, and to our knowledge has not before been noted in the matching literature (cf. Imai, King and Stuart, 2008, p.495). Then, in a separate subsection, we give intuition why, in common situations, PSM approximates random pruning.

4.1.1 Random Pruning Increases Imbalance

We define *random pruning* as a process of deleting observations in a data set that is independent of X. To offer intuition for how random pruning increases imbalance, we now give a discrete and a continuous example.

Consider first a simple discrete data set that happened to be perfectly balanced, with a treatment group composed of one male and one female, M_1, F_1, and a control group with the same composition, M_0, F_0. Then, randomly dropping two of the four observations leaves us with one matched pair among $\{M_1, M_0\}$, $\{F_1, F_0\}$, $\{M_1, F_0\}$, or $\{F_1, M_0\}$, with equal probability. This means that with 1/2 probability the resulting data set will be balanced ($\{M_1, M_0\}$ or $\{F_1, F_0\}$) and with 1/2 probability it will be completely imbalanced ($\{M_1, F_0\}$ or $\{F_1, M_0\}$). Thus, on average in these data random matching will increase imbalance.

For a simple continuous example, consider a randomly assigned T and a fixed univariate X. Consider, as a measure of imbalance, the squared difference in means between the treated and control group of X. The expected value of this measure (which equals the squared standard error of the difference in means) is proportional to $1/n$. Thus, as we prune from this sample randomly, n declines and our measure of imbalance increases.

Of course, if all the matching discrepancies are of the same size, pruning at random or by any other means will not change the average matching discrepancy (or most other measures of imbalance). But in more realistic simulations, and real data we have examined, random pruning increases imbalance. We also introduce a higher dimensional example with real data in Section 4.3.

4.1.2 PSM Approximates Random Matching

In a simple simulation, we provide intuition for how relatively balanced data makes PSM, but not MDM or CEM, highly sensitive to trivial changes in the covariates, often producing nonsensical results approximating random matching. In the left panel of Figure 2, we generate data with 12 observations and two covariates, with one covariate plotted by the other. The data are well balanced between treated (black disks) and control (open circles) units. From these initial data, we generate 10 data sets, where we add to each of the 12 observations a small 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 nearly on top of one another, but it only appears that one set is there). Next, we run CEM 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 linked by curved solid lines).

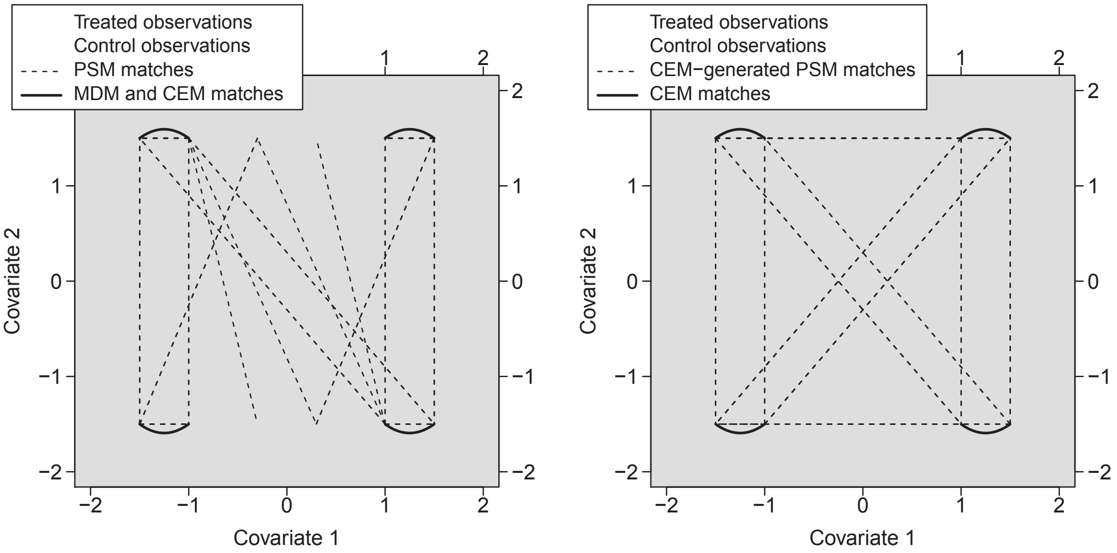

Figure 2: Ten data sets (differing from each other by imperceptibly small amounts of random error) with 4 treated units (black disks) and 8 control units (open circles). CEM and MDM match the closest control units to each treated (curved black lines). The two-step procedures match different control units for each data set, as can be seen for PSM (dashed lines, left panel) and PS-CEM (dashed 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.)

However, when we run PSM on each of the 10 data sets generated for Figure 2, the four treated units are each matched to *different* control units (as portrayed by the maze of dashed 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; it is blind to the space of X that is not represented in $\hat{\pi}$.

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: to do this, 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*b*) and, second, without running CEM as usual, we match directly on the nonparametric estimate of the propensity score. The right panel in Figure 2 is constructed the same way as the left panel except that instead of the dashed lines representing propensity score matches, they represent PS-CEM matches. The result is almost as bad as PSM. The dashed 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. This suggests that ignoring X and only matching based on the scalar propensity score generates the PSM paradox.

4.2 How PSM Generates Model Dependence and Bias

We now turn to a demonstration of how PSM generates model dependence and bias. We begin by hiding a completely randomized experiment within an imbalanced data set. Unlike Figure 1, we do not include a fully blocked experiment within these data. For each of two covariates, we randomly and independently draw 100 control units from a Uniform(0,5) and 100 treated units from Uniform(1,6). This leaves the overlapping $[1, 5] \times [1, 5]$ square as a completely randomized experiment and observations falling outside adding imbalance. We generate the outcome as $Y_i = 2T_i + X_{i1} + X_{i2} + \epsilon_i$, where $\epsilon \sim N(0, 1)$. We repeat the entire simulation 100 times. We assume, as usual, that the analyst knows the covariates necessary to achieve unconfoundedness but does not know the functional form.

To evaluate the methods, we compute both model dependence and potential for bias, each averaged over 100 simulated data sets. We measure model dependence by the variance in the estimate of the causal effect over 512 separate models (linear regression using every possible combination of X_1 and X_2 and their 3 second order and 4 third order effects) from the same simulated data set for each given caliper level; we do this for PSM and then also MDM as a comparison. The results, which appear in the top-left panel of

Figure 3, show that at first the degree of model dependence drops for both MDM and PSM, but then, after PSM has pruned enough so that the PSM paradox kicks in, model dependence dramatically increases. Instead of benefiting from units being dropped, PSM is causing damage. Model dependence for MDM, as expected, declines monotonically as stricter calipers are applied and more units are pruned.

To show how the combination of model dependence and analyst discretion can result in bias, we implemented an estimator meant to simulate the common situation where the analyst chooses a preferred estimate to publish from many possible estimates. Suppose the researcher's preferred hypothesis is that the causal effect is large, and that this preference intentionally or unintentionally affects their choice. Thus, for each caliper level of PSM and then MDM, we select the largest estimated treatment effect from among the estimates provided by the 512 possible specifications. The results, in the top-right panel of Figure3, show that calipering initially does what we would expect by reducing the potential for bias for both MDM and PSM, with PSM even slightly outperforming MDM. However, as calipering continues, the PSM paradox kicks in, and PSM increases model dependence (as indicated in the top left graph), the potential for bias with PSM dramatically grows even while the bias under MDM monotonically declines as we would expect and desire. (Although we do not show the graph, these patterns are unchanged for mean squared error as well.)

To provide intuition for how the paradox occurs in these data, we show which observations are matched and in which order in one of the 100 simulated data sets. Thus also in Figure 3, we plot X_1, X_2 points from one simulated data set, with matches from MDM (bottom-left panel) and PSM (bottom-right panel) denoted by lines drawn between the points, colored in by when they were matched or pruned in the calipering process. (The outcome variable is ignored during the matching process, as usual.) Darker colors were pruned later (i.e., matched earlier).

As expected, the MDM results (bottom-left panel) show that treated (circles) and control (dots) pairs that are close to each other are matched first (or pruned last). These darker blue lines mostly appear within the (completely randomized) square in the middle.

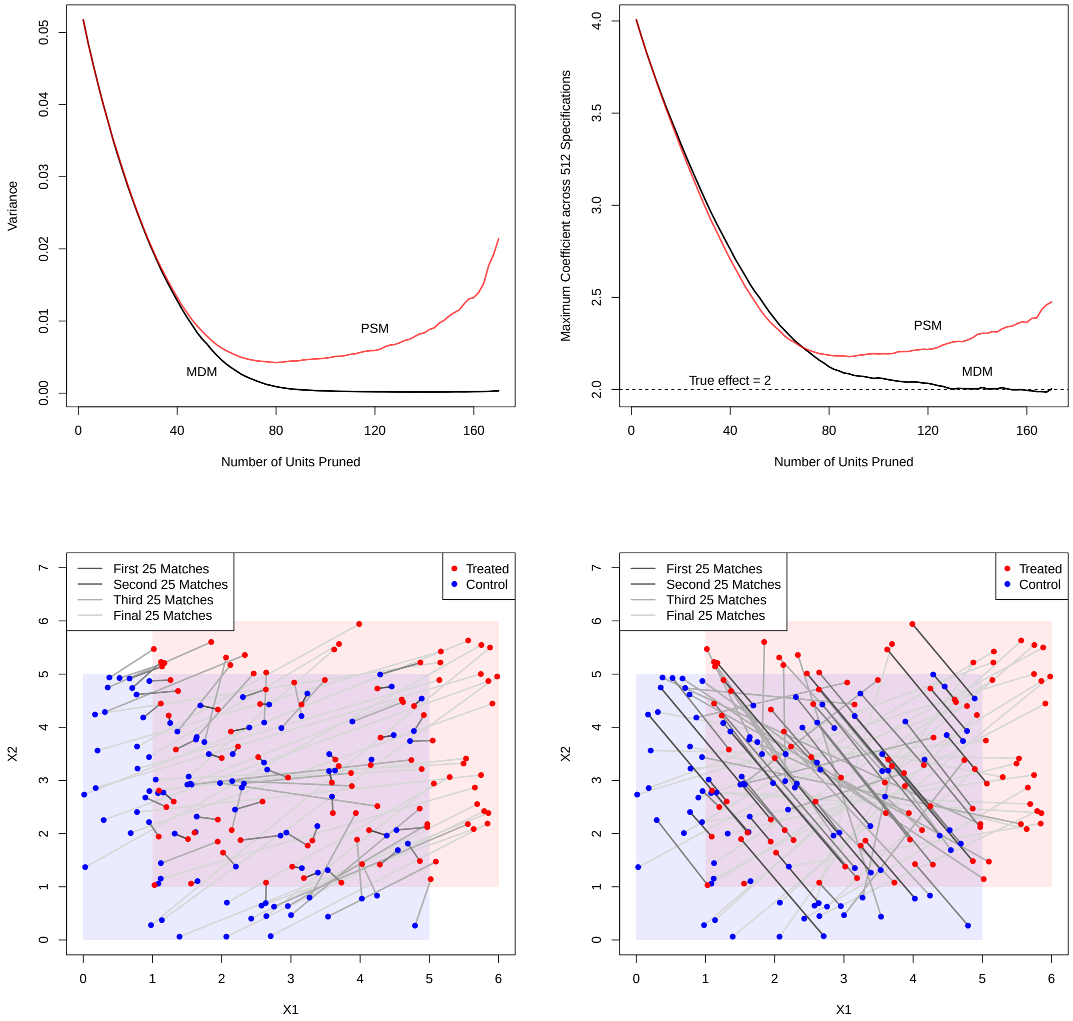

Figure 3: PSM Increases Model Dependence and Potential Bias. Top-left panel: The vertical axis measures model dependence as the average, over 100 data sets, of the variance in the causal effect estimate across 512 models applied to the same data. Top-right panel: The vertical axis shows the maximum estimated causal effect from 512 models applied to each of 100 data sets. For one simulated data set, the order of matches is indicated for MDM (bottom-left panel) and PSM (bottom-right panel).

In stark contrast, PSM, in the bottom-right panel, finds many matches seemingly unrelated to local closeness of treated and control units and many even outside the middle square. The diagonal pattern in PSM dark lines comes from the propensity score logit which cannot distinguish high values of X_1 and low values of X_2 from low values of X_1 and high values of X_2.

Overall, the figure shows that PSM is trying to match globally — meaning it essentially has only one chance to get it right, rather than matching locally like other methods and having some additional degree of robustness. In fact, because the propensity score is outside the space of the original data, using it for analysis violates the *congruence principle*. This principle which holds that the data space and analysis space should be the same. Statistical methods which violate this principle are known to generate nonrobust and counterintutive properties (Mielke and Berry, 2007).

4.3 Damage Caused in Real Data

In this section, we reveal the PSM paradox in real applications, with data selected and analyzed by others, including two published studies and a large number of others in progress. We obtained the data from the studies in progress by advertising to assist scholars in making causal inferences, in return for access to their data and a promise not to redistribute their data or publish their substantive results (or identities). For almost all the more than 20 data sets in progress we analyzed, we found patterns similar or analogous to the two we are about to present in detail. From this, we conclude that the PSM Paradox is prevalent in many real applications.

In this first published study we reanalyze, Finkel, Horowitz and Rojo-Mendoza (2012) showed that civic education programs in Kenya cause citizens to have more civic competence and engagement and be more supportive of the political system, with $n = 3,141$ survey responses, 1,347 of which received the program. They also measured a large number of socioeconomic, demographic, and leadership covariates. Second, Nielsen et al. (2011) show that a sudden decrease in international foreign aid to a developing country (an "aid shock") increases the probability of the onset of lethal conflict within that country. They collect data on developing countries from 1975 to 2006, in total representing $n = 2,627$ country-years, including 393 (treated) aid shocks. The authors measure 18 pre-treatment covariates representing national levels of democracy, wealth, population, ethnic and religious fractionalization, and prior upheaval and violence. Finally, we analyzed a large number of data sets obtained from scholars doing work in progress, which we received by trading offers of help with their analyses and promising not to cite or scoop them. The results of all sources of data yielded very similar conclusions to that from the two data sets we now reanalyze.

For both of the published studies we replicate, Figure 4 plots imbalance (vertically) by the number of pruned observations (horizontally). We measure imbalance (i.e., the difference between the empirical distribution of X in the treated and control groups) by the "Mahalanobis Discrepancy," proposed by Abadie and Imbens (2006), which as per Section 2.3 measures imbalance in the space of X.[2] In each plot, the open circle at the left summarizes the imbalance in original data set. For reference, we also add a solid triangle that summarizes the level of imbalance that would be present if T were assigned via complete randomization.

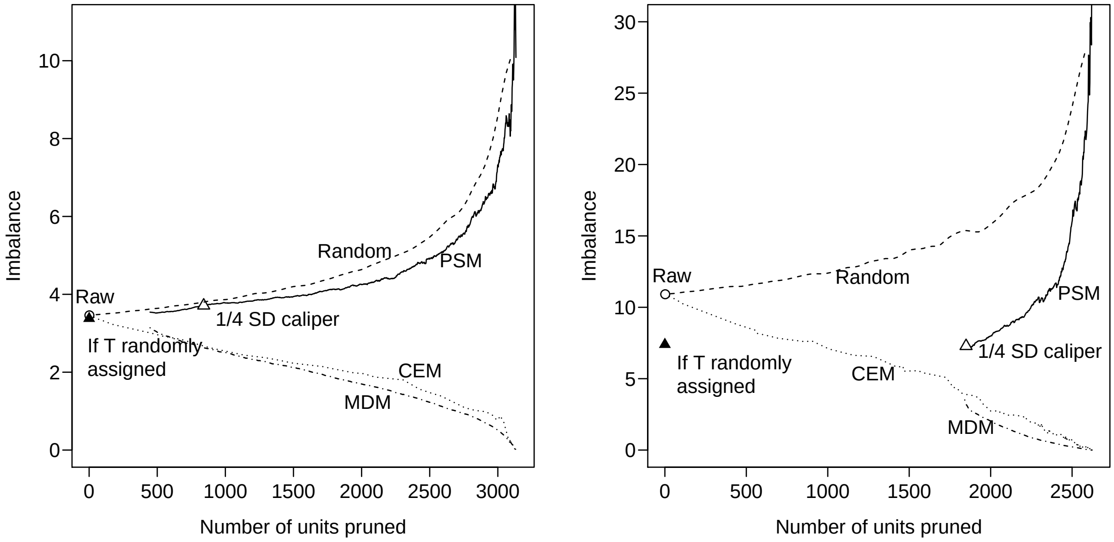

Figure 4: Imbalance-Matched Sample Size Graph, with data from Finkel, Horowitz and Rojo-Mendoza (2012) for the left panel and Nielsen et al. (2011) for the right.

The one-to-one PSM analysis (which in the case of Nielsen et al. 2011 is the published result and in both cases is estimated by all main effects in a logit model) is represented by the left end of the solid line. In the left panel, PSM's initial result is worse than the original data; in the right panel it is an improvement, as Nielsen et al. (2011) report. However, consider what happens in either data set if we caliper off the worst match according to

[2] We repeated the analysis the L_1 imbalance metric proposed by Iacus, King and Porro 2011*b*, and the average absolute difference in means for the columns of X, the components of which are often used in applied articles. Essentially the same conclusions result from each of these and other measures we have tried. We also repeated the analysis with various direct measures of model dependence, and found similar conclusions, although the large number of covariates in these applications mean that numerous measures could be chosen.

the propensity score metric (i.e., the largest value of $|\hat{\pi}_i - \hat{\pi}_j|$ across all matched pairs), recalculate the level of imbalance, and repeat. These results, which are represented by the full black line in each panel, reveal the PSM paradox kicking in immediately and continuing until no data is left: That is, the more strict we are in applying PSM, the worse imbalance gets. (In a few of the unpublished data sets we analyzed that had much worse initial imbalance, PSM helped for initial pruning and then started increasing imbalance as in these graphs; simulated examples of this pattern appear in Section 4.4.)

If we use the venerated practice of setting the caliper to 1/4 of a standard deviation of the propensity score, imbalance is worse than the basic PSM solution for the left panel and provides no improvement for the right panel. Following the strictures of PSM even more closely, in the hopes of finding better balance and less model dependence, accomplishes precisely the opposite.

For comparison, in each graph, we also prune via MDM (dot-dashed line) and CEM (dotted line). For MDM, we do one-to-one matching (and so the line starts at the same horizontal point as PSM) and then caliper off the observations with the largest Mahalanobis distance, recompute imbalance, and repeat. For CEM, we begin with the loosest possible coarsening, so that all data fall in a single stratum and no observations are pruned (and so the line starts from the original data). 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 the imbalance metric. Additional cutpoints eventually lead to more observations being dropped.

As can be seen in both panels in Figure 4, the MDM and CEM lines both tend downward through their entire range with no hint of the paradox that would be represented by an upward turn like PSM: in this case, the trade off is as it should be, in that one can reasonably choose any point along this frontier to do an analysis. (The figure also includes a dashed line marked “Random” representing the average of a sequence of data sets constructed by random pruning; as with the simpler examples in Section 4.1, the figure shows that random pruning increases imbalance.)

4.4 Damage Caused In Data Generated to Fit PSM Theory

We now study different types of simulations generated consistent with PSM theory. Results vary by the number of covariates, levels of imbalance, and matching method.

4.4.1 Data Generation Processes

We generate data by following Gu and Rosenbaum (1993). Covariates are drawn from a multivariate normal (meeting the data requirements of EPBR) with variances of 1 and covariances of 0.2. Data sets with high, medium, and low levels of balance result from setting the control group mean vector to (0,0,0) and different treated group mean vectors to (0.1,0.1,0.1), (1,1,1), and (2,2,2), respectively. We draw 250 treated and 250 control units, which is equivalent for our purposes to generating a larger pool of controls and pruning down to 250 to achieve 1-to-1 matching with the treated units. We then prune from that point by calipering off additional units.

We draw 50 random data sets, for each of the nine combinations of 1, 2, and 3 covariates and low, medium, and high levels of imbalance. (Analyses with more covariates and higher levels of imbalance predictably produce even more dramatic patterns than presented here and so we do not present them.) For each data set generated, we analyze the same data with PSM, CEM, and MDM, following the procedures from Section 4.3. We repeat the procedure for each level of pruning and for each of the 50 data sets, average, and put a point on a graph we present below. All our results apply to estimating both SATT and FSATT; for simplicity, we present results here for the latter, which is the most commonly recommended and used approach.

4.4.2 Results

Figure 5 gives the results for the methods in three rows (PSM, MDM, and CEM from top to bottom) and different numbers of covariates in separate columns (1,2,3, from left to right). Each of the nine graphs in the figure gives results from data generated with low (dotted line), medium (dashed line), and high (solid line) levels of initial imbalance. For graphical clarity, individual matching solutions do not appear and instead we average over the 50 simulations in each graph for a given matched sample size and level of imbalance. The PSM paradox is revealed whenever one of the lines increase from left to right (i.e., with imbalance increasing as the number of observations drops).

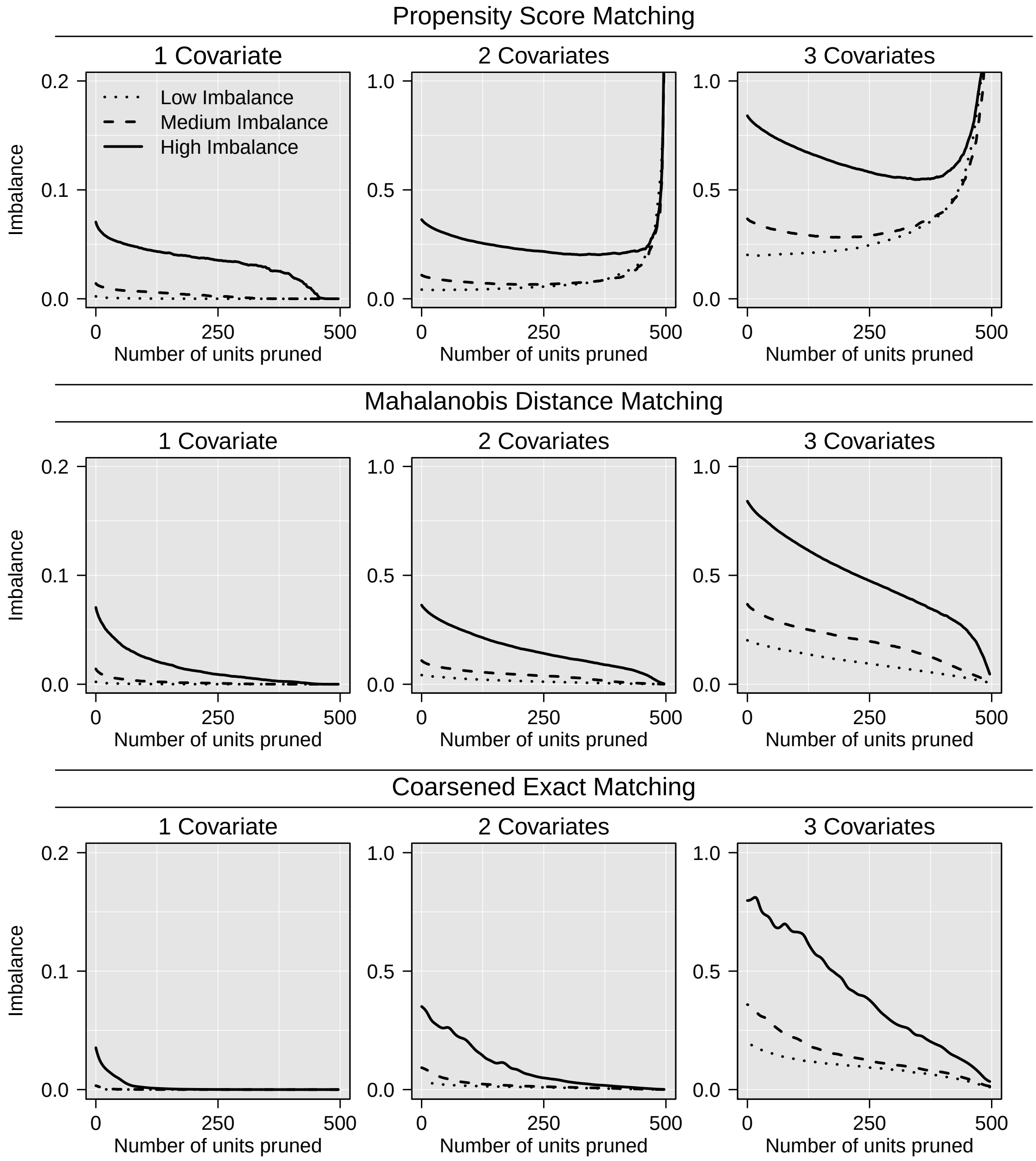

Figure 5: For PSM, MDM, and CEM in rows, and 1, 2, and 3 covariates in columns, these graphs give average values from 50 simulations with low (dotted), medium (dashed), and high (solid) levels of initial imbalance between the treated and control groups. The paradox is revealed for portions of lines that systematically increase. This can be seen for PSM with more than one covariate but not for CEM and MDM.

As expected, the usual curse of dimensionality reduces the performance of all three matching methods, as can be seen by the level of imbalance increasing from graphs at the left to the graphs at the right in any one row. Also as expected, the second and third rows of the Figure 5 show that CEM and MDM do not suffer from the paradox in these data: for all the lines in all the graphs in these rows, imbalance never increases as more observations are pruned, just as we would want to be the case.

However, for PSM in the first row, three important patterns emerge, all of which occur when the propensity score paradox kicks in (where a line changes direction from heading downward to where it starts heading upwards). First, no paradox emerges with PSM and one covariate (top left graph) because PSM does no dimension reduction; in this case, the propensity score is merely a rescaling of a scalar X. Second, the paradox point appears earlier, that is with fewer observations pruned, the more covariates are included in the propensity score regression (as we go from left to right in the top row of Figure 5). This problem is worse for 3 than 2 covariates and, although we do not show it, the paradox intensifies with more covariates. Third, the paradox kicks in earlier as the data become more balanced, approximating a completely randomized experiment. This can be seen by comparing the dotted (well balanced) and solid (least balanced) data in the two graphs at the top right, and noting that the point where the paradox starts moves to the left for better balanced data.

5 Guidance for Users

In this section, we offer guidance for those accustomed to using PSM, and prefer to keep using it, and for users of other matching methods. We also show how other methods can also generate a paradox but only under much more extreme circumstances. The main lesson using any matching method is that because observations are dropped, the technique will only be beneficial if something does enough good to counterbalance the negative effects of the loss of information. For PSM, the problems outweigh the good when pruning after complete randomization has been approximated, while for other methods the problems occur much later during pruning — only when they approximate a fully blocked experiment.

5.1 Advice for PSM Users

Our results indicate that those who wish to continue to use PSM would improve their work by adhering to the following seven points.

First, PSM can be used without increasing imbalance — if one is careful to check imbalance after running PSM and to verify that the PSM paradox is not in evidence. Setting a more restrictive caliper may well increase imbalance. Diagnostic plots like Figure 4 provide an easy way to assess whether PSM has reached this point. If the researcher is careful to stop pruning before imbalance starts to increase, the paradox can be avoided; the resulting PSM solution will likely be suboptimal compared to other matching methods, but it will be an improvement relative to the original data.

Second, researchers should be aware that PSM can help the most in data where valid causal inferences are least likely (i.e., with high levels of imbalance) and may do the most damage in data that are already well suited to making causal inferences (i.e., with low levels of imbalance).

Third, PSM is better justified in very large samples, and where relatively few observations are pruned, both so as not to go past the point of complete randomization and so that the difference after matching between fully blocked and completely randomized experiments is smaller and less consequential.

Fourth, the model dependence that remains after PSM can usually be reduced further if the researcher switches to another matching method, although this could be done after an initial application of PSM.

Fifth, some applications precede PSM with exact matching on a few covariates. The advantage of this procedure is that its first step can take one closer to full blocking than PSM alone is capable of. Its disadvantage is that it makes PSM even more dangerous: The closer the exact matching step comes to including all variation in X, the quicker the PSM step will run into the PSM paradox and begin to increase imbalance, model dependence, and bias.

Sixth, researchers who use PSM and are left with unnecessary model dependence and bias, and do not correct this with another matching method, should explicitly clarify how much imbalance, and therefore model dependence and bias, was left after applying PSM so inadvertent biases that may creep in due to researcher discretion can be evaluated by the reader.

Finally, to the extent that the iterative methods discussed in Section 2.4 match more on X than the propensity score, it seems plausible that they would perform better than PSM alone, although not as well as methods that match directly on X without the constraint of passing through a one dimensional modified propensity score function. In addition to being used only rarely in the applied literature, the theoretical properties of most of these approaches have rarely been studied. We briefly study this issue by replicating the PSM analyses in this paper with the version of the automated iterative procedure proposed in Imbens and Rubin (2015, Ch.13). We find little improvement, and not much change overall. For one example, we replicated the top three graphs from Figure 5 with the Imbens-Rubin iterative procedure and found almost impreceptible differences from those three graphs. This, however, is merely one illustration rather than a general result and so these methods remain worthy of further study.

5.2 Advice for Users of Other Matching Methods

Any matching method that prunes in a manner independent of the covariates (and thus is pruning randomly) can increase imbalance. With PSM, this point, which we call the PSM paradox, kicks in after the point of complete randomization is reached, since PSM is blind to information in X not represented in the propensity score.

For other matching methods that can detect all differences in X, pruning after approximating complete randomization will continue to help reduce imbalance. Much later, after we prune enough to approximate a fully blocked experimental design, all information in X has been exhausted. At that point, all the units are exchangable aside from their treatment assignment and so any further pruning can only take place at random (with respect to X), which would increase imbalance. Of course, at full blocking — such as for example exact paired matching — it is obvious that further pruning serves no purpose. We can however contrive two instances where researchers might be fooled. To illustrate, we offer two simulations that involve pruning with MDM after using all information in X.

In the first simulation, we contrive a data set where nature is malicious. We begin by generating 100 values of a single covariate X deterministically, in pairs along the number line as $X = 1, 2, 4, 5, 7, 8, \ldots, 145, 146, 148, 149$. We then assign observations with even values of X to receive treatment and those with odd values to receive control. If we stopped here, each treated unit would match best to the control observation 1 unit away and T would be independent of X in sample (and where both treated and control units of X have a mean of 75). Then to each value of X, we add a tiny amount of jitter drawn from a uniform on the interval $[-0.00001, 0.00001]$. This results in some pairs being slightly better matches than others, although solely due to random jitter. We then introduce confounding (which can be productively fixed via matching) by taking the three treated units with the lowest values of X and reassigning them to control, and taking the three control units with the highest X values and assign them to treatment. For example, this creates a substantial difference between the mean value of X for the treated (≈83) and control (≈67). We generate the outcome variable as $Y = T + 0.01X + \epsilon$, where $\epsilon \sim N(0, 1)$.

The resulting data set has important levels of imbalance (and confounding) due to the units at the low and high values of X. The rest of the data will have matches that are effectively at random. The idea is that any method of matching will first prune the extreme (imbalanced) observations first for good reason and then start pruning at random.

We measure model dependence by first estimating the regression of Y on a constant, T, and elements of one of the subsets of $\{X, X^2, X^3, X^4, X^5\}$, and then repeat for all the other subsets. Then our measure is the range of estimates of the coefficient on T across all these regressions. Results appear in Figure 6, with model dependence plotted vertically and the number of treated units pruned by MDM horizontally.

Thus, MDM first prunes the six extreme values of X which causes model dependence to drop. After that point, when all pairs differ by pure randomness, MDM continues to prune without accomplishing anything of value. Matching in this way does not overcome the fact that pruning itself increases imbalance, and so the overall imbalance line starts heading upward.

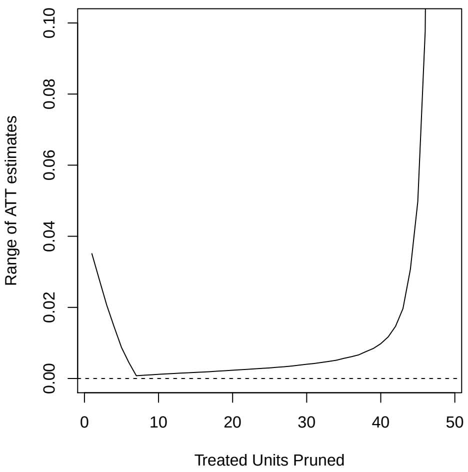

Figure 6: The Paradox with Mahalanobis Distance Matching

For a second illustration, we create a very small data set in high dimensional space so that points are so far spread out that few good matches are available. This is easy to see in MDM since Mahalanobis distances in this situation have the characteristic property of differing by tiny, essentially random amounts, only after many digits to the right of the decimal point. Thus, we generate a small data set, $n = 200$, with k covariates, for $k = 2, 3, 4, 5, 10$. For each k, we generate 100 data sets with covariates drawn from independent standard normals with means drawn from a uniform on the interval $[-10, 10]$. Then, for units designated as control, we add an independent draw for each covariate from a normal with mean zero and standard deviation 5.

We then define a set $\mathcal{M}$ of linear regression models that includes all possible specifications that include subsets of covariates, squared terms, and interactions, with squared terms and interactions included only if the main effects are included. We draw one model from $\mathcal{M}$ to define the true data generating process. We use this one true model to generate Y as a linear function of the treatment times its effect of 100, the covariates with coefficients drawn from a uniform distribution on the interval of [0,500], a constant term of 500, and a normal error term with mean 0 and standard deviation 500.

For each of the 100 data sets and each sample size, we run PSM and MDM, using all main effects only. To compute model dependence for a (matched) data set, we draw 1,000 models from $\mathcal{M}$, estimate the treatment effect for each as the coefficient on the treatment variable, and then compute the variance across these estimates. In order to have a comparable measure, the subset of 1,000 models is fixed across all runs (within a fixed k). We then average the standardized estimates of model dependence within each run, over the 100 runs, and plot scaled estimates.

Figure 7 gives our results, in parallel to previous figures, so that number of units pruned is on the horizontal axis and model dependence on the vertical axis. With PSM in red and MDM in blue, one panel appears for each k. Four patterns are apparent. First, PSM has higher levels of model dependence than MDM throughout all five graphs. Second, the advantage of MDM over PSM increases in all five graphs as more observations are pruned. Third, the PSM paradox is evident in all five graphs. And finally, a paradox, with more units pruned leading to higher levels of imbalance, also affects MDM in 10 dimensional space in the last graph (and to some small extent right at the end of the some of the others).

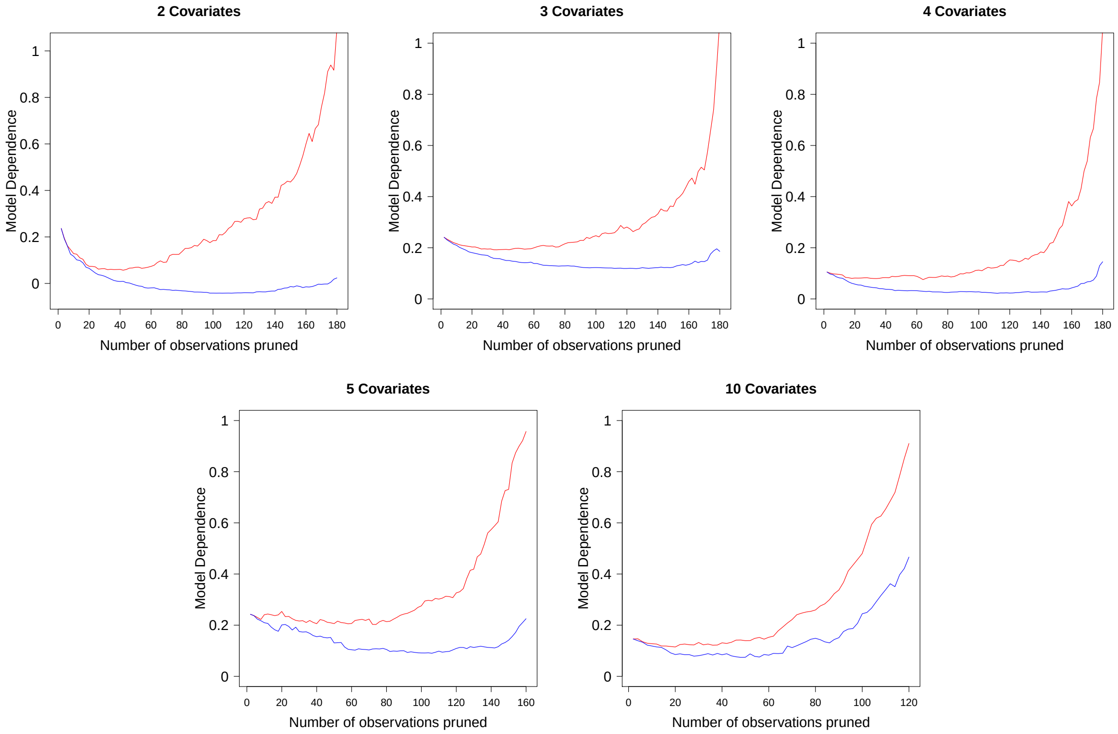

Figure 7: Model Dependence by Number of Covariates, with PSM in red and MDM in blue.

In all applications, with all matching methods, it pays to study the data, the units that are pruned, how much imbalance and model dependence is left, and whether the process of pruning is improving or degrading inferences.

6 Concluding Remarks

The important insight behind PSM is to analyze an observational data set by approximating as closely as possible a completely randomized experiment. However, when feasible, approximating a fully blocked randomized experiment can be substantially better than approximating a completely randomized experiment. The consequence of not doing so will in some situations merely mean that important information is left on the table — just as those who actually design experiments know to block on all available pre-treatment covariates whenever feasible to avoid wasting research resources, statistical power, and efficiency. However, in the case of PSM, the problem is not merely information discarded but the damage PSM creates by continuing to prune after it has nearly accomplished its goal of approximating a completely randomized experiment; in this situation, the PSM paradox will kick in and pruning observations will discard information and also increase imbalance, model dependence, researcher discretion, and the potential for bias.

Fortunately, these problems are usually easy to avoid by switching to one of the other popular methods of matching. However, the same paradox of matching increasing imbalance can occur with other methods when enough observations have been pruned to approximate full blocking. Although few researchers would prune observations that are exactly matched, it is important to not be fooled by problems with few observations being matched in very high dimensional space, where no matches may exist by any method.

In any matching method, the researcher should closely follow the advice in the literature about these other methods. For example, researchers should use information about the substance of the problem being analyzed and measurement characteristics of the variables included, such as encoded in coarsenings in CEM or data measurements in Euclidean distance matching, rather than the automatic standardization in MDM that is convenient for a methods paper like this.

An open question worth following up is whether the PSM paradox discussed here explains some of the difficulties scholars have noticed that PSM has caused, or not solved, in real data analyses. For example, Peikes, Moreno and Orzol (2008), Glazerman, Levy and Myers (2003), and Smith and Todd (125) have each pointed to PSM requiring many more observations than they expected as one source of PSM's problems, which is the difference one would experience when running a completely randomized experiment instead of a fully blocked randomized experiment.

References

Abadie, Alberto and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74(1):235–267.

Athey, Susan and Guido Imbens. 2015. "A Measure of Robustness to Misspecification." *American Economic Review Papers and Proceedings* .

Austin, Peter C. 2008. "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003." *Journal of the American Statistical Association* 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.

Banaji, Mahzarin R and Anthony G Greenwald. 2013. *Blindspot: Hidden biases of good people*. Random House LLC.

Box, George E.P., William G. Hunger and J. Stuart Hunter. 1978. *Statistics for Experimenters*. New York: Wiley-Interscience.

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.

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. 2004. Estimating Causal Effects in Nonexpermental Studies. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. Andrew Gelman and Xiao-Li Meng. New York: Wiley.

Diamond, Alexis and Jasjeet S Sekhon. 2012. "Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies." *Review of Economics and Statistics* 95(3):932–945.

Drake, C. 1993. "Effects of misspecification of the propensity score on estimators of treatment effects." *Biometrics* 49:1231–1236.

Efron, Bradley. 2015. "Estimation and accuracy after model selection." *Journal of the American Statistical Association* .

Finkel, Steven E, Jeremy Horowitz and Reynaldo T. Rojo-Mendoza. 2012. "Civic Education and Democratic Backsliding in the Wake of Kenya's Post-2007 Election Violence."

Journal of Politics 74(01):52–65.

Glazerman, Steve, Dan M. Levy and David Myers. 2003. “Nonexperimental versus experimental estimates of earnings impacts.” *The Annals of the American Academy of Political and Social Science* 589(September):63–93.

Greevy, Robert, Bo Lu, Jeffrey H. Silver and Paul Rosenbaum. 2004. “Optimal multivariate matching before randomization.” *Biostatistics* 5(2):263–275.

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.

Hansen, Ben B. 2004. “Full Matching in an Observational Study of Coaching for the SAT.” *Journal of the American Statistical Association* 99(467):609–618.

Hill, Jennifer. 2008. “Discussion of research using propensity-score matching: Comments on A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003by Peter Austin, Statistics in Medicine.” *Statistics in medicine* 27(12):2055–2061.

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. 2011*a*. “Causal Inference Without Balance Checking: Coarsened Exact Matching.” *Political Analysis* . http://gking.harvard.edu/files/abs/cem-plus-abs.shtml.

Iacus, Stefano M., Gary King and Giuseppe Porro. 2011*b*. “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. 2015. “A Theory of Statistical Inference for Matching Methods in Applied Causal Research.”. http://j.mp/Nt9TkZ.

Imai, Kosuke, Gary King and Clayton Nall. 2009. “The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation.” *Statistical Science* 24(1):29–53. http://gking.harvard.edu/files/abs/cluster-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.

Imai, Kosuke and Marc Ratkovic. 2014. “Covariate balancing propensity score.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1):243–263.

Imbens, Guido. 2000. “The role of the propensity score in estimating the dose-response functions.” *Biometrika* 87:706–710.

Imbens, Guido and Donald Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences An Introduction*. Cambridge University Press.

Imbens, Guido W. 2004. “Nonparametric estimation of average treatment effects under exogeneity: a review.” *Review of Economics and Statistics* 86(1):4–29.

Ioannidis, John PA. 2005. "Why most published research findings are false." *PLoS medicine* 2(8):e124.

Kahneman, Daniel. 2011. *Thinking, fast and slow*. Macmillan.

Kallus, Nathan. 2015. "Optimal A Priori Balance in The Design of Controlled Experiments.".

Kang, Joseph D. Y. and Joseph L. Schafer. 2007. "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." *Statistical Science* 22(4):523–539.

King, Gary and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2):131–159. http://gking.harvard.edu/files/abs/counterft-abs.shtml.

King, Gary and Langche Zeng. 2007. "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference." *International Studies Quarterly* (March):183–210. http://gking.harvard.edu/files/abs/counterf-abs.shtml.

Lechner, Michael. 2001. Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption. In *Econometric Evaluation of Labour Market Policies*, ed. M. Lechner and F. Pfeiffer. Heidelberg: Physica pp. 43–58.

Lunceford, Jared K and Marie Davidian. 2004. "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study." *Statistics in medicine* 23(19):2937–2960.

Mahoney, Michael J. 1977. "Publication prejudices: An experimental study of confirmatory bias in the peer review system." *Cognitive therapy and research* 1(2):161–175.

Mielke, P.W. and K.J. Berry. 2007. *Permutation Methods: A Distance Function Approach*. New York: Springer.

Morgan, Stephen L. and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research, 2nd edn*. Cambridge: Cambridge University Press.

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. 2010. "The foundations of causal inference." *Sociological Methodology* 40(1):75–149.

Peikes, Deborah N, Lorenzo Moreno and Sean Michael Orzol. 2008. "Propensity score matching." *The American Statistician* 62(3).

Robins, James M, Miguel Angel Hernan and Babette Brumback. 2000. "Marginal structural models and causal inference in epidemiology." *Epidemiology* 11(5):550–560.

Robins, JM and H Morgenstern. 1987. "The foundations of confounding in epidemiology." *Computers & Mathematics with Applications* 14(9):869–916.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.

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.

Rubin, Donald B. and Neal Thomas. 2000. "Combining propensity score matching with additional adjustments for prognostic covariates." *Journal of the American Statistical Association* 95:573–585.

Simmons, Joseph P, Leif D Nelson and Uri Simonsohn. 2011. "False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological science* 22(11):1359–1366.

Smith, Jeffrey A. and Petra E. Todd. 2005. "Does matching overcome LaLonde's critique of nonexperimental estimators?" *Journal of Econometrics* 125(1-2, March-April):305–353.

Smith, Jeffrey and Petra Todd. 125. "Rejoinder." *Journal of Econometrics* 2005:365–375.

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. "Matching with multiple control groups with adjustment for group differences." *Journal of Educational and Behavioral Statistics* . Forthcoming.

Tetlock, Philip E. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton: Princeton University Press.

Vansteelandt, Stijn and RM Daniel. 2014. "On regression adjustment for the propensity score." *Statistics in Medicine* 33(23):4053–4072.

Wilson, Timothy D and Nancy Brekke. 1994. "Mental contamination and mental correction: unwanted influences on judgments and evaluations." *Psychological bulletin* 116(1):117.

Zhao, Zhong. 2008. "Sensitivity of propensity score methods to the specifications." *Economic Letters* 98(3, March):309–319.