A Theory of Statistical Inference for Matching Methods in Causal Research*

Stefano M. Iacus† Gary King‡ Giuseppe Porro§

May 10, 2019

Abstract

Researchers who generate data often optimize efficiency and robustness by choosing stratified over simple random sampling designs. Yet, all theories of inference proposed to justify matching methods are based on simple random sampling. This is all the more troubling because, although these theories require exact matching, most matching applications resort to some form of ex post stratification (on a propensity score, distance metric, or the covariates) to find approximate matches, thus nullifying the statistical properties these theories are designed to ensure. Fortunately, the type of sampling used in a theory of inference is an axiom, rather than an assumption vulnerable to being proven wrong, and so we can replace simple with stratified sampling, so long as we can show, as we do here, that the implications of the theory are coherent and remain true. Properties of estimators based on this theory are much easier to understand and can be satisfied without the unattractive properties of existing theories, such as assumptions hidden in data analyses rather than stated up front, asymptotics, unfamiliar estimators, and complex variance calculations. Our theory of inference makes it possible for researchers to treat matching as a simple form of preprocessing to reduce model dependence, after which all the familiar inferential techniques and uncertainty calculations can be applied. This theory also allows binary, multicategory, and continuous treatment variables from the outset and straightforward extensions for imperfect treatment assignment and different versions of treatments.

*Our thanks to Alberto Abadie, Adam Glynn, Kosuke Imai, and Molly Roberts for helpful comments on an earlier draft. Replication code can be found here S. Iacus, King, and Porro, 2018.

†Department of Economics, Management and Quantitative Methods, University of Milan, Via Conservatorio 7, I-20124 Milan, Italy; stefano.iacus@unimi.it

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

§Department of Law, Economics and Culture, University of Insubria, Via S.Abbondio 12, I-22100 Como, Italy; giuseppe.porro@uninsubria.it

1 Introduction

Matching is a powerful nonparametric approach for improving causal inferences, especially in observational studies — that is, where assignment of units to treatment and control groups is not under the control of the investigator and not necessarily random. Matching is increasingly popular among applied researchers because it can be simple to apply and easy to understand. The basic idea is that certain serious statistical problems in a data set can be sidestepped by limiting inferences to a carefully selected subset. In particular, by reducing the strength of the relationship between pre-treatment covariates and the treatment assignment variable, statistical methods applied to the matched subset have reduced model dependence, estimation error, and bias (Cochran and Rubin, 1973; Ho, Imai, King, and Stuart, 2007; Rubin, 1974). By removing heterogeneous observations, matching can sometimes reduce variance but, when variance increases, the bias reduction usually more than compensates in typically large observational data sets. See Imbens, (2004), Morgan and Winship, (2014), and Stuart, (2010).

In this article, we discuss the theories of statistical inference that justify the statistical properties of estimators applied to matched data sets. We begin by observing that every theory of statistical inference involves an *axiom* about alternative realities where many hypothetical data sets could have been generated, and which we are supposed to imagine also generated our data, under the same conditions at the same moment in time. This data generation axiom can be modeled after how the one observed data set was actually drawn, on the theory that it is sometimes easier to imagine how hypothetical data sets might also have been generated. More common in the observational data sets to which matching is often applied, the data generation process is not known, and so researchers arbitrarily choose a data generation process for the observed and hypothetical data sets. In either case, these hypothetical realities are axioms that *define* the nature of our inferences, and the meaning of quantities such as standard errors, rather than being claims that could in principle be proven wrong. Stating the sampling axiom then clarifies the specific *assumptions* necessary for causal identification and unbiased estimation, which of course can be violated and which thus need to be justified by researchers. Applied researchers

must therefore understand that the specific assumptions to be justified, and how they may be justified, depends on this data generation axiom.

Until now, theories of statistical inference discussed in the literature on matching use axiom that the data are generated by *simple random sampling*, where each population unit has the same probability of selection (e.g., Abadie and Imbens, 2006). This is a simple-to-understand data generation process but, under finite sample inference, turns out to require that treated and control units match exactly on all measured pre-treatment covariates or on the propensity score (Lechner 2001, Imbens 2000, and Imai and Dyk 2004) — conditions impossible to meet in finite data with at least one continuous variable. In practice, empirical analysts routinely violate this exact matching requirement by applying various forms of approximate matching. Interestingly, they do this within a simple random sampling framework by stratifying ex post on the original covariate space, or a propensity score or distance metric on that space, and treating approximate matches within strata as if they were exact. Unfortunately, the assumptions necessary to make this procedure appropriate are virtually never discussed or justified by those implicitly using them. In other words, theorists assume no stratification in repeated sampling when they are being explicit about their theory of inference, but they actually do assume stratification in almost all real applications implicitly during applied data analyses.

In this article, we bring stratification into a theory of causal inference for matching in an explicit and visible way. Instead of burring the assumption ex post in the data analysis stage, we include it ex ante via an alternative formally stated axiom about the data generating process following a stratified sampling framework. We then make explicit all the assumptions necessary for valid causal inference given this axiom, which must be followed by researchers if they are to proceed as they analyze data as they do now. Because the strata under this theory are defined explicitly, ex ante, and in the space of the investigator's original variables, rather than ex post on the basis of more complicated derived variables like a propensity score or standardized (Mahalanobis) distance, it is easier to understand and, as with the congruence principle (Mielke and Berry, 2007), more intuitive and statistically robust.

Other theories of inference that work well in a stratified framework, like the one we propose, include novel finite sample approaches based on Neyman's randomization-based theory (Imai, 2008) and Fisher's permutation-based inference (Rosenbaum, 1988). These are not as easy to use as the stratified theory we propose, but easier than those based on simple random sampling. Alternatively, one can use asymptotic results which, in addition to the approximations necessary, unfortunately also must assume that the observational data grows in size at given arbitrary rates that depend upon the number of continuous covariates (Abadie and Imbens, 2012). These alternative approaches can be of value in some instances, but none allow researchers the convenience of using whatever point and uncertainty estimates they might have without a prior matching step.

Section 2 outlines our theory of statistical inference for matching based on stratified random sampling, and Section 3 gives the properties of estimators that satisfy it. We discuss what can go wrong in applications in Section 4. Then, in Section 5, we work through a real data set to show how using matching methods designed for simple random sampling are, as used, implicitly allowing for approximate matching, and how this step leads to uncontrolled imbalance and bias. This section also shows that by choosing directly the stratified random sampling matching theory of this paper, researchers can estimate the same treatment effect without hiding the approximation step. Section 6 concludes. Appendix A gives the proofs and Appendix B extends the theory to situations where the true and observed treatment status diverge and where different versions of treatment are evident.

2 Causal Inference under Stratified Random Sampling Theory

2.1 Data Generation Process

Theories of statistical inference require an axiom about the assumed data generation process in hypothetical repeated samples. In the matching literature, existing theories of inference for matching assume (usually implicitly) simple random sampling, which we define formally as follows:

Axiom A0' [Simple Random Sampling]**:** *Consider a population of units Θ with covariates $\mathcal{X}$. Draw repeated hypothetical samples, of fixed size $n < \infty$, at random from this population (i.e., so that each sample of n observations has equal probability of selection).*

In this article, we offer a new theory of inference for matching that replaces Axiom A0' with an axiom based on stratified random sampling. Stratification is a well known technique in statistics that has had a role in matching since at least Cochran, (1968) (see also Rubin, 1977). To ease the exposition below, we denote by $\mathcal{X}$ the space of pre-treatment covariates and offer a formal definition of stratification as:

Definition 1. *Let $\Pi(\mathcal{X})$ be a finite partition of the covariate space $\mathcal{X}$, and let $A_k \in \Pi(\mathcal{X})$ $(k = 1, \ldots, K < \infty)$ be one generic set of the partition, i.e. $\cup_k A_k = \mathcal{X}$ and $A_l \cap A_m = \emptyset$ for $l \neq m$.*

For example, suppose that $\mathcal{X}$ consists of the variables `age`, `gender` and `earnings`, i.e. $\mathcal{X} = \{\texttt{age}, \texttt{gender}, \texttt{earnings}\}$. Then $\Pi(\mathcal{X})$ can be interpreted as the product (space) of variables $\texttt{age} \times \texttt{gender} \times \texttt{earnings} = \Pi(\mathcal{X})$. Therefore, in the example, one of the sets A_k might be the subset of "young adult males making greater than $25,000", i.e, $A_k = \{\texttt{age} \in (18, 24]\} \times \{\texttt{gender} = M\} \times \{(\texttt{earnings} > \$25000)\}$. When no ambiguity is introduced, we drop the subscript k from A_k. Stratified random sampling involves random sampling from within strata A with given quotas proportional to the relative weight of the strata $\{W^A, A \in \Pi(\mathcal{X})\}$.

Finally, we offer our alternative data generating process axiom:

Axiom A0 [Stratified Random Sampling]**:** *Consider a population of units Θ, and denote the space of covariates as $\mathcal{X}$. Let $\Pi(\mathcal{X})$ be a partition of $\mathcal{X}$ that stratifies Θ into disjoint subpopulations of units. Let $\{W^A, A \in \Pi(\mathcal{X})\}$ be fixed weights for the strata. Draw repeated hypothetical samples of $[n \cdot W^A]$ observations, $n < \infty$, via simple random sampling (defined in Axiom A0', above) in each stratum $A \in \Pi(\mathcal{X})$, so that the total number of observations is n (and where $[x]$ is the integer part of x).*

In this alternative Axiom A0, the strata and the total number of observations for each hypothetical repeated sample and the observed sample are fixed. Then, the data set within

each stratum is drawn according to simple random sampling from Axiom A0'.[1]

The axioms described in Axioms A0 and A0' cannot be proven true or false on the basis of comparisons to a single observed data set, arguments about plausibility, or information about how matching methods are used. Because the repeated samples are strictly hypothetical, A0 and A0' are not even statements that could be true or false in principle. Instead, the choice of an axiom merely defines how to interpret one's causal inferences and uncertainty estimates, the specific type of repeated hypothetical samples, and the ultimate inferential target. As all matching methods use some kind of stratification of the covariates $\mathcal{X}$, Axiom A0 highlights this fact and clarifies the theoretical assumptions necessary for valid inferences, rather than, as under Axiom A0', keeping it hidden and left to applied researchers to deal with outside of the process of statistical inference.

2.2 Treatment Assignment

Consider now the data generated in Axiom A0, where subject i ($i = 1, \ldots, n$) has been exposed to treatment $T_i = t$, for $t \in \mathcal{T}$, where $\mathcal{T}$ is either a subset of $\mathbb{R}$ or a set of (ordered or unordered) categories, T_i is a random variable, and t one possible value of it. Then $\mathcal{Y} = \{Y_i(t) : t \in \mathcal{T}, i = 1, \ldots, n\}$ is the set of *potential outcomes*, the possible values of the outcome variable when T takes on different values. For each observation, we observe one and only one of the set of potential outcomes, that for which the treatment was actually assigned: $Y_i \equiv Y_i(T_i)$. In this setup, T_i is a random variable, the potential outcomes are fixed constants for each value of T_i, and $Y_i(T_i)$ is a random variable, with randomness stemming solely from the data generation process for T determining which of the potential outcomes is observed for each i. Let X_i be the $p \times 1$ vector ($X \in \mathcal{X}$) of pre-treatment covariates for subject i.[2]

[1] Let $M_j^A = \{i : T_i = t_j, X_i \in A\}$ be the set of indexes of all observations for treatment level $T_i = t_j$ within stratum $A \in \Pi(\mathcal{X})$ and $M_j = \bigcup_{A \in \Pi(\mathcal{X})} M_j^A$ be the set of all indexes of the observations corresponding to treatment $T = t_j$. Denote the number of observations in each set by $m_j^A = |M_j^A|$ and $m_j = |M_j|$, respectively and define the weights introduced in Axiom A0 as $W_j^A = m_j^A/m_j$, $j = 1, 2$. We assume that, in our stratified random sampling data generation process, the proportions W_j^A are fixed across repeated samples, and hence the weights in A0 are defined by $W^A = (m_1^A + m_2^A)/(m_1 + m_2)$ for $A \in \Pi(\mathcal{X})$.

[2] We can clarify Axioms A0 and A0' by giving a contrasting axiom where the repeated hypothetical sampling distributions are based on the use of the randomized treatment assignment mechanism. This

2.3 Treatment Effect

Let t_1 and t_2 be distinct values of T that happen to be of interest, regardless of whether T is binary, multicategory, or continuous (and which, for convenience, we refer to as the treated and control conditions, respectively). Assume T is observed without error (an assumption we relax in Appendix B). Define the *treatment effect* for each observation as the difference between the corresponding two potential outcomes, $\text{TE}_i = Y_i(t_1) - Y_i(t_2)$, of which at most only one is observed (this is known as the "Fundamental Problem of Causal Inference"; Holland 1986). (Problems with multiple or continuous values of treatment variables have multiple treatment effects for each observation, but the same issues apply.)

The object of statistical inference is usually an average of treatment effects over a given subset of observations. Researchers then usually estimate one of two types of quantities. The first is the sample average treatment effect on the treated, for which the potential outcomes and thus TE_i are considered fixed, and inference is for all treated units in the sample at hand: $\text{SATT} = \frac{1}{|M_1|}\sum_{i \in M_1} \text{TE}_i$, with the control units used to help estimate this quantity (Imbens, 2004, p.6). Other causal quantities of this first type are averaged over different subsets of units, such as from the population, the subset of the population similar to X, or all units in the sample or population regardless of the value of T_i. Since a good estimate of one of these quantities will usually be a good estimate of the others, usually little attention is paid to the differences for point estimation, although there may be differences with respect to uncertainty estimates under some theories of inference (Imai, 2008; Imbens and Wooldridge, 2009).

The second type of causal quantity is when some treated units have no acceptable

axiom is used for Fisher's permutation-based inference of sharp null hypotheses (Rosenbaum, 1988) and Neyman's randomization-based theory for average treatment effects (Imai, 2008) (See also Ding, (2016).)

Axiom A0" [Randomized Treatment Assignment]**:** *Consider an observed data set of n observations, with treated variable $T_i \in \{0, 1\}$, covariates $X_i \in \mathcal{X}$, outcome Y_i, and $i = 1, \ldots, n$. Define hypothetical repeated samples that reassign the vector of values of T by randomly drawing a permutation of T, such that each of the $n!$ possible permutations have equal probability of selection.*

We focus on developing a theory of inference around the use of Axiom A0, and so do not use Axiom A0" further. Nevertheless, the differences among these three axioms help clarify the meaning of each and to suggest potential avenues for future research.

matches among a given control group and so are pruned along with unmatched controls, a common situation which gives rise to "feasible" versions of SATT (which we label FSATT) or of the other quantities discussed above. This formalizes the common practice in many types of observational studies by focusing on quantities that can be estimated well (perhaps in addition to estimating a more model dependent estimate of one of the original quantities) (see Crump, Hotz, Imbens, and Mitnik, 2009; S. M. Iacus, King, and Porro, 2011; Rubin, 2010), an issue we return to in Section 3.2. (In multi-level treatment applications, the researcher must choose whether to keep the feasible set the same across different treated units so that direct comparison of causal effects is possible, or to let the sets vary to make it easier to find matches.)

2.4 Assumptions

We now describe Assumptions A1–A3, which establish the theoretical background needed to justify valid causal inference under finite data with stratified random sampling as defined in Axiom A0; this set of Assumptions can be seen as a natural strata-wide extension of the pointwise theory by Rosenbaum and Rubin, (1983) which differs because it builds off of Axiom A0' instead.

The first assumption (which we generalize in Appendix B) helps to precisely define the variables used in the analysis:

Assumption A1 [SUTVA: Stable Unit Treatment Value Assumption (Rubin, 1980, 1990, 1991)]**:** *A complete representation of all potential outcomes is* $\mathcal{Y} = \{Y_i(t) : t \in \mathcal{T}, i = 1, \ldots, n\}$.

SUTVA can be interpreted in at least three ways (see VanderWeele and Hernan, 2012). First is "logical consistency," which connects potential outcomes to the observed values and thus rules out a situation where say $Y_i(0) = 5$ if $T_i = 1$ but $Y_i(0) = 12$ if $T_i = 0$ (Robins, 1986). Second is "no interference," which indicates that the observed value T_i does not affect the values of $\{Y_i(t) : t \in \mathcal{T}\}$ or $\{Y_j(t) : t \in \mathcal{T}, \forall\ j \neq i\}$ (Cox, 1958). And finally, SUTVA requires that the treatment assignment process produce one potential outcome value for any (true) treatment value (Neyman, 1935).

To use our theory to justify a matching method requires that the information in these strata, and the variables that generate them, be taken into account. The theory does not require that our specific formalization of these strata be used in a matching method, only that the information is controlled for in some way. This can be done by directly matching on A, using some function of A in covariates to control for, or some type of weighting that takes account of A. An example is given in Section 5.

We now introduce the second assumption, which ensures that the pre-treatment covariates defining the strata are sufficient to adjust for any biases. (This assumption serves the same purpose as the "no omitted variable bias" assumption in classical econometrics, but without having to assume a particular functional form.) Thus, by conditioning on the values of X encoded in the strata A, we define:

Assumption A2 [Set-wide Weak Unconfoundedness]**:** $T \perp Y(t)|A$*, for all* $t \in \mathcal{T}$ *and each* $A \in \Pi(\mathcal{X})$.

For example, under A2, the distribution of potential outcomes under control $Y(0)$ is the same for the unobserved treated units and as the observed control units; below, this will enable us to estimate the causal effect by using the observed outcome variable in the control group.

Apart from the sampling framework, Assumption A2 can be thought of as a degenerate version of the Conditioning At Random (CAR) assumption in Heitjan and Rubin, (1991) with conditioning fixed. CAR was designed to draw inferences from coarsened data, when the original uncoarsened data are not observed. In the present framework, $\Pi(\mathcal{X})$ represents only a stratification of the reference population and each stratum A in that definition is fixed in repeated sampling. A special case of Assumption A2, with sets A fixed to singletons (i.e. taking $A = \{X = x\}$), is known as "weak unconfoundedness" used under exact matching theory (Imai and Dyk, 2004; Imbens, 2000; Lechner, 2001) and first articulated in Rosenbaum and Rubin, (1983).

Finally, any matching theory requires a version of the *"common support"* assumption, i.e. for any unit with observed treatment condition $T_i = t_1$ and covariates $X_i \in A$, it is also *possible* to observe a unit with the counterfactual treatment, $T_i = t_2$, and the covariate

values in the same set A. This is the assumption that rules out, for example, being able to estimate the causal effect of United Nations interventions in civil wars on peace building success when the UN intervenes only when they are likely to succeed (King and Zeng, 2006). In less extreme cases, it is possible to narrow the quantity of interest to a portion of the sample space (an thus the data) where common support does exist. More formally, we introduce this version that works under the stratified random sampling Axiom A0:

Assumption A3 [Set-wide Common Support]**:** *For all measurable sets* $B \in \mathcal{T}$ *and all sets* $A \in \Pi(\mathcal{X})$ *we have* $p(T \in B | X \in A) > 0$.

Assumptions A2 and A3 make the search for counterfactuals easier since all observations in the vicinity of (i.e., in the same strata as) a unit, rather than only those with exactly the same covariate values, are now acceptable matches. (The combination of the pointwise versions of both A2 and A3 is often referred as "strong ignorability" (Abadie and Imbens, 2002; Rosenbaum and Rubin, 1983).) Assumption A3 also requires that at least one treated and one control unit (or one in each treatment regime) appear within every stratum, and so A3 imposes constraints on the weights.

2.5 Identification of the Treatment Effect

We show here that Assumptions A1-A3 enable point identification of the causal effect in the presence of approximate matching. Identification for the expected value of this quantity can be established under the new assumptions by noting, for each $A \in \Pi(\mathcal{X})$, that

$$E\{Y(t)|A\} \overset{\mathbf{A2}}{=} E\{Y(t)|T = t, A\} = E\{Y|T = t, A\},$$

which means that within set A_k, we can average over the observed Y corresponding to the observed values of the treatment T rather than unobserved potential outcomes for which the treatment was not assigned. The result is that the average causal effect within the set A, which we denote by τ^A, can be written as the difference in two means of observed variables, and so is easy to estimate:

$$\tau^A = E\{Y(t_1) - Y(t_2)|A\} = E\{Y|T = t_1, A\} - E\{Y|T = t_2, A\}, \tag{1}$$

for any $t_1 \neq t_2 \in \mathcal{T}$. That is, (1) simplifies the task of estimating the causal effect in approximate matching in that it allows one to consider the means of the treated and control groups separately, within each set A, and to take the weighted average over all strata $A \in \Pi(\mathcal{X})$ afterwards. To take this weighted average, we use Assumption A3:

$$E(Y(t)) \overset{\mathbf{A3}}{=} E(E\{Y(t)|A\}) \tag{2}$$

which is exactly what we need to calculate the average causal effect $\tau = E(Y(t_1)) - E(Y(t_2))$. Assumption A3 is required because otherwise $E\{Y(t)|A\}$ may not exist for one of the two values of $t = t_1$ or $t = t_2$ for some stratum A, in which case $E(Y(t))$, would not exist and the overall causal effect would not be identified.

3 Properties of Estimators After Matching

Current estimation practice after one-to-one matching involves using estimators for the difference in means or with regression adjustment that follows matching. In j-to-k matching for $j > 0$ and $k > 1$ varying over units, the same procedures are used after averaging within strata for treatment and control groups or, equivalently, without strata but with unit-level weights. Either way, the same estimation procedures that might have been used without matching can now be used as is, along with familiar uncertainty estimates and diagnostic techniques. We now give some details of how our theory of inference justifies these simple procedures.

3.1 Difference in Means Estimator

To describe the property of the estimators, we adapt the notation of Abadie and Imbens, (2011) (which operates under axiom A0') and rewrite the causal quantity of interest as the weighted sum computed within each stratum A from (1):

$$\begin{aligned}\tau &= \frac{1}{m_1}\sum_{i\in M_1} E\{\text{TE}_i\} = \frac{1}{m_1}\sum_{A\in\Pi(\mathcal{X})}\sum_{i\in M_1^A} E\{Y_i(t_1) - Y_i(t_2)|X_i \in A\} \\ &= \frac{1}{m_1}\sum_{A\in\Pi(\mathcal{X})}\sum_{i\in M_1^A}(\mu_1^A - \mu_2^A) = \frac{1}{m_1}\sum_{A\in\Pi(\mathcal{X})}(\mu_1^A - \mu_2^A)m_1^A = \sum_{A\in\Pi(\mathcal{X})}\tau^A W_1^A,\end{aligned} \tag{3}$$

where $\mu_k^A = E\{Y(t_k)|X \in A\}$ $(k = 1, 2)$ and τ^A is the treatment effect within set A as in (1). Consider now an estimator $\hat{\tau}$ for τ based on this weighted average:

$$\hat{\tau} = \sum_{A \in \Pi(\mathcal{X})} \hat{\tau}^A W_1^A = \frac{1}{m_1} \sum_{i \in M_1^A} (Y_i(t_1) - \hat{Y}_i(t_2)) \tag{4}$$

where $\hat{\tau}^A$ is the simple difference in means within the set A, i.e.:

$$\begin{aligned} \hat{\tau}^A &= \frac{1}{m_1^A} \sum_{i \in M_1^A} \left(Y_i - \hat{Y}_i(t_2) \right) = \frac{1}{m_1^A} \sum_{i \in M_1^A} \left(Y_i - \frac{1}{m_2^A} \sum_{j \in M_2^A} Y_j \right) \\ &= \frac{1}{m_1^A} \sum_{i \in M_1^A} Y_i - \frac{1}{m_2^A} \sum_{j \in M_2^A} Y_j. \end{aligned} \tag{5}$$

Finally, we have the main result (see the appendix for a proof):

Theorem 1. *The estimator $\hat{\tau}$ is unbiased for τ.*

Given that the sets of the partition $\Pi(\mathcal{X})$ are disjoint, it is straightforward to obtain the variance $\sigma^2_{\hat{\tau}} = \mathrm{Var}(\hat{\tau})$ of the causal effect. If we denote by $\sigma^2_{\hat{\tau}^A}$ the variance of the stratum-level estimates $\hat{\tau}^A$ in (5), we have $\sigma^2_{\hat{\tau}} = \sum_{A \in \Pi(\mathcal{X})} \left(\sigma_{\hat{\tau}^A} W_1^A \right)^2$.

3.2 Simplified Inference Through Weighted Least Squares

The direct approach to estimating the treatment effect by strata and then aggregating is useful to define the matching estimator, but it is more convenient to rewrite the estimation problem in an equivalent way as a weighted least squares problem. This approach provides a easy procedure for computing standard errors, even for multi-level treatment (see Section 3.3) or when one or more strata contain only one treated unit and one control unit (see Section 4. In this latter case, one cannot directly estimate the variance within the strata $\sigma^2_{\hat{\tau}^A}$ but we can still obtain an estimate of it by applying whatever estimator one would have applied to the data set without matching.

We now introduce the weights we use to simplify the estimator in (4) and re-express it as the difference in weighted means. For all observations, define the weights w_i as

$$w_i = \begin{cases} 1, \text{ if } T_i = t_1, \\ 0, \text{ if } T_i = t_2 \text{ and } i \notin M_2^A \text{ for all } A, \\ \frac{m_1^A}{m_2^A} \frac{m_2}{m_1}, \text{ if } T_i = t_2 \text{ and } i \in M_2^A \text{ for one } A. \end{cases} \tag{6}$$

Then, the estimator $\hat{\tau}$ in (4) can be rewritten as

$$\hat{\tau} = \frac{1}{m_1} \sum_{i \in M_1} Y_i w_i - \frac{1}{m_2} \sum_{j \in M_2} Y_j w_j,$$

where the variance is the sum of the variances of the two quantities. Therefore, the standard error of $\hat{\tau}$ is the usual standard error of estimates for regression analysis with weights. For example, consider the linear regression model:

$$Y_i = \beta_0 + \beta_1 T_i + \epsilon_i, \quad \epsilon_i \sim N(0,1) \text{ and i.i.d}$$

where $\hat{\tau} \equiv \hat{\beta}_1$ if weights w_i are used in estimation. So the standard error of $\hat{\beta}_1$ can be obtained as the output of this weighted least squares (WLS) model, and is the correct estimate of $\sigma^2_{\hat{\tau}}$. Other models, such as GLM with weights, can also be estimated in a similar fashion. The only change that needs to be made to the estimator without matching is to include these weights.

3.3 Estimation with Multi-level Treatments

For more than two treatments we define the multi-treatment weights as

$$w_i(k) = \begin{cases} 1, \text{ if } T_i = t_1, \\ 0, \text{ if } T_i = t_k \text{ and } i \notin M_k^A \text{ for all } A, \\ \frac{m^A}{m_k^A} \frac{m_k}{m_1}, \text{ if } T_i = t_k \text{ and } i \in M_k^A \text{ for one } A. \end{cases}$$

Then, for each $k = 2, 3, \ldots$, the treatment effect $\tau(k)$ can be estimated as $\hat{\beta}_1(k)$ in

$$Y_i = \beta_0 + \beta_1(k) T_i + \cdots + \epsilon_i$$

with weights $w_i(k)$ and, again, the usual standard errors are correct as is.

3.4 Additional Regression Adjustment for Further Covariates

If Assumption A2 holds, then adjusting for covariates is unnecessary to ensure unbiasedness. If Assumption A2 holds but the analyst is unsure if it does, and so adjusts for pre-treatment covariates (with interactions), then the downside is trivial (Lin, 2013; Miratrix, Sekhon, and Yu, 2013). But sometimes, the researcher may need to adjust for covariates via a model, even if they were not used during matching. In this situation, it sufficient

to proceed as one would without the matching step by including all the variables in the regression model

$$Y_i = \beta_0 + \beta_1 T_i + \gamma_1 X_{i1} + \cdots \gamma_d X_{id} + \epsilon_i.$$

and using the weights in (6) to run the WLS regression. The estimated coefficient $\hat{\beta}_1$ is then the estimator of the treatment effect $\hat{\tau}$ and its standard error is an unbiased estimator of its standard deviation, under the model.

3.5 Defining Strata in Observational Data

One question that may arise in this framework, as in any stratified sampling, is how to choose the strata $A \in \Pi(\mathcal{X})$ in a given problem? The answer is by definition application-specific, which can be an advantage in that it relies on variables in the original investigator-defined units of measurement, reflecting knowledge the investigator must have.

To show the applicability of our approach in observational studies, we take advantage of the fact that in many data sets variables referred to as "continuous" in fact often have natural breakpoints that may be as or more important than the continuous values. These may include grade school, high school, and college degrees for the variable "years of education"; the official poverty level for the variable "income"; or puberty, official retirement age, etc., for the variable "age". This understanding of measurement recognizes that, for another example, 33° Fahrenheit may be closer to 200° than to 31°, at least for certain purposes. Most data analysts not only know this distinction well but use it routinely to collapse variables in their ordinary data analyses. Indeed, in analyses of sample surveys, continuous variables with no natural breakpoints, or where authors never use breakpoints to collapse variables or categories, are uncommon.

For another example, consider estimating the causal effect of the treatment variable "taking one introductory statistics course" on the outcome variable "income after college", and where we also observe one pre-treatment covariate "years of education", along with its natural breakpoints at high school and college degrees. Assumption A2 says that it is sufficient to control for the coarsened three-category education variable (no high school degree, high school degree and possibly some college courses but no college degree, and

college degree) rather than the full "years of education" variable. In this application, A2 is plausible if, as seems common, employers at least at first primarily value degree completion in setting salaries. Then, post-stratification and matching within the strata is appropriate. If, instead, a major difference in expected income exists between those who have, for example, one versus three years of college, then there can be some degree of bias induced.

4 How to Avoid Violating Assumptions A2 and A3

When a data set has at least one stratum A that does not contain all levels of the treatment, the now prevalent view in the literature is that the best approach is to change the quantity of interest and switching from SATT to FSATT, where we use only strata where A3 is satisfied (Crump, Hotz, Imbens, and Mitnik, 2009; S. M. Iacus, King, and Porro, 2011; Rubin, 2010). Yet, this absence of evidence for A3 does not necessarily imply that the assumption itself is false; it could instead have been the case that we happen not to have sufficient samples from those strata.

In the situation when switching to FSATT is not an option, because only an inference about the original quantity of interest will do, bias may arise if, for example, we merge two or more strata into a new larger strata, match within this larger strata, and violate A2, and possibly also A3. This same issue arises under stratified sampling A0 as under simple random sampling A0', but we discuss how to think about it under stratified sampling in this section.

4.1 How Bias Arises?

To understand where bias may arise under Axiom A0 when some strata A need to be enlarged or changed, we study the following bias decomposition, by adapting ideas designed to work under Axiom A0' from Abadie and Imbens, (2006, 2011, 2012). Let $\mu_t(x) = E\{Y(t)|X = x\}$ and $\mu(t_k, x) = E\{Y|X = x, T = t_k\}$. Under Assumption A2 we know that $\mu_{t_k}(x) \overset{\mathbf{A2}}{=} \mu(t_k, x) \equiv \mu_k^A$ for all $\{X = x\} \subseteq A$. Then the bias is written as:

$$\hat{\tau}^A - \tau^A = \sum_{A \in \Pi(\mathcal{X})} \left\{ (\bar{\tau}^A - \tau^A) + E^A + B^A \right\} W_1^A,$$

where

$$\bar{\tau}^A = \frac{1}{m_1^A} \sum_{i \in M_1^A} (\mu_{t_1}(X_i) - \mu_{t_2}(X_i))$$

$$E^A = \frac{1}{m_1^A} \sum_{i \in M_1^A} \left((Y_i - \mu_{t_1}(X_i)) - \frac{1}{m_1^A} \sum_{i \in M_1^A} \frac{1}{m_2^A} \sum_{j \in M_2^A} (Y_j - \mu_{t_2}(X_j)) \right)$$

and

$$B_A = \frac{1}{m_1^A} \sum_{i \in M_1^A} \frac{1}{m_2^A} \sum_{j \in M_2^A} (\mu_{t_2}(X_i) - \mu_{t_2}(X_j))$$

where $\mu_{t_k}(X) = \mu_k^A$ for $X \in A$. Therefore, both $(\bar{\tau}^A - \tau^A)$ and E^A have zero expectation inside each set A and $B^A = 0$. But if some of the sets A' are different from the original partition A, or combined or enlarged, then assumption A2 may not apply any longer, in general, $\mu_{t_k}(X) \neq \mu_k^A$ for $X \in A' \neq A$.

4.2 Nonparametric Regression Adjustment

One way to proceed is with the following regression adjustment, as in Abadie and Imbens, (2011), that compensates for the bias due to the difference between A and A'. Let $\hat{\mu}_{t_2|A}(x)$ be a (local) consistent estimator of $\mu_{t_2}(x)$ for $x \in A$. In this case, one possible estimator is the following

$$\hat{\tau}^A = \frac{1}{m_1^A} \sum_{i \in M_1^A} (Y_i - \hat{\mu}_{t_2|A}(X_i)) - \frac{1}{m_2^A} \sum_{j \in M_2^A} (Y_j - \hat{\mu}_{t_2|A}(X_j)). \tag{7}$$

This estimator is asymptotically unbiased if the number of control units in each stratum grows at the usual rate. If instead of using a local estimator $\hat{\mu}_{t_2|A}(x)$ we use a global estimator $\hat{\mu}_{t_2}(x)$, i.e. using all control units in the sample as in Abadie and Imbens, (2011), then the calculation of the variance of the estimator is no longer obtained by simple weighting and the validity of the approach requires a treatment similar to the asymptotic theory of exact matching. More technical assumptions and regularity on the unknown functions $\mu_t(x)$ are needed to prove that the regression type estimator in (7) can compensate for the bias asymptotically but, essentially, it is required that, for some $r \geq 1$, we impose $m_1^r/m_2 \to \kappa$, with $0 < \kappa < \infty$. A simplified statement is that $m_1/m_2^{4/k} \to 0$,

where k is the number of continuous covariates in the data and this condition is equivalent to $m_1^{k/4}/m_2 = m_1^r/m_2 \to \kappa$. The proof of these results can be found in Abadie and Imbens, (2011).

4.3 Asymptotic Filling of the Strata

If Assumption A3 is apparently violated because there are not enough observations in one or more strata, but we still believe A1–A3 to be true and we happen to be able to continue to collect data, then it is worth knowing that it is theoretically possible to fill all the strata in $\Pi(\mathcal{X})$ and obtain unbiased estimates of the treatment effect. This is theoretically possible under the additional assumption that $m_1^r/m_2 \leq \kappa$, with $0 < \kappa < \infty$, $r > k$, and k the number of continuous covariates. By Proposition 1 in Abadie and Imbens, (2012), all the strata A will be filled with probability one. This result is enough to obtain asymptotically unbiased estimates of the causal effect under the original assumptions A2–A3, without changing the initial partition $\Pi(\mathcal{X})$ or other technical smoothness assumptions on the functions $\mu_t(x)$ and $\hat{\mu}_{t|A}(x)$. As such, one could use an asymptotic approximation to obtain estimates and standard errors, but it is considerably safer to use these results as a guide to future data collection.

5 Approximate Matching in Practice

In this section, we apply commonly used matching methods to the same real data set in order to highlight five important points[3]. First, we emphasize how the application of all matching methods, in almost all real data sets, require approximations that may violate the corresponding theory of inference. Second, the assumptions do not fail gracefully: Even small deviations from the requirements of any theory of inference can yield large biases or misinterpretations. Third, there is reason to believe that our alternative (stratified random sampling) theory will often be more robust to incorrect approximations than existing (simple random sampling) theories. And finally, common usage of some existing theories of inference typically ignore the essential approximations, making it difficult

[3]Replication code can be found here S. Iacus, King, and Porro, 2018.

or impossible for applied researchers in most situations to apply the theory with fidelity. Applied researchers typically march forward anyway, inappropriately burying the approximations, and the assumptions necessary to make the theories valid, usually without comment as part of commonly used data analysis practices. In contrast, under our alternative stratified-based theory of inference, all necessary assumptions are stated explicitly, up front, and before any data analysis. These assumptions, and any deviations from them, are also considerably easier to understand and use under our alternative than under existing theories. Finally, as emphasized in previous sections, the choice of stratified sampling in Axiom A0 vs simple random sampling in Axiom A0' is a statement about a hypothetical sampling process, rather than a claim that can be proven right or wrong. In this situation, the critical task for the analyst is to completely *understand* how the theory of inference is applied in the context of their data, and to interpret it correctly, rather than to justify whether it is correct, "plausible," or appropriate for an application. As such, the far greater simplicity of the stratified over simple random sampling theory can be a major advantage.

Data For data, we consider the National Supported Work (NSW) training program used in the seminal paper by Lalonde, (1986). In these data, the outcome variable is the real earnings of workers in 1978 (`re78`) and the treatment is the participation in the program. Pre-treatment control variables include `age`; years of `education`; indicators for `black` and `hispanic`; an indicator for marital status, `married`; an indicator for not possessing a high school degree (`nodegree`); earnings in 1975, `re75`, and 1974, `re74`; and unemployment status in both years, `u74` and `u75`. The data set contains 297 individuals exposed to treatment and 425 control units. This is in fact an experimental sample, although Lalonde, (1986) analyzed it as an observational study to provide insights about matching methods. The quantity of interest is the sample average treatment effect on the treated (ATT), the increase in earnings in 1978 due to treatment.

Applying Simple Random Sampling-Based Theory We now show how three ways of satisfying the existing simple random sampling-based theory of inference all fail in these

data. We begin with the simple random sampling Axiom A0', and assume SUTVA, along with pointwise unconfoundedness and common support.

First, we apply exact matching theory, the hardest to satisfy but best case scenario, requiring exact matching on X. In most applications with rich sets of covariates, few matches are available. In our application, we do happen to have 55 treated units that exactly match 74 control observations (this occurs because the otherwise continuous variables, `re74` and `re75`, have a point mass at zero, and other variables are discrete). Unfortunately, this small sample would leave confidence intervals on the ATT too wide to make useful inferences.

More generally, this exact matching approach may work for very large samples, when there is high probability that match occurs without replacement for one-to-one nearest neighbors matching, and imbalance is zero (or very small according to some distance). In this situation, a simple regression model, with pre-treatment covariates and a treatment indicator, will normally be able to take into account the remaining bias in either simple or stratified random sampling.

Second, we consider exact matching on the propensity score. If successful, this approach would yield less efficient estimates than exact matching on X, and would introduce a variety of other serious problems, but causal estimates would at least be ex ante unbiased (King and Nielsen, 2017). To try this, we use the propensity score specification in Dehejia and Wahba, (1999), a logistic model for all the indicator variables, as well as `age`, `education`, `re74`, and `re75` and their squares. Unfortunately, as is typical in data sets with continuous covariates, lowering the bar for what constitutes a match in this way buys us *zero* additional matched observations. This is not a surprise, since propensity score matching requires exact matching on the propensity score, which does not happen with any higher probability than exact matching on X as long as we have some continuous variables.

A final option to follow existing theory would be to have a very large data set. Although we do not have a large data set, in observational data analysis, the data set is whatever one chooses to include. In this case, we could add new data by gathering con-

temporaneous surveys on the same subject, of similar people, and treat them as part of the pool of potential control units (see Dehejia and Wahba, 1999). In fact, adding external data has been tried in this application but turns out not to help because it greatly increases heterogeneity, does not markedly increase the information in the data as n increases about the ATT, and so does not satisfy the conditions for the theory of inference to apply (see King, Lucas, and Nielsen, 2017; Smith and Todd, 2005).

Applying Approximate Simple Random Sampling-Based Theory At this point, we can see that applying an existing theory of inference based on simple random sampling, to generate valid causal inferences in these data without approximations, is not possible. Researchers in this situation typically try to come up with an approximate matching solution, but this leads to two problems.

First, approximate matching is not justified by the simple random sampling-based theory of inference, as the formal properties of the resulting statistical estimators do not hold. Second, one might think that small deviations from the theoretical requirement would be approximately unbiased, but this is untrue. No known theorem supports this claim and, since even exactly matched propensity scores implies only approximate matching on X, which can greatly increase the variance across samples and drive any one sample farther from the truth.

By looking at how imbalanced a dataset is, we can get a feel for at least the potential bias due to failing to exactly matching on X or on the propensity score. To illustrate, we use one-to-one nearest neighbor propensity score matching (NN-PSM) with a caliper of 0.001. This results in 100 treated units and 100 control units. The closest (inexact) match allowed by this procedure has a difference in propensity scores of only 0.000003, but yet still has substantial imbalance:

treated	age	education	black	hispanic	married	nodegree	u74	u75	re74	re75
1	20	9	0	1	0	1	0	0	8740.939	8015.442
0	23	8	1	0	0	1	1	1	0.000	0.000

As can be seen, `education` (at 9 years for treated and 8 for control) is not far off, at least for a job training program. Also apparently close is `age` but, at 20 and 23, the impact could be determinative if the legal age of adulthood (21) impacts prospective employers

hiring decisions. More serious is that the treated person in this pair is `hispanic` and employed in both 1974 and 1975, whereas the control person is `black` and unemployed in the same years. In this typical example, a practitioner would have to implicitly admit to not controlling for several of the variables they designated as pre-treatment confounders, thus violating ignorability or to hope that the biases of one confounder miraculously cancels out the biases for another.

We also repeat here the same analysis for a larger caliper to further increase the number of matched units, and show its cost in terms of increasing imbalance further. Here we choose a larger caliper of 0.01, and find 219 treated units matched with 219 control units. The next best pair of matched units this brings in has a "small" propensity score difference of 0.000622, but with an obviously large imbalance on X:

treated	age	education	black	hispanic	married	nodegree	u74	u75	re74	re75
1	27	8	1	0	0	1	1	1	0.00	0.00
0	27	12	0	1	0	0	0	0	27913.66	24276.97

The differences between the treated and control groups on X here are even more substantial, even with an only slightly larger caliper. Here we match a treated African American who dropped out of junior high school with no income, to a control group Hispanic who graduated from high school with more than $27,000 of income.

The two matched pairs of units we describe here are each intuitive and the degree of approximation is easy to understand. However, to understand the full degree of approximation for the entire matching solution requires performing this identical comparison on *every* pair of matched observations (100, 219, or 274, depending on the choice of caliper).

Although we do not offer an example until later, we also note that running NN-PSM with a caliper of 0.1 matches 274 treated units to 274 control units (i.e. 548 units).

As is clear from this discussion, the size of the propensity score caliper alone provides little intuition about the quality of the match, the degree of approximation to the requirements of the theory of inference, the resulting level of imbalance on X, or the degree of statistical bias in the ATT. Since the required Axiom A0' is an axiom, being able to clearly convey what it means is the only real requirement in applications.

Applying Stratified Random Sampling-Based Theory We now illustrate three advantages of replacing Axiom A0' with Axiom A0, and thus thinking of the data in terms of stratification rather than simple random sampling.

The first advantage of stratification-based matching is ease of interpretation, which is essential in matching. Understanding assumptions — which by definition are unverifiable — is important in any empirical analysis. However, the sampling axiom in matching *defines* the inferences being made and thus also the meaning of the sampling distribution and standard errors. As such, without some clear understanding of the sampling process, making any inferences at all makes little sense.

To convey how one would interpret an application under this stratification-based theory of inference, and why matching is easier to understand than under simple random sampling-based approaches, we now give an example involving analysis choices in a real data set. For stratification-based inference, the key choice is the partition of X, which we have been referring to as A. In principle, this choice must be made prior to examining the data, or else the weights will not be fixed in repeated samples. We discuss different ways of interpreting this requirement so that it may be used in practice when not generating the data oneself.

A reasonable way to define A, before seeing the data, is to define it based on information in the data set's codebook. In the case of these data, a natural choice is to match all binary variables exactly, `age` according to the official U.S. Bureau of Labor Statistics stratification (i.e., 16–24, 25–54, 55 and over), and for the variable `education` to coarsen by formal degrees — *elementary* [0,6], *high school* [7–12], *undergraduate* [13–16], *graduate* [17,). The covariate `u74` (and `u75`) is an indicator variable which is nonzero when `re74` (and `re75`) is nonzero. As a result, this continuous counterpart of the unemployment status (`re74` and `re75`) can be in principle dropped from the matching stage and eventually included in the model specification for the ATT estimation.

If we use this definition of A, which we could plausibly have arrived at before examining the data, then we can think of the data generation process as stratified random sampling within this given partition. Then all hypothetical repeated samples, the resulting

sampling distribution, and associated standard errors, confidence intervals, and hypothesis tests, is defined as a consequence. As it happens, when we can try this stratification with our one observed data set, we find that we have 221 treated units matched with 313 control units.

Now suppose we examine the data and prefer to drop `re74` and `re75` from the partition in order to prune fewer observations. To make this decision statistically justifiable, two conditions must hold. The first condition is that we must ensure we do not violate set-wide ignorability (i.e., Assumptions A2 and A3), which will be satisfied in one or more of three situations: the two variables are unrelated to the treatment variable, unrelated to the outcome given the treatment, or included in an appropriate model during the estimation stage. The second condition involves conceptualizing the resulting strata A. If the choice of A is determined from the data (not merely the codebook), then the weights are random and, as a result, more complicated methods must be used for uncertainty estimates (point estimates remain unchanged). However, we may still be able to conceptualize A as fixed ex ante if we can argue that we would have interpreted the partitions the same way if we had thought of the same reasoning before seeing the data. That is, sometimes seeing the data causes one to surface ideas that could easily have been specified ex ante. Of course, we should try to avoid the lure of post hoc, just-so stories, but if we do, we would be justified in interpreting A as fixed, and then all the familiar methods are available for computing uncertainty estimates, such as standard errors and confidence intervals. Either way, the advantage of stratification is that understanding the sampling axiom, and how to think about the resulting data generation process, is straightforward. In this case, dropping `re74` and `re75` result in matching 278 treated units matched to 394 control units.

Now suppose we go another step and try to interpret our analyses without much prior knowledge of A. Here, we first generate a large set of matching solutions. This need not be done in practice, but we find it useful here for illustrative purposes. To do this, we create 500 random stratifications of the covariate space by dividing the support of each pre-treatment covariate into a random number of subintervals (chosen uniformly on the integers 1,. . . ,15). For comparison, we also generate 500 matching solutions from NN-PSM

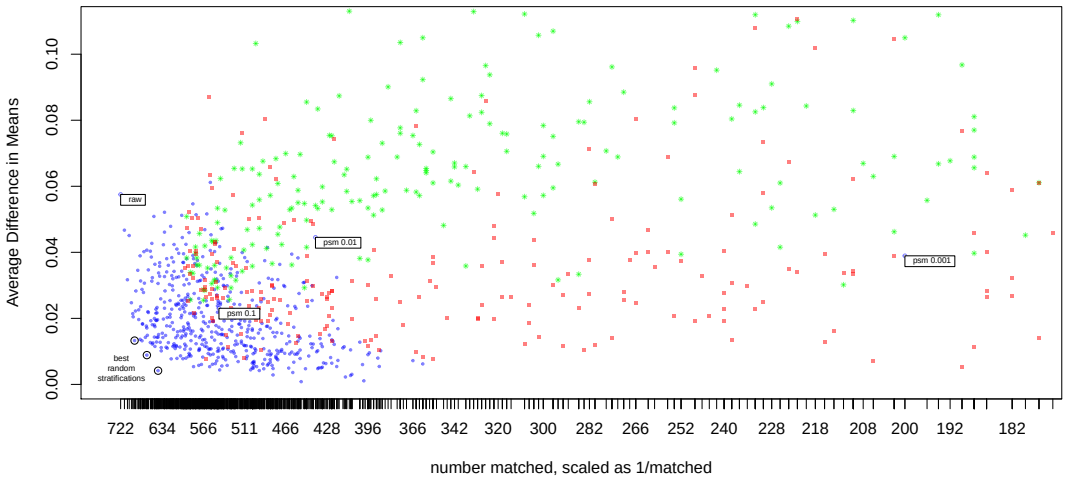

Figure 1: Randomly created matching solutions: imbalance (vertically) by matched sample size (horizontally). Each dot represents a matching solution, including the original unmatched data (Raw); three solutions with lower imbalance and more matched treated units than any other (best random stratifications, at the lower left); NN-PSM solutions with calipers of 0.1, 0.01, and 0.001; 500 solutions based on random stratifications of the covariate space (dots); 500 random NN-PSM solutions (stars), and 500 random NN-MDM solutions (squares). The plot represents solutions with at least 200 matched units.

(nearest neighbor, propensity score matching) models, by randomly selecting propensity score models and its caliper, and 500 NN-MDM (nearest neighbor, Mahalanobis distance matching) solutions, by randomly selecting input variables and its caliper (both selecting from the set of all main, polynomial, and interaction terms up to the second degree, with a logistic specification as usual for the propensity score model). In real applications, imbalance is best measured on the scale of the original variables but, to save space for our methodological purposes, we use the average of the standardized difference in means applied to each matching variable. (Other measures of imbalance do not materially change our conclusions.) Then, in Figure 1, the vertical axis is this measure of imbalance and the horizontal axis is the number of matched units (scaled according to $1/n$). Each point in the plot corresponds to one randomly selected matching solution (with stratification solutions in blue, NN-PSM in green, and NN-MDM in red). Stratification solutions here are all based on coarsened exact matching (CEM), but our stratified theory of inference applies to any member of the class of "monotonic imbalance bounding" matching methods (S. M. Iacus, King, and Porro, 2011).

The `raw` dot (at the middle left) corresponds to the original, unmatched data. In the same figure, we also include and label the three different NN-PSM matching solutions discussed above with calipers set to 0.1, 0.01 and 0.001 (across the middle of the graph from left to right). The dots marked with "best stratifications" (at the bottom left) represent matching solutions based on stratifications with the lowest imbalance for a given number of matched units or the largest number of matched units for any given imbalance. These solutions do not necessarily represent the theoretical frontier of imbalance and matched sample size, since they were generated randomly, but they are the best solutions among those in this graph (King, Lucas, and Nielsen, 2017), but are still the best among those randomly generated.

Then, to convey how easy it is to understand a stratified matching solution, consider only the central dot of this sequence of "best stratifications". This matching solution was constructed (by chance, i.e. randomly) using the cross products of following strata:

variable	class cut-points
`age`	(min=17, 29.67, 42.33, max=55), i.e. three classes
`education`	(min=3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, max=16)
`re74`	(min=0, 19785.34, max=39570.68), i.e. two classes
`re75`	(min=0, max=37431.66), i.e. one large class
`black`	all in the same class, i.e. no matching
`hispanic`	classes 0 and 1, i.e. exact matching
`married`	classes 0 and 1, i.e. exact matching
`nodegree`	classes 0 and 1, i.e. exact matching
`u74`	classes 0 and 1, i.e. exact matching
`u75`	classes 0 and 1, i.e. exact matching

The advantage of presenting these strata is that they convey *all* information necessary about the entire matching solution in an easy-to-understand and compact display. To use our stratified theory of inference, we need to imagine that our data, and all the repeated hypothetical samples, were generated by a stratified sampling design, based on these strata. In fact, the data are observational, and the hypothetical distributions do not and will not exist. However, we can still conceptualize what this distribution means as if these strata are fixed. The argument should be recognized as more of a stretch, since we did arrive at this stratification directly from the data, but stating this axiom about hypothetical (stratified) sampling replications cannot be proven wrong and so it is reasonable to use it as a way to interpret a matching-based estimator. Our main point here is that axiom itself is easy to understand: all we need to do is to understand the strata defined above.

In contrast, to convey all information in a NN-PSM matching solution, we would need

to understand every individual matched set in a matching solution, as we did above, but for 100, 219, or 274 individual matched sets (corresponding to calipers of 0.001, 0.01, and 0.1). This of course would be infeasible to comprehend all at once. With stratification, a researcher can more easily, quickly, and concisely understand the approximation and what assumptions are necessary to believe that bias is being constrained, without hundreds of separate evaluations. These stratifications might still be implausible as ex ante definitions for A, but researchers will be able to understand, if appropriate justify, the assumptions more easily. In this example, we can see that this particular matching solution does not control for `black` or `re75`, as was the case for particular pair of NN-PSM matched sets above, but this time we can see all the compromises from the entire data set at once, so that one can judge whether this approximation is justifiable. The problem of course is that even if one can understand a hundred or more stratifications, the axiom of simple random sampling requires exact matching on X or the propensity score, not a nearest neighbor solution, or one within some caliper.

A second advantage of stratification-based matching is that imbalance tends to be lower than under other matching methods for any given number of matched observations. This is not a general claim, but it is a typical pattern in many applications (King, Lucas, and Nielsen, 2017). This can be seen in Figure 1 by the blue stratification-based matching points appearing to the lower left — indicating lower imbalance given higher numbers of observations — whereas the green NN-PSM and red NN-MDM matching solutions, appearing above and to the right — indicating more imbalance or fewer matched observations.

A final advantage of stratification-based matching is that estimated treatment effects are often less variable, and thus somewhat more robust, than under other matching methods. To see this common, but also not universal, pattern we compute, for each of the matching solutions in Figure 1, an estimate of the ATT and standard error (by regressing the outcome variable on the treatment indicator and all pretreatment variables included in the matching solution). We then present, in Figure 2, all the ATT estimates (vertically) by the matched sample size (horizontally). Because of the enormous variability of NN-

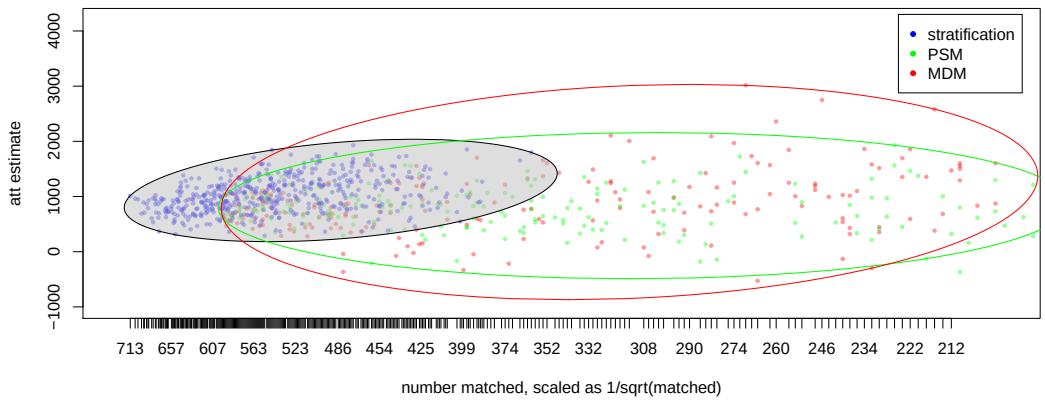

Figure 2: Estimated ATT (vertically) by number of matching units (horizontally), for matching solutions in Figure 1. The ellipses are convex hulls of the plotted points and represent roughly the dispersion of the ATT estimates for each matching method. The plot represents solutions with at least 200 matched units.

Mahalanobis, the plot is zoomed in, excluding points outside of the range of the visible axes.

This figure shows that for any given matched sample size (i.e., any point on the horizontal axis), the vertical variability of the ATT estimates are much larger for NN-PSM and NN-MDM than for stratification. Because Figure 2 does not reveal all the data, we present Table 1, which summarizes aspects of these same estimates. Table 1 demonstrates that stratification, in addition to having lower overall imbalance (i.e., finding better matched subgroups of treated and control units), is also the method that on average produces more matched units, a less variable and more robust ATT estimate, with a smaller standard error. In addition, the one-sigma Monte Carlo confidence interval for average treatment effect contains zero for both PSM and MDM, but not under stratification.

As we exemplify with this analysis, the choice of a theory of inference defines the nature of the hypothetical repeated samples used for statistical inference. Whether these samples are based on simple or stratified random sampling is not an assumption vulnerable to being proven wrong, but rather than axiom that defines how we interpret standard errors and confidence intervals. As such, the critical question is not which is more appropriate but whether we are able to clarify the meaning of one's uncertainty calculations. As

Method	min ATT	average ATT	median ATT	max ATT
MDM	-108673.5	819.1	886.0	31915.9
PSM	-1693.3	681.3	710.7	5586.3
Stratification	287.4	1031.1	999.7	1931.0

Method	min Std. Error	average Std.Error	median Std. Error	Max Std.error
MDM	510.2	2031.6	740.9	296298.1
PSM	194.3	870.2	696.0	4056.7
Stratification	461.6	529.6	524.2	681.3

Method	min n	average n	median n	max n
MDM	4	288	279	594
PSM	4	302	307	592
Stratification	360	550	553	713

Table 1: Distribution of estimated ATTs, their standard error and number of matched units for the data in Figure 2.

we show here, under stratified random sampling, the assumptions and inferences are considerably clearer and easier to understand, and do not require asymptotic results, which is quite unlike the situation with most methods for simple random sampling-based inference.

6 Concluding Remarks

In this paper, we highlight the assumptions and estimators necessary for identification and unbiased causal estimation when, as is usually the case in practice, matches are approximate rather than exact, and treatment variables are assumed known and applied without error. The theory of statistical inference we develop here justifies the common practice among applied researchers of using matching as preprocessing and then applying the same convenient and familiar methods of estimation and inference. Only with formally stated assumptions like those presented here can applied researchers begin to assess whether they are meeting the requirements necessary for valid causal inference in real applications. By moving the nearly universal stratification assumption made ex post into an explicit ex ante assumption, the assumptions that must be met are taken out of the shadows and made explicit. Researchers are still responsible for meeting these assumptions, and in observational data causal inference is always hazardous, but researchers should now be able to see more clearly the conditions necessary for generating valid inferences.

Appendix A Proofs

Proof of Theorem 1. This is true because, for each A, $\hat{\tau}^A$ is an unbiased estimator of τ^A. In fact,

$$E\{\hat{\tau}^A\} = \frac{1}{m_1^A}\sum_{i\in M_1^A} E(Y_i) - \frac{1}{m_2^A}\sum_{j\in M_2^A} E\{Y_j\} = \frac{1}{m_1^A}\sum_{i\in M_1^A}\mu_1^A - \frac{1}{m_2^A}\sum_{j\in M_2^A}\mu_2^A = \mu_1^A - \mu_2^A$$

now

$$E\{\hat{\tau}\} = \sum_{A\in\Pi(\mathcal{X})} E\{\hat{\tau}^A\}W_1^A = \sum_{A\in\Pi(\mathcal{X})}(\mu_1^A - \mu_2^A)W_1^A = \tau.$$

□

Proof of Theorem 2. Recall that $T^* = T - u$. If $Y(t)$ is a generalized additive function of T linearly and X, then it has a form like $a + bt + c\cdot h(X)$, for any deterministic function $h(\cdot)$ independent of t. Hence $E\{Y(T)\} - E\{Y(T^*)\} = a + bE\{T\} + c\cdot h(X) - a - bE\{T\} - c\cdot h(X) + bE(u) = bE(u) = 0$. □

Proof of Theorem 3. Recall that $Y(t) = a_0 + \sum_{k=1}^p a_k t^k$ with coefficients $a_0, a_1, \ldots, a_k$. Using independence of T and u and the fact that $T^* = T - u$, we write

$$\begin{aligned} E\{Y(T^*)\} &= a_0 + \sum_{k=1}^p a_k E\{(T-u)^k\} = a_0 + \sum_{k=1}^p a_k \sum_{i=0}^k \binom{k}{i} E\{T^i\}E\{(-u)^{k-i}\} \\ &= a_0 + \sum_{k=1}^p a_k \left(E\{T^k\} + \sum_{i=0}^{k-1}\binom{k}{i} E\{T^i\}E\{(-u)^{k-i}\}\right) \end{aligned}$$

and the result follows. □

Lemma 1. *[Mean Value Theorem (De Crescenzo, 1999)] Let X and Y be nonnegative random variables, with X stochastically smaller than Y. Let g be some measurable and differentiable function such that $E[g(X)]$ and $E[g(Y)]$ are finite; let g' be measurable and Riemann-integrable on $[x, y]$ for all $y \geq x \geq 0$. Then*

$$E\{g(Y)\} - E\{g(X)\} = E\{g'(Z)\}\,(E\{Y\} - E\{X\})$$

with Z a non-negative random variable with distribution function

$$F_Z(z) = \frac{F_X(z) - F_Y(z)}{E\{Y\} - E\{X\}}, \quad z \geq 0,$$

and F_X, F_Y and F_Z the distribution functions of X, Y and Z respectively.

Proof of Theorem 4. A direct application of Lemma 1, with $Y = T = T^* + u$, $X = T^*$ and $g = Y$. □

Appendix B Allowing True and Observed Treatment Status to Diverge

We show here how the stratified sampling-based theory of statistical inference is easy to extend in several ways. In particular, thus far, the observed treatment variable T has been assumed (here and the matching literature generally) to equal the true treatment actually applied, T^*, so that $T^* = T$. In most applications, this assumption is implausible and so we now let these two variables diverge. To do this, we offer definitions, assumptions for identification, and, when T is continuous, assumptions for estimation.

B.1 Definitions

Consider the following three cases:

i) *Versions of treatments*: Observing treatment variable $T = t_j$ implies that the unobserved true treatment $T^* = t^*$ belongs to a known set U_j. For example, if treatment group members are assigned to receive a medicine, say $T^* = t_1^*$, we know they take the medicine but, unbeknownst to the researcher, they take the medicine at different times of day, or with different foods, or in slightly different amounts, etc., within the constraints defined by set U_1. That is, we assume that all possible variations of the treatment belong to a set U_1. In this case, if the prescribed assignment to the treatment was $T^* = t_j^*$ but actually $t^* \in U_j$ was the true treatment received, then $T = t_j$ is observed, T^* and its realization t^* are unobserved, $Y(T)$ is a random variable (with variation depending on T^*), and its realization $Y(t^*)$ is observed.

ii) *Discretization*: In this situation, T^* is an observed (continuous or discrete) treatment, which the investigator chooses to discretize for matching as T. We set $T = t_j$ if $T^* \in U_j$, with U_j a prescribed (nonrandom) set. In this framework, $T = t_j$ and $T_i^* = t^* \in U_j$ are observed; $Y(T)$ is an observed random variable (with variation depending on the known T^*), and $Y(t^*)$ is an observed point.

iii) *Discretization with error*: Given the unobserved true treatment level T^*, we observe $\bar{T}^* = T^* + \epsilon$, where ϵ is unobserved error. Then, for the purpose of matching (again based on some substantive criteria so matches can be found), the observed value of $T = t_j$ corresponds to a discretized version of $\bar{T}^*$, i.e $T = t_j$ if $\bar{T}^*$ belongs to the interval U_j. As a result, $T = t_j$ is observed, T^* and ϵ are unobserved, $Y(T)$ is an observed random variable (with variation depending on the observed $\bar{T}^*$) and $Y(T^*)$ is an unobserved point.

The above cases correspond to an analysis based a discretized version of T^* which we denote by T. The distinguishing feature of these cases is that the discretization is controlled by unobserved features of the data generation process in case i), the investigator in case ii), and both in case iii). The discretization of T^* (in case ii) and $\bar{T}^*$ (in case iii) may be temporary for the purpose of matching and can be reversed when a modeling step follows matching.

When T and T^* diverge, we redefine the treatment effect as averaging over the variation (observed for ii and unobserved for i and iii) in $Y(T^*)$ for each observed treatment level so that analyzing a discretized version of the treatment variable rules out the problem of uncertainty about the true value of the treatment. That is, instead of comparing two treatment levels t_1 and t_2, we compare the average effect between two sets of unobserved true treatment sets U_1 and U_2. Thus, for two chosen observed levels, $T = t_1$ and $T = t_2$, the corresponding true treatment levels are $T^* = t^* \in U_1$ and $T^* = t^* \in U_2$, respectively. Then, the redefined treatment effect is

$$\text{TE}_i = E[Y_i(t^*) \mid t^* \in U_1] - E[Y_i(t^*) \mid t^* \in U_2] = E[Y_i(T_i = t_1)] - E[Y_i(T_i = t_2)]$$

with the averages SATT, FSATT, and others defined as in Section 2.3.

B.2 Assumptions

We keep the usual SUTVA assumption A1 but extend the framework of the previous sections to where the true treatment level T^* may diverge from the observed treatment level T. In what follows, we denote this mechanism as a map φ of the form $t = \varphi(t^*)$ which includes case i), ii) and iii) above.

We now introduce one additional assumption which ensures that different treatment levels remain distinct:

Assumption A4 [Distinct Treatments]**:** *Partition $\mathcal{T}$ into disjoints sets, U_j, $j = 1, \ldots$, and define φ as a map from T^* to T be such that $\varphi(t') \neq \varphi(t'')$ for $t' \in U_j$ and $t'' \in U_k$, $j \neq k$.*

Assumption A4 is enough to ensure the identifiability of the true treatment effect despite the divergence of T and T^*; it can usually be made more plausible in practice by choosing treatment levels that define the causal effect farther apart. A4 also says that discretizing the true treatment T^* into the observed value T does not affect the distribution of the potential outcomes; that is, if $T = 1 = \varphi(T^* = 2)$, the relevant potential outcome (which is observed if $T = 1$) is based on the (true) treatment actually applied, $Y(T^* = 2)$. Assumption A4 can also be replaced with instrumental variables and other assumptions where the divergence between observed and true treatment levels is conceptualized as noncompliance (e.g., Angrist, Imbens, and Rubin, 1996; Imai, King, and Nall, 2009), or different types of constancy assumptions (VanderWeele and Hernan, 2012).

To complete the setup, we make Assumption A2 compliant with Assumption A4. Let $D_U(z)$ be an indicator variable of the set U of $\mathcal{T}$ such that $D_U(z) = 1$ if $z \in U$ and $D_U(z) = 0$ otherwise. Then we replace Assumption A2 with A2', which we refer to as "double set-wide" because of the sets for the treatment and covariates:

Assumption A2' [Double Set-wide Weak Unconfoundedness]**:** *Assignment to the treatment T^* is weakly unconfounded, given pre-treatment covariates in set $A \in \Pi(\mathcal{X})$, if $D_U(t^*) \perp Y(t^*) | A$, for all $t^* \in U$ and each $U \subset \mathcal{T}$ and $A \in \Pi(\mathcal{X})$.*

A2' is again an extension of the notion of weak unconfoundedness suggested by Rosenbaum and Rubin, (1983).

B.3 Identification

Under coarsening of a continuous treatment, Assumptions A1, A2', A3 and A4 allow for identification and estimation of the treatment effect. For each $A \in \Pi(\mathcal{X})$ and $t^* \in U_i$, we

have

$$E\{Y(T^*)|A\} \overset{\mathbf{A2'}}{=} E\{Y(T^*)|D_{U_i}(T^*)=1,A\} = E\{Y|D_{U_i}(T^*)=1,A\}$$
$$= E\{Y|T^* \in U_i, A\} \overset{\mathbf{A4}}{=} E\{Y|T=t_i,A\} = E\{Y(t_i)|A\}$$

Hence, the average casual effect for $t^* \in U_1$ versus $t^* \in U_2$, within set A, is

$$E\{Y(t_1^*) - Y(t_2^*)|A\} = E\{Y(t_1)|A\} - E\{Y(t_2)|A\}$$

Then, under Assumption A3, we average over all strata as in (2), which enables us to compute the average treatment effect even when conditioning on an observed treatment assignment that differs from the true treatment.

B.4 Assumptions for Estimation when T is Continuous

In case iii) where the observation is continuous, a meaningful quantity of interest is $E\{Y(t_1^*) - Y(t_2^*)\}$, given the comparison of two chosen levels of the treatment t_1^* and t_2^*. After matching, $E\{Y(t)\}$ is modeled and used to estimate $E\{Y(T^*)\}$. Our goal here is to evaluate the discrepancy $E\{Y(t_1) - Y(t_2)\} - E\{Y(t_1^*) - Y(t_2^*)\}$, which of course we want to be zero. We begin with an assumption on the type of measurement error, u:

Assumption A5 [Berkson's type measurement error]**:** *Let $T = T^* + u$, with $E(u) = 0$ and u independent of the observed treatment status T and $\mathcal{X}$.*

(We name Assumption A5 in honor of Berkson, (1950), although we have added the condition, for our more general context, of independence with respect to $\mathcal{X}$; see also Hyslop and Imbens 2001.) We now offer three theorems that prove, under different conditions, the validity of using T for estimation in place of T^*. We begin with the simplest by assuming that $Y(t)$ is linear in t, although it may have any relationship with X.

Theorem 2. *Under Assumptions A1, A2', A3, A4, and A5, when $Y(t)$ is linear in t, and any function of X is independent of t, $E\{Y(T)\} = E\{Y(T^*)\}$.*

Theorem 2 enables us to work directly with the observed treatment T because $E\{Y(T)\} = E\{Y(T^*)\}$. With Assumption A5, we can write $E\{Y(T^*)|A\} = E\{Y(T)|A\}$ by a parallel argument. Therefore, Assumptions A1, A2', A3, A4, and A5 allow for valid causal

estimation even in the presence of approximate matching and a divergence between the observed and true treatment. The average causal effect for t_1^* versus t_2^* when $t_1 \in U_1$ and $t_2 \in U_2$ is then

$$E\{Y(t_1^*) - Y(t_2^*)|A\} = E\{Y(t_1) - Y(t_2)|A\}$$

Linearity in t, which is part of the basis of the assumption's reliance on the difference in means estimator, is not so restrictive because the Theorem 2 does not constrain the functional relationship with $\mathcal{X}$. Nevertheless, we can generalize this in two ways. First, consider a polynomial relationship:

Theorem 3. *Under Assumptions A1, A2', A3, A4 and A5, when $Y(t)$ is a polynomial function of t of order p, it follows that*

$$E\{Y(T)\} - E\{Y(T^*)\} = \sum_{k=1}^{p} a_k \sum_{i=0}^{k-1} \binom{k}{i} E\{T^i\} E\{(-u)^{k-i}\}.$$

If, in addition, we assume a structure for the error u such that the moments of u are known (e.g., $u \sim N(0,1)$ or the truncated Gaussian law to satisfy Assumption A4), then the moments of T can be estimated. With estimators of $a_0, a_1, \ldots, a_p$, we can estimate and correct for the bias term. For example, if $p = 2$ and $u \sim N(0,1)$ then the bias has the simple form $a_2(2E\{u^2\} + 2E\{T\}E\{u\}) = 2a_2$. So one estimates a generalized additive model for $E\{Y(T)\} = a_0 + a_1 T + a_2 T^2 + h(X)$ (with $h(X)$ any function of X) and adjust the result by $-2\hat{a}_2$. This makes valid estimation possible under this less restrictive polynomial process, once one assumes Assumptions A1, A2', A3, A4, and A5.

Our final generalization works under a special type of measurement error:

Assumption A6 [Stochastically ordered measurement error]**:** *Let $T = T^* + u$, with T^* a non-negative random variable and u a non-negative random variable independent of the observed treatment status T and $\mathcal{X}$.*

Then, we have our final theorem justifying how estimation can proceed:

Theorem 4. *Let Y be differentiable with respect to t. Then given Assumptions A1, A2', A3, A4 and A6,*

$$E\{Y(T)\} - E\{Y(T^*)\} = \int_0^\infty Y'(z)(F_{T^*}(z) - F_T(z))\mathrm{d}z$$

with and F_T and F_{T^*} the distribution functions of T and T^* respectively.

Theorem 4 allows one to estimates the bias due to the measurement error. If the distribution functions of u (or T) and T^* are known, this bias can be evaluated analytically or via Monte Carlo simulation. In Assumption A6, the measurement error cannot be zero mean and T^* is nonnegative. The measurement error u is still independent of T and, even though T is systematically larger than T^*, it is not deterministic. Note that if u is a negative random variable, a similar result apply with a change of sign in the above expression. Thus, Assumptions A1, A2', A3, A4, A5, and A6 allow for valid causal estimation if we can adjust for the bias, as in Theorem 3.

References

Abadie, Alberto and Guido W. Imbens (2002). "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects". In: *NBER Technical Working Paper* 283.

— (2006). "Large Sample Properties of Matching Estimators for Average Treatment Effects". In: *Econometrica* 74.1, pp. 235–267.

— (2011). "Bias-corrected matching estimators for average treatment effects". In: *Journal of Business & Economic Statistics* 29.1.

— (2012). "A Martingale Representation for Matching Estimators". In: *Journal of the American Statistical Association* 107.498, pp. 833–843.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables (with discussion)". In: *Journal of the American Statistical Association* 91, pp. 444–455.

Berkson, Joseph (1950). "Are there two regressions?" In: *Journal of the american statistical association*, pp. 164–180.

Cochran, William G. (1968). "The effectiveness of adjustment by subclassification in removing bias in observational studies". In: *Biometrics* 24, pp. 295–313.

Cochran, William G. and Donald B. Rubin (1973). "Controlling bias in observational studies: A review". In: *Sankhya: The Indian Journal of Statistics, Series A* 35, Part 4, pp. 417–466.

Cox, David R. (1958). *Planning of Experiments*. New York: John Wiley.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar Mitnik (2009). "Dealing with limited overlap in estimation of average treatment effects". In: *Biometrika* 96.1, p. 187.

De Crescenzo, Antonio (1999). "A Probabilistic analogue of the mean value theorem and its applications to reliability theory". In: *Journal of Applied Probability* 36, pp. 706–719.

Dehejia, Rajeev H. and Sadek Wahba (1999). "Causal Effects in Nonexperimental Studies: Re-Evaluating the Evaluation of Training Programs". In: *Journal of the American Statistical Association* 94.448, pp. 1053–62.

Ding, Peng (2016). "A paradox from randomization-based causal inference". In: *arXiv preprint arXiv:1402.0142*.

Heitjan, D.F. and Donald B. Rubin (1991). "Ignorability and Coarse Data". In: *The Annals of Statistics* 19.4, pp. 2244–2253.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart (2007). "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference". In: *Political Analysis* 15, pp. 199–236. URL: j.mp/matchP.

Holland, Paul W. (1986). "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81, pp. 945–960.

Hyslop, Dean R. and Guido W. Imbens (2001). "Bias from classical and other forms of measurement error". In: *Journal of Business and Economic Statistics* 19.4, pp. 475–481.

Iacus, Stefano M., Gary King, and Giuseppe Porro (2011). "Multivariate Matching Methods that are Monotonic Imbalance Bounding". In: *Journal of the American Statistical Association* 106, pp. 345–361. URL: j.mp/matchMIB.

Iacus, Stefano, Gary King, and Giuseppe Porro (2018). *Replication script for Iacus, King, Porro (2018), "A Theory of Statistical Inference for Matching Methods in Causal Research"*. DOI: 10.7910/DVN/AOY452. URL: https://doi.org/10.7910/DVN/AOY452.

Imai, Kosuke (2008). "Variance identification and efficiency analysis in randomized experiments under the matched-pair design". In: *Statistics in medicine* 27.24, p. 4857.

Imai, Kosuke and David A. van Dyk (2004). "Causal Inference with General Treatment Treatment Regimes: Generalizing the Propensity Score". In: *Journal of the American Statistical Association* 99.467, pp. 854–866.

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". In: *Statistical Science* 24.1, pp. 29–53. URL: j.mp/essrole.

Imbens, Guido W. (2000). "The role of the propensity score in estimating the dose-response functions". In: *Biometrika* 87 (3), pp. 706–710.

— (2004). "Nonparametric estimation of average treatment effects under exogeneity: a review". In: *Review of Economics and Statistics* 86.1, pp. 4–29.

Imbens, Guido W. and J.M. Wooldridge (2009). "Recent Developments in the Econometrics of Program Evaluation". In: *Journal of Economic Literature* 47 (1), pp. 5–86.

King, Gary, Christopher Lucas, and Richard A. Nielsen (2017). "The Balance-Sample Size Frontier in Matching Methods for Causal Inference". In: *American Journal of Political Science* 61.2, pp. 473–489.

King, Gary and Richard A. Nielsen (2017). "Why Propensity Scores Should Not Be Used for Matching". In: Working Paper. URL: http://j.mp/PSMnot.

King, Gary and Langche Zeng (2006). "The Dangers of Extreme Counterfactuals". In: *Political Analysis* 14.2, pp. 131–159. URL: j.mp/dangerEC.

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

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. by M. Lechner and F. Pfeiffer. Heidelberg: Physica, pp. 43–58.

Lin, Winston (2013). "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique". In: *The Annals of Applied Statistics* 7.1, pp. 295–318.

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

Miratrix, Luke W, Jasjeet S Sekhon, and Bin Yu (2013). "Adjusting treatment effect estimates by post-stratification in randomized experiments". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75.2, pp. 369–396.

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

Neyman, J. (1935). "Statistical problems in agricultural experimentation". In: *Journal of the Royal Statistical Society* II 2.107–154.

Robins, James M. (1986). "A new approach to causal inference in mortality studies with sustained exposure period - application to control of the healthy worker survivor effect". In: *Mathematical Modelling* 7, pp. 1393–1512.

Rosenbaum, Paul R. (1988). "Permutation tests for matched pairs with adjustments for covariates". In: *Applied Statistics*, pp. 401–411.

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

Rubin, Donald B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies". In: *Journal of Educational Psychology* 6, pp. 688–701.

— (1977). "Assignment to Treatment Group on the Basis of a Covariate". In: *Journal of Educational Statistics* 2.1-26, p. 1.

— (1980). "Comments on "Randomization Analysis of Experimental Data: The Fisher Randomization Test", by D. Basu". In: *Journal of the American Statistical Association* 75, pp. 591–593.

— (1990). "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies". In: *Statistical Science* 5.4, pp. 472–480.

— (1991). "Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism". In: *Biometrics* 47, pp. 1213–1234.

— (2010). "On the Limitations of Comparative Effectiveness Research". In: *Statistics in Medicine* 29.19, pp. 1991–1995.

Smith, Jeffrey A. and Petra E. Todd (2005). "Does matching overcome LaLonde's critique of nonexperimental estimators?" In: *Journal of Econometrics* 125.1-2, pp. 305–353.

Stuart, Elizabeth A. (2010). "Matching Methods for Causal Inference: A Review and a Look Forward". In: *Statistical Science* 25.1, pp. 1–21.

VanderWeele, Tyler J. and Miguel A Hernan (2012). "Causal Inference Under Multiple Versions of Treatment". In: *Journal of Causal Inference* 1, pp. 1–20.