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

Stefano M. Iacus† Gary King‡ Giuseppe Porro§

November 22, 2015

Abstract

To reduce model dependence and bias in causal inference, researchers usually use matching as a data preprocessing step, after which they apply whatever statistical model and uncertainty estimators they would have without matching. Unfortunately, this approach is appropriate in finite samples only under exact matching, which is usually infeasible, or approximate matching only under asymptotic theory if large enough sample sizes are available, but even then requires unfamiliar specialized point and variance estimators. Instead of attempting to change common practices, we show how those analyzing certain specific (but extremely common) types of data can instead appeal to a much easier version of existing theory. This alternative theory is substantively plausible, requires no asymptotic theory, and is simple to understand. Its core conceptualizes continuous variables as having natural breakpoints, which are common in applications (e.g., high school or college degrees in years of education, a governmental poverty level in income, or phase transitions in temperature). The theory 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.

†Department of Economics, Business and Statistics, 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 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 involves pruning observations to improve balance between the treated and control groups (solely as a function of measured pre-treatment covariates so as to avoid inducing selection bias) and then estimating the causal effect from the remaining sample. By eliminating or moderating the strength of the relationship between pre-treatment covariates and the treatment assignment variable, matching can reduce model dependence, estimation error, and bias (Cochran and Rubin, 1973; Rubin, 1974). By removing heterogeneous observations, matching can sometimes reduce variance but, when variance increases, the bias reduction offered usually more than compensates in typically large observational data sets. See Ho et al. (2007); Imbens (2004); Morgan and Winship (2014); Stuart (2010).

From the applied researcher's perspective, matching is convenient because it is treated an easy-to-use preprocessing step that does not disturb existing work flows and can be used even when, as usual, exact matching is infeasible. That is, after pruning observations that do not approximately match, researchers apply whatever statistical methods they would have without matching, such as a parametric regression modeling. In order to improve the performance of some nearest neighbor matching methods, such as based on propensity scores or Mahalanobis distance, applied researchers also often follow an ad hoc procedure of iterating between these formal methods and informal balance checking in the space of the covariates.

From the theoretical perspective, these practices are problematic because no finite sample theory of inference justifies them: Under existing theories of inference, researchers must either stick to exact matching, which will leave them with too few or no observations to infer anything, or, if enough observations are available, rely on asymptotic theory and switch from familiar analysis methods and variance estimators to specialized ones

(Abadie and Imbens, 2012).

The goal of this paper is to help applied researchers understand existing statistical theory so that know when they can use matching as they have, without new complicated approaches to learn; this enables them to use the techniques they know well, such as regression modeling and diagnostics. The theory of inference which enables these practices is simple to understand and apply, and does not require asymptotic making sampling assumptions or imposing distributional assumptions on the data researchers choose to analyze. Researchers are still responsible for meeting the assumptions of the theory in any application, but the simplicity of the theory may make this easier in the class of applications to which this theory applies.

Most of the change we propose is to merely use existing theory, but to recognize 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 some purposes. Variables with natural breakpoints are such an omnipresent feature of observational data that they are rarely even explicitly discussed. Most data analysts not only know this distinction well but use it routinely to collapse variables in their ordinary data analyses. For example, in analyses of sample surveys, which account for about half of all quantitative work in political science and a large fraction of work in rest of the social sciences (King et al., 2001, fn.1), examples of continuous variables with no natural breakpoints, and even without any examples where the authors used the breakpoints to collapse variables or categories, are uncommon.

Thus, much of our goal is to help researchers evaluate whether their work fits into a large class of applications we identify, for which existing theory works well, so commonly used matching practices need not be altered.

Section 2 presents our theory of statistical inference for matching, and Section 3 gives

the properties of estimators that satisfy it. We discuss what can go wrong and what to do about it in Section 4. Section 5 explains which matching methods and associated procedures are justified by the theory. Section 6 then extends the theory to situations where the true and observed treatment status diverge and where different versions of treatment are evident. Section 7 concludes.

2 A Theory of Causal Inference for Approximate Matching

Consider a sample of $n < \infty$ observations 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. (The potential outcomes could also be treated as random variables with a distribution induced by sampling from a given superpopulation or data generation process.) We also observe a $p \times 1$ vector of pre-treatment covariates X_i for subject i, and for some purposes consider this to be a random variable drawn from a superpopulation, where $X \in \mathcal{X}$.

In most applications, repeated sampling from a given (super)population is either a convenient fiction or real but unobserved. In either case, the data generation process is at least partly an axiom rather than a substantive assumption. In theoretical discussions, researchers have made progress by treating the data as having been generated via *simple random sampling* (i.e., "complete randomization"), see e.g. Abadie and Imbens (2006). An alternative approach we use in Section 2.2 is *stratified random sampling* (i.e., "block randomization"), which, when designing a data generation strategy, is preferred

on grounds of bias, model dependence, and variance; in many observational data sets of interest here, however, either option is plausible and so may be a reasonable choice.

2.1 Quantities of Interest

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 (until Section 6). 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: $\text{SATT} = \frac{1}{\#\{T_i=t_1\}} \sum_{i\in\{T_i=t_1\}} \text{TE}_i$ (Imbens, 2004, p.6). (The control units are used to help estimate this quantity.) 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 (Imbens and Wooldridge, 2009).

The second type of causal quantity is when some treated units have no acceptable 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 et al., 2009; 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.)

2.2 Assumptions

The existing finite sampling theory of causal inference in observational studies is based on the assumption that it is possible to match treated and control units *exactly* on all measured pre-treatment covariates (Lechner 2001, Imbens 2000, and Imai and van Dyk 2004.) Exact matching in relatively informative data sets normally yields no (or too few) observations and so empirical analysts routinely violate its basic principles and match only approximately. Approximate matching can be justified under asymptotic theory, if enough data are available, but then specialize point and variance estimators are required (Abadie and Imbens, 2012). We introduce here a theory of statistical inference that does not require resorting to asymptotic theory unless some of the assumptions below fail to hold for a given observational study.

We now describe Assumptions A1–A3, which establish the theoretical background needed to justify causal inference under the standard practice of approximate matching in finite samples; this theory can be seen as a natural extension of the pointwise theory by Rosenbaum and Rubin (1983). The first assumption (which we generalize further to more realistic situations in Section 6) 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 suggests three interpretations (see VanderWeele and Hernan, 2012). First is "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).

Prior to introducing the next two fundamental assumptions, we clarify the link between the stratified sampling assumption and this theoretical framework. To be specific, we take discrete variables as they are and coarsen continuous variables at their natural breakpoints, which we take as fixed. As discussed above, natural breakpoints exist for almost all apparently continuous variables in real applications. Then the product space of all the discrete and coarsened continuous variables form a set of strata, within which observations are drawn randomly and repeatedly.

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. As described in Section 5, the theory does not require that our specific formalization of these strata be used in a matching method, only that the information is accounted for. With this in mind, we offer one clear formalization of this idea, in terms of a partition of the product space of the covariates.

Matching by discretization of continuous variables dates at least to Cochran (1968) (see also Rubin, 1977). We formalize these notions here:

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, if $\mathcal{X}$ is the product space of variables $\texttt{age} \times \texttt{gender} \times \texttt{earnings} = \mathcal{X}$, then one of the sets, A_k, might be the subset of young adult males making greater than \$25,000: $\{\texttt{age} \in (18, 24]\} \times \{\texttt{gender} = M\} \times \{(\texttt{earnings} > \$25000)\}$. When not required for clarity, we drop the subscript k from A_k and write A.

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, given 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. Assumption A2 is designed to obtain unconfoundedness within each set A instead of for every single point $\{X = x\}$. 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 (Imbens, 2000; Imai and van Dyk, 2004; Abadie and Imbens, 2006; Lechner, 2001) and was firstly articulated in Rosenbaum and Rubin (1983).

As an example, consider estimating the causal effect of the treatment variable "taking one introductory statistics course" on the outcome variable "income at 22 years old", 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.

Finally, any matching theory requires the condition of *"common support"*, 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,

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

Assumption A3 makes the search for counterfactuals easier since those in the vicinity of (i.e., with the same strata as), rather than exactly equal to, a given covariate vector $X \in A$ are now acceptable.

This hypothesis was introduced by Rosenbaum and Rubin (1983); the combination of the pointwise versions of both A2 and A3 is often referred as "strong ignorability" (Rosenbaum and Rubin, 1983; Abadie and Imbens, 2002).

2.3 Identification

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 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 simple and commonly used estimation procedures are used as is, along with familiar diagnostic techniques. We now give some details of how our theory of inference justifies these simple procedures.

Let $M_j^A = \{i : T_i = t_j, X_i \in A\}$ be the set of indexes of all matched 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. We assume A and m_j^A, and thus m_j, remain fixed under repeated sampling. However, m_j^A could be estimated via the first (observed) random draw and then fixed for the remaining (hypothetical) samples, which is analogous to the common practice of conditioning on a pretest in sample survey design.

3.1 Difference in Means Estimator

To describe the property of the estimators we adapt the approach of Abadie and Imbens (2011) and rewrite the causal quantity of interest as the weighted sum computed within each stratum A from (1):

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

where $\mu_k^A = E\{Y(t_k)|X \in A\}$ $(k = 1, 2)$, with weights are $W^A = m_1^A/m_1$, 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:

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

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 $\sigma^2_{\hat{\tau}} = \text{Var}(\hat{\tau})$ 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 $\sigma^2_{\hat{\tau}} = \sum_{A\in\Pi(\mathcal{X})}\left(\sigma_{\hat{\tau}^A}W^A\right)^2$. The weights W^A in (3) are fixed given that, in our stratified random sampling data generation process, the number of treated units ($T_i = t_1$) per strata A (m_1^A), is fixed.

3.2 Estimators With Small Strata

If one or more strata contains only one treated unit and one control unit, one cannot directly estimate the variance within the strata, but we can still obtain an estimate of it

by applying whatever estimator one would have applied to the data set without matching. To show this, we introduce a new set of weights to simplify the estimator in (4) as the difference in weighted means. For all observations, we define the weights w_i as

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

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

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

and the variance of this estimator is just the sum of the variances of the two quantities.

3.3 Robust Estimation via Regression Adjustment

If Assumption A2 holds, then adjusting for covariates is unnecessary. If Assumption A2 holds but the analyst is unsure, and so adjusts for pre-treatment covariates (with interactions), then the downside is trivial (Lin et al., 2013; Miratrix, Sekhon and Yu, 2013). If A2 does not hold, then adjusting for covariates after preprocessing may still produce unbiased estimates. In this sense, current practice is doubly robust. In this latter case, if researchers follow the rule of including in X any covariate that affects either T or Y, then this set will satisfy A2 if any subset satisfies A2 (VanderWeele and Shpitser, 2011).

As an example of covariate adjustment, we 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 $\hat{\tau} \equiv \hat{\beta}$ if weights w_i are used in estimation. So the standard error of $\hat{\beta}_1$ obtained as output of this simple weighted least squares (WLS) model is the correct estimate of $\sigma^2_{\hat{\tau}}$.

To introduce covariates $(X_1, X_2, \ldots, X_d)$, let

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

and again, by WLS, the estimated coefficient $\hat{\beta}_1$ is the estimator of the treatment effect $\hat{\tau}$ and the estimated variance is the standard error of $\hat{\beta}_1$. Other models, such as GLM with weights, can be used as well in a similar fashion: The only change to the estimator that one would have used without matching is to include these weights, as with any weighted analysis.

3.4 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.

4 What Can Go Wrong and What to Do About It

When a data set has no controls sufficiently close to a treated unit, or in our framework a stratum A does not include a sufficient number of treated and control units, the now prevalent view in the literature is that changing the quantity of interest and switching from SATT to FSATT is often the best approach (Crump et al., 2009; Iacus, King and Porro, 2011; Rubin, 2010). This is the motivation for the common application of calipers applied to existing methods, the only qualification being that the new estimand should be fully characterized (Iacus, King and Porro, 2011, Section 6.3.3).

In the relatively unusual situation when switching to FSATT is not an option, because only an inference about the original quantity of interest will do, we have four options.

	Fixed, finite sample	Asymptotic sample
Assume A2–A3 (keep strata A)	Case 1	Case 4
Violate A2–A3 (enlargen strata A)	Case 2	Case 3

Table 1: Four Cases for Estimating SATT With Unfilled Strata

Case 1: Fixed Strata with Adjustment via Extrapolation. As Iacus, King and Porro (2011) detail, the ultimate quantity of interest can be computed as a weighted average

with two parts: strata that contain both treated and controls units, and strata that contain only treated units (which require modeling and thus risk model dependence). We then use a parametric model, such as in Section 3.3, to extrapolate the missing potential outcomes. This weighted average approach will normally be more robust than fitting a structural model to the entire data set since, as in approaches like multiple imputation, the model is used only where needed and is thereby restricted from causing any damage to portions of the data set without problems to begin with. In this case the sample size and strata remains fixed, and A2–A3 are assumed.

Case 2: Strata Enlargement with Regression Adjustment. For each stratum without at least one treated and one control unit, create a new stratum from the union of sufficient adjacent sets, which, when combined, contain sufficient treated and control units. This union implies a modification of the original Assumptions A2–A3, and should of course only be done if appropriate theoretically or if the likely bias induced would be less than the precision needed in the final estimate. If this relaxed version of the assumption is substantively plausible, we can proceed without changing the theory or estimation procedure. If the alternative assumption is not likely correct, the result will be additional model dependence, even after conditioning on A. In this situation, a regression adjustment within the strata or a global model with weights determined as in the previous section may be used to obtain an acceptable or possibly unbiased estimate of the quantity of interest. In this case, the sample size remains fixed, and alternative forms of Assumptions A2–A3 are used, and unbiasedness may require assuming the veracity of the model chosen after matching.

Case 3: Strata Enlargement with Asymptotic Adjustment. We now adapt the nonparametric adjustment approach justified by the innovative asymptotic theory of Abadie and Imbens (2006, 2011, 2012). This theory establishes the speed at which the space of observations must be filled in order to find control observations to serve as counterfactuals Abadie and Imbens (2012, Prop. 1), so that approximate matching estimators work asymptotically almost as well as exact matching after nonparametric adjustment. Un-

like other asymptotic approaches, which are specific to a given method of matching, this approach does not require any distributional assumptions on the data.

In this view, one can merge those sets A which contain only control units or only treated units into other sets of the partition, or create a new stratum which is the union of both the original ones. This increased coarsening may introduce bias. To understand where bias may arise when some strata A need to be enlarged or changed, we study the following bias decomposition. 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:

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

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 a mix or simply enlarged, then assumption A2 no longer applies and, in general, $\mu_{t_k}(X) \neq \mu_k^A$ for $X \in A' \neq A$. Thus we proceed 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

$$\hat{\tau}^A = \frac{1}{m_1^A} \sum_{i \in M_1^A} (Y_i - \hat{\mu}_{t_2|A}(X_i)) - \frac{1}{m_2^A} \sum_{j \in M_2^A} (Y_j - \hat{\mu}_{t_2|A}(X_j)). \tag{6}$$

This estimator is asymptotically unbiased if the number of control units in each strata 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 (6) 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).

Case 4: Fixed Strata with Asymptotic Sampling and No Adjustment. If relaxing Assumptions A2–A3 is not an option but there is the possibility of increasing the sample size of the control group using a sufficiently large reservoir of control units, we can still produce unbiased estimates of the causal effect τ, without any post hoc adjustment.

More specificaly, under the assumption that $m_1^r/m_2 \leq \kappa$, with $0 < \kappa < \infty$, $r > k$, and k the number of continuous covariates, then by Proposition 1 in Abadie and Imbens (2012), all the strata A will be filled with probability one. This result is enough to obtain unbiased estimates of the causal effect under the original assumptions A2–A3 and without changing the initial partition $\Pi(\mathcal{X})$ and without other technical smoothness assumptions on the functions $\mu_t(x)$ and $\hat{\mu}_{t|A}(x)$. Notice that this result requires assumptions A2–A3: That is, under other approximate matching theories, even if the strata will be filled at the given rate, the bias would not vanish asymptotically for $k > 2$ and further nonparametric regression adjustment is required as in point 3. Instead, all we need here is Assumptions A2–A3 and standard asymptotics, with no bias correction.

5 How Specific Matching Methods Fit the Theoretical Framework

We now show that most common methods of matching, as used in practice, are justified by the theory of inference proposed here. Of course, whether appeal to this theory justifies the

specific inferences a researcher draws from any one data set depends on their appropriate use of an appropriate matching method to meet Assumptions A1–A3.

Thus, in this section, we focus on the prior question of which matching methods, if used appropriately, can in principle be justified by our theory of inference. The answer to this question for any one method depends on whether it makes use of the knowledge conveyed in the strata we denote A (within which we imagine observations are drawn over repeated samples). The particular representation we choose for this information (i.e., the strata A) is less important than knowing that the method includes this information encoded in it in some way.

The existence of the information itself and the fact that most researchers have this knowledge is rarely at issue. Almost all applied researchers have a great deal of knowledge about their data. They usually understand which covariates are discrete, recognize the natural breakpoints in their continuous variables, and thus perceive intuitively the strata in their data, and when observations within these strata are essentially equivalent (i.e., up to Assumption A2). The issue, then, is less whether the researchers are aware of this information, and instead whether the matching methods they choose use this information.

In this light, the theory of inference we proposed justifies Coarsened Exact Matching (CEM) if the chosen coarsenings correspond with the strata A (Iacus, King and Porro, 2011). Moreover, in real data sets, even though the number of strata grow fast in the number of variables, no more than n of these are populated; and, in practice, observations within a data set tend to cluster much more tightly than any random calculation would indicate. CEM falls in the class of Monotonic Imbalance Bounding (MIB) methods, and some other methods within this class are also easily justified by this theory of inference (Iacus, King and Porro, 2011). These include when familiar matching methods — such as propensity score matching, Mahalanobis distance-based matching, or others — or parametric methods — such as linear, logistic, or other regression analysis — are applied within CEM strata.

An approach not fully justified by our theory is a one-shot application of nearest neigh-

bor methods, such as based on propensity score or Mahalanobis distances. These methods do define strata and, if the strata happen to respect the strata A, then we might think that the theory can be used. However, the strata are defined only as a function of the data, without any integral way to add prior information about natural breakpoints in variables or other features represented in fixed strata A. In this sense, like Bayesian modeling without the ability to include known prior information, the methods used in this one-shot way have an impoverished representation of our knowledge of the data and so cannot be justified by our theory.

In practice, applied researchers seem to understand intuitively that these existing one-shot applications of some matching methods exclude considerable information, and it turns out are able to avoid the problem. We can see their intuition in their efforts to compensate by the common practice of iteratively switching between one of these methods, most often propensity score matching, and a direct examination of the imbalance in X between the treated and control groups. The iterations ensure that the deep prior knowledge analysts have is used, for example, by verifying that we not match a college dropout with a first year graduate student even if they have been in school only slightly different amounts of time. This can happen with the iterative method somewhat more automatically as they stratify more and more finely directly or on the propensity score.

Statisticians also understand the problem and have made numerous suggestions for how to perform this iterative process manually in order to try to include this crucial prior information (which we choose to represent in A) in their analyses. See for example Austin (2008), Caliendo and Kopeinig (2008), Rosenbaum, Ross and Silber (2007), Stuart (2008), and Imai, King and Stuart (2008). Similarly, Ho et al. (2007, p.216) recommend searching across matching solutions and using the one with the best balance on X. Rosenbaum and Rubin (1984) try to compensate for the missing prior information in an application of propensity score matching (a technique they invented) by including and excluding covariates in their propensity score regression until sufficient balance on X information is included. Finally, Imbens and Rubin (2009) propose a mostly automated algorithm to iteratively adjust until convergence between propensity score matching and

specific types of balance checking, along with a warning to include manual checking.

Thus, although many one-shot applications of nearest neighbor matching methods are not justified by our theory of inference, the application of these methods when combined with the iterative procedure used in practice are much better justified. When an analyst is able to include all the relevant prior information (which we summarize as A) during the stage in which they check balance on X, then iterative application of matching methods are justified by the theory of matching described in this paper.

Finally, for clarity, we note that this iterative procedure when used with propensity score matching in particular requires an assumption (even before considering the theory we propose here) that has not been formally stated. That is, the validity of the iterative procedure depends on the assumption (in addition to A1–A3) that a subinterval $[p_1, p_2]$ of the (0,1) range of the propensity score scale corresponds to a unique set A. More precisely, for all i such that the propensity score $e(X_i) \in [p_1, p_2]$, there exists a set $A \in \Pi(\mathcal{X})$ such that $X_i \in A$; that is, $e^{-1}([p_1, p_2]) = A$, where $e^{-1}(\cdot)$ is the inverse image of the propensity score function. This additional assumption is required to ensure that essential information about closeness of units the k-dimensional space of $\mathcal{X}$ is not obliterated when transformed into the scalar propensity score before matching.

6 Allowing True and Observed Treatment Status to Diverge

Thus far, the observed treatment variable T has been assumed (by us 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.

6.1 Definitions

Consider the following three cases:

i) *Versions of treatments*: Observing treatment variable $T = t_j$ implies that the unob-

served 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 2.1.

6.2 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).

6.3 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

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

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.

6.4 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 $E\{Y(T)\} = a_0 + a_1T + a_2T^2 + h(X)$ (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.

7 Concluding Remarks

This paper highlights 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 not 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 forgetting it while applying other models and methods, as well as the common practice of iterating between formal matching methods and informal balance checks. 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. Adding this approach to the tools available may enable applied researchers in appropriate situations to harvest the power of matching without changing their well known data analytic procedures.

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

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

□

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 $g = Y$. □

References

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

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

Abadie, Alberto and Guido W Imbens. 2011. "Bias-corrected matching estimators for average treatment effects." *Journal of Business & Economic Statistics* 29(1).

Abadie, Alberto and Guido W Imbens. 2012. "A Martingale Representation for Matching Estimators." *Journal of the American Statistical Association* 107(498):833–843.

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

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.

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

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

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

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

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." *Biometrika* 96(1):187.

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

Heitjan, D.F. and D.B. Rubin. 1991. "Ignorability and Coarse Data." *The Annals of Statistics* 19(4):2244–2253.

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.

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

Hyslop, Dean R. and Guido W. Imbens. 2001. "Bias from classical and other forms of measurement error." *Journal of Business and Economic Statistics* 19(4):475–481.

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

Imai, Kosuke and David A. van Dyk. 2004. "Causal Inference with General Treatment Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99(467):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." *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.

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

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

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

Imbens, G.W. and J.M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47:5–86.

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1):49–69. http://gking.harvard.edu/files/abs/evil-abs.shtml.

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.

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.

Lin, Winston et al. 2013. "Agnostic notes on regression adjustments to experimental data:

Reexamining Freedmans critique." *The Annals of Applied Statistics* 7(1):295–318.

Miratrix, Luke W, Jasjeet S Sekhon and Bin Yu. 2013. "Adjusting treatment effect estimates by post-stratification in randomized experiments." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(2):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." *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." *Mathematical Modelling* 7:1393–1512.

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, 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, DB. 1990. "On the Applicaiton of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science* 5(4):472–480.

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

Rubin, Donald B. 1977. "Assignment to Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2(1-26):1.

Rubin, Donald B. 1980. "Comments on "Randomization Analysis of Experimental Data: The Fisher Randomization Test", by D. Basu." *Journal of the American Statistical Association* 75:591–593.

Rubin, Donald B. 1991. "Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism." *Biometrics* 47:1213–1234.

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

Stuart, Elizabeth A. 2008. "Developing practical recommendations for the use of propensity scores: Discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003'." *Statistics in Medicine* 27(2062–2065).

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

VanderWeele, Tyler J and Ilya Shpitser. 2011. "A new criterion for confounder selection." *Biometrics* 67(4):1406–1413.

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