J. R. Statist. Soc. A (2008)
171, *Part* 2, *pp.* 481–502

Misunderstandings between experimentalists and observationalists about causal inference

Kosuke Imai,

Princeton University, USA

Gary King

Harvard University, Cambridge, USA

and Elizabeth A. Stuart

Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

[Received January 2007. Final revision August 2007]

Summary. We attempt to clarify, and suggest how to avoid, several serious misunderstandings about and fallacies of causal inference. These issues concern some of the most fundamental advantages and disadvantages of each basic research design. Problems include improper use of hypothesis tests for covariate balance between the treated and control groups, and the consequences of using randomization, blocking before randomization and matching after assignment of treatment to achieve covariate balance. Applied researchers in a wide range of scientific disciplines seem to fall prey to one or more of these fallacies and as a result make suboptimal design or analysis choices. To clarify these points, we derive a new four-part decomposition of the key estimation errors in making causal inferences. We then show how this decomposition can help scholars from different experimental and observational research traditions to understand better each other's inferential problems and attempted solutions.

Keywords: Average treatment effects; Blocking; Covariate balance; Matching; Observational studies; Randomized experiments

1. Introduction

Random treatment assignment, blocking before assignment, matching after data collection and random selection of observations are among the most important components of research designs for estimating causal effects. Yet the benefits of these design features seem to be regularly misunderstood by those specializing in different inferential approaches. Observationalists often have inflated expectations of what experiments can accomplish; experimentalists ignore some of the tools that observationalists have made available; and both regularly make related mistakes in understanding and evaluating covariate balance in their data. We attempt to clarify some of these issues by introducing a general framework for understanding causal inference.

As an example of some of the confusion in the literature, in numerous references across a diverse variety of academic fields, researchers have evaluated the similarity of their treated and control groups that is achieved through blocking or matching by conducting hypothesis tests, most commonly the t-test for the mean difference of each of the covariates in the two

Address for correspondence: Kosuke Imai, Department of Politics, Princeton University, Princeton, NJ 08544, USA.
E-mail: KImai@Princeton.Edu

 0964–1998/08/171481

groups. We demonstrate that when these tests are used as stopping rules in evaluating matching adjustments, as frequently done in practice, they will often yield misleading inferences. Relatedly, in experiments, many researchers conduct such balance tests after randomization to see whether additional adjustments need to be made, perhaps via regression methods or other parametric techniques. We show that this procedure is also fallacious, although for different reasons.

These and other common fallacies appear to stem from a basic misunderstanding that some researchers have about the precise statistical advantages of their research designs, and other paradigmatic designs with which they compare their work. We attempt to ameliorate this situation here.

To illustrate our points, we use two studies comparing the 5-year survival of women with breast cancer who receive breast conservation (roughly, lumpectomy plus radiation) *versus* mastectomy. By the 1990s, multiple randomized studies indicated similar survival rates for the two treatments. One of these was Lichter *et al.* (1992), a study by the National Institutes of Health which randomized 237 women to mastectomy or breast conservation, within blocking strata defined by age, clinical node status and the presence or absence of cardiac disease. To study whether this result generalized to women more broadly, the US Government Accounting Office used observational data from women being treated in general medical practices across the USA (US General Accounting Office, 1994; Rubin, 1997). The data came from the National Cancer Institute's 'Surveillance, epidemiology, and end results' database, with information on 5000 cancer patients, which includes nearly all women who were diagnosed with breast cancer in five states and four metropolitan areas. We illustrate our results by examining the design of these studies, rather than their findings, but note that the General Accounting Office study did find that the results from the randomized trials also held in the broader population. However, our results apply to all key designs for making causal inferences and not only to these studies.

2. Quantities of interest

Consider an observed sample of n units taken from a finite population of N units, where typically $N \gg n$. Stochastic processes that may not be fully observed or known generate variables representing the sample selection I_i and treatment assignment T_i mechanisms. As a result of these mechanisms, unit i is in our sample if $I_i = 1$ and not if $I_i = 0$; unit i received the treatment if $T_i = 1$ and not if $T_i = 0$. Without loss of generality, assume that the treated and control groups in the sample are each of size $n/2$ so that n is an even number. For each unit, two potential outcome variables exist, $Y_i(1)$ and $Y_i(0)$, which represent the fixed values of the outcome variable when T_i is 1 or 0 respectively. In the sample, the potential outcome variable that corresponds to the actual value of the treatment variable is observed, whereas the other is not observed, and so we write the observed outcome as $Y_i \equiv T_i\, Y_i(1) + (1 - T_i)\, Y_i(0)$ for units with $I_i = 1$. In our framework, therefore, (I_i, T_i, Y_i) are random variables.

We define the (unobserved) *treatment effect* for unit i as

$$\mathrm{TE}_i \equiv Y_i(1) - Y_i(0). \tag{1}$$

The quantity TE_i may vary across units as a function of the observed X_i and unobserved U_i pretreatment characteristics of unit i. We observe the covariates X_i but not U_i in the sample, and possibly neither in the remainder of the population. In practice, researchers often do not attempt to estimate TE_i for each i, and instead they estimate only its average over either the

sample, producing the *sample average treatment effect* SATE,

$$\text{SATE} \equiv \frac{1}{n} \sum_{i \in \{I_i = 1\}} \text{TE}_i,$$

or over the population, producing the *population average treatment effect* PATE (Imbens, 2004),

$$\text{PATE} \equiv \frac{1}{N} \sum_{i=1}^{N} \text{TE}_i.$$

In the breast cancer studies, SATE is the effect of mastectomy *versus* breast cancer for the women in a particular study. PATE, which is the quantity of real interest for women who are subsequently diagnosed with breast cancer, is the effect of breast conservation *versus* mastectomy among a larger population, e.g. all women who are diagnosed with breast cancer for whom either treatment would be an appropriate therapy.

3. A decomposition of causal effect estimation error

A simple baseline estimator of either SATE or PATE is the difference in the sample means of the observed outcome variable between the treated and control groups:

$$D \equiv \frac{1}{n/2} \sum_{i \in \{I_i = 1, T_i = 1\}} Y_i - \frac{1}{n/2} \sum_{i \in \{I_i = 1, T_i = 0\}} Y_i.$$

Then, the difference between PATE and this estimator, which we call *estimation error*, is

$$\Delta \equiv \text{PATE} - D. \tag{2}$$

By studying estimation error, we focus on the most basic goal of statistical inference—the deviation of an estimate from the truth—rather than all of the various commonly used approximations to this goal, such as unbiasedness, consistency, efficiency, asymptotic distribution, admissibility and mean-square error. These statistical criteria can each be computed from our results (by taking expectations, limits, variances, etc.), but all are secondary to understanding and ultimately trying to reduce estimation error in a particular study.

We simplify the decomposition of estimation error by considering an additive model that rules out interactions between the observed X and unobserved U covariates:

$$Y_i(t) = g_t(X_i) + h_t(U_i), \tag{3}$$

with unknown functions g_t and h_t, for $t = 0, 1$. Then, the key result is that estimation error Δ can be decomposed into additive terms

$$\Delta = \Delta_{\text{S}} + \Delta_{\text{T}} = \Delta_{\text{S}_X} + \Delta_{\text{S}_U} + \Delta_{\text{T}_X} + \Delta_{\text{T}_U}, \tag{4}$$

where $\Delta_{\text{S}} = \text{PATE} - \text{SATE}$ and $\Delta_{\text{T}} = \text{SATE} - D$ represent *sample selection* and *treatment imbalance* respectively (see Heckman *et al.* (1998) and King and Zeng (2006)). In the second line of equation (4), we further decompose sample selection error Δ_{S} into two components, Δ_{S_X} and Δ_{S_U}, due to selection on observed (X) and unobserved (U) covariates respectively. Treatment imbalance Δ_{T} similarly decomposes into components Δ_{T_X} and Δ_{T_U} due to imbalance with respect to these observed and unobserved covariates.

We now derive and interpret each of the components under the additive model of equation (3). To focus on the key issues in this paper, our decomposition assumes away other forms of estima-

tion error that often need to be attended to in actual empirical analysis, such as post-treatment bias, measurement error, simultaneity, lack of compliance with the treatment assignment and missing data, among others.

3.1. Sample selection

The first component, sample selection error, is given by

$$\Delta_{\mathrm{S}} \equiv \mathrm{PATE} - \mathrm{SATE} = \frac{N-n}{N}(\mathrm{NATE} - \mathrm{SATE}),$$

where NATE is the *non-sample average treatment effect* and is defined by applying the SATE formula to the observations in the population but not in the sample, i.e.

$$\mathrm{NATE} \equiv \sum_{i \in \{I_i = 0\}} \frac{\mathrm{TE}_i}{N-n}.$$

Thus, the sample selection error component of causal estimation error vanishes if one of three conditions holds:

(a) the sample is a census of the entire population, so that $I_i = 1$ for all observations and thus $n = N$;
(b) the treatment effect in the sample is the same as in the rest of the population, SATE = NATE;
(c) we redefine the problem so that the population of interest is coincident with the sample, in which case SATE and PATE are equivalent.

A special case of no sample selection error occurs when TE_i is constant over i, in which case SATE = NATE. In the presence of heterogeneous treatment effects, random sampling guarantees no sample selection bias rather than no sample selection error, i.e. $E(\Delta_{\mathrm{S}}) = 0$.

From the definition of TE_i in equation (1), under the additive model of equation (3) and after a little algebra, the sample selection error Δ_{S} as defined above can be decomposed into the two additive components relating to observed and unobserved covariates:

$$\Delta_{\mathrm{S}_X} = \frac{N-n}{N}\left[\frac{1}{N-n}\sum_{i \in \{I_i=0\}}\{g_1(X_i) - g_0(X_i)\} - \frac{1}{n}\sum_{i \in \{I_i=1\}}\{g_1(X_i) - g_0(X_i)\}\right],$$

$$\Delta_{\mathrm{S}_U} = \frac{N-n}{N}\left[\frac{1}{N-n}\sum_{i \in \{I_i=0\}}\{h_1(X_i) - h_0(X_i)\} - \frac{1}{n}\sum_{i \in \{I_i=1\}}\{h_1(X_i) - h_0(X_i)\}\right].$$

Alternatively, these components can be expressed in the form

$$\Delta_{\mathrm{S}_X} = \frac{N-n}{N}\int \{g_1(X) - g_0(X)\}\, \mathrm{d}\{\tilde{F}(X|I=0) - \tilde{F}(X|I=1)\}, \tag{5}$$

$$\Delta_{\mathrm{S}_U} = \frac{N-n}{N}\int \{h_1(U) - h_0(U)\}\, \mathrm{d}\{\tilde{F}(U|I=0) - \tilde{F}(U|I=1)\}, \tag{6}$$

where $\tilde{F}$ represents the empirical (possibly multivariate) cumulative distribution function. Since, by equation (3), the potential outcomes are deterministic functions of X and U, the treatment effect in the sample is the same as in the population when the distributions of X and U are identical in each. Specifically, when the empirical distribution of the observed pretreatment covariates X is identical in the population and sample—$\tilde{F}(X|I=0) = \tilde{F}(X|I=1)$—then Δ_{S_X}

vanishes. Similarly, when the empirical distribution of all unobserved pretreatment covariates is identical in the population and sample—$\tilde{F}(U|I=0)=\tilde{F}(U|I=1)$—then Δ_{S_U} vanishes. Since X is observed only in sample (and U is not observed at all) these conditions cannot be verified from the observed sample alone. However, if the population distribution of X is known, weighting or imputing can be used to adjust for the bias due to Δ_{S_X}. Alternatively, if we assume that the treatment effect is constant over X_i, then $g_1(X_i)-g_0(X_i)$ is constant, implying that $\Delta_{S_X}=0$. Similarly, if the treatment effect is assumed to be constant over U_i, then $\Delta_{S_U}=0$.

In the breast cancer studies, sample selection error refers to differences between the women in each study and those in the general population who are candidates for either treatment. We might expect sample selection error to be smaller in the observational study with 5000 patients who are broadly representative of at least five states and four metropolitan areas than in the small random assignment study with just 237 women, all of whom agreed to participate and were willing and able to travel to the National Institutes of Health for follow-up visits. In fact, the published studies on this experiment do not even contain information on exactly how the patients were selected. For the randomized study, observable sample selection error might include differences in income, information, education and severity of disease, whereas selection error that would be difficult for us to observe and adjust for might include psychological conditions that are related to a woman's decision to participate in a randomized trial.

3.2. *Treatment imbalance*

From previous definitions under the additive model of equation (3) and after a little algebra, the treatment imbalance error term $\Delta_T=\mathrm{SATE}-D$ as defined above can be decomposed into the two additive components,

$$\Delta_{T_X}=\frac{1}{n/2}\left\{\sum_{i\in\{I_i=1,T_i=0\}}\frac{g_1(X_i)+g_0(X_i)}{2}-\sum_{i\in\{I_i=1,T_i=1\}}\frac{g_1(X_i)+g_0(X_i)}{2}\right\}$$

for observed covariates and a corresponding expression for Δ_{T_U} for unobserved covariates, with $h_j(\cdot)$ and U_i replacing $g_j(\cdot)$ and X_i respectively. These terms can also be expressed as

$$\Delta_{T_X}=\int\frac{g_1(X)+g_0(X)}{2}\,\mathrm{d}\{\tilde{F}(X|T=0,I=1)-\tilde{F}(X|T=1,I=1)\},\tag{7}$$

$$\Delta_{T_U}=\int\frac{h_1(U)+h_0(U)}{2}\,\mathrm{d}\{\tilde{F}(U|T=0,I=1)-\tilde{F}(U|T=1,I=1)\}.\tag{8}$$

These components vanish if the treatment and control groups are balanced (i.e. have identical empirical distributions) for the observed X_i and unobserved U_i covariates. For example, $\Delta_{T_X}=0$ if the following equality holds:

$$\tilde{F}(X|T=1,I=1)=\tilde{F}(X|T=0,I=1),\tag{9}$$

which is entirely in sample and observable. If this condition is not met, we need to adjust the data to meet this condition so that valid inference can be made. In contrast, verifying the exact value of Δ_{T_U} is impossible since U is by definition unobserved.

In the breast cancer example, treatment imbalance error arises from observable and unobservable differences between women who receive breast conservation *versus* mastectomy. The randomization in Lichter *et al.* (1992) ensures that, if the study is sufficiently large, no systematic differences exist between the women who receive the two therapies. Their Table 1 compares the characteristics of women in the two treatment groups and shows few differences. In contrast,

because it was not randomized, the General Accounting Office breast cancer study is likely to suffer from some treatment imbalance since doctors do not base treatment decisions on random-number generators. Matching methods that were described in US General Accounting Office (1994) and Rubin (1997) attempt to deal as well as possible with observed differences but, without randomization, the samples may of course still differ in unobserved (and thus unadjusted) ways.

4. Generalizations

Blocking in experimental research involves the random assignment of units to treatment and control groups within strata (blocks) that are defined by a set of observed pretreatment covariates (Fisher, 1935). Blocking guarantees that the treated and control groups are identical with respect to these covariates so that they cannot affect our inferences. In contrast, *matching* is a procedure that involves dropping, repeating or grouping observations from an observed data set to reduce covariate imbalances between the treated and control groups that were not avoided during data collection (Rubin, 1973). Blocking takes place before randomization of treatments, whereas matching is implemented only after treatments have been assigned. Although their goals are so close that the terms are often used interchangeably, we keep the distinction here.

In this section, we show how the essential logic of our decomposition remains unchanged when blocking on all observed covariates, and when the quantity of interest is the average causal effect for the treated units rather than all units. (Changing to an infinite population perspective requires imagining a superpopulation from which the N population units are randomly drawn, and then averaging over this extra variation. Our resulting estimand changes from PATE to the *superpopulation average treatment effect* SPATE, i.e. $\text{SPATE} \equiv E\{Y(1) - Y(0)\} = E(\text{PATE})$. We denote the estimation error for SPATE as Δ^* and define it as $\Delta^* \equiv \text{SPATE} - D = \Delta_{\text{S}} + \Delta_{\text{T}} + \text{SPATE} - \text{PATE}$, which directly extends our decomposition in Section 3. No other results or analyses need change.)

4.1. Decomposition with complete blocking

Suppose that we select our n observations, completely block on X, and then randomly assign T to half of the units within each block. Letting $\mathcal{X}$ denote the set of unique observed values of the rows of X, our decomposition in equation (4) then becomes

$$\Delta = \Delta_{\text{S}} + \text{SATE} - D = \Delta_{\text{S}_X} + \Delta_{\text{S}_U} + \sum_{x \in \mathcal{X}} w_x \Delta_{\text{T}_{U|x}},$$

where w_x is the proportion of units in each stratum x of $\mathcal{X}$, and

$$\Delta_{\text{T}_{U|x}} = \int \frac{h_1(U) + h_0(U)}{2} \, \text{d}\{\tilde{F}(U|T=0, X=x, I=1) - \tilde{F}(U|T=1, X=x, I=1)\}.$$

This result demonstrates that some basic intuition of our decomposition in equation (4) remains the same, where blocking eliminates Δ_{T_X} and does not affect Δ_{S}. It also shows that Δ_{T_U} changes to the weighted average of $\Delta_{\text{T}_{U|x}}$, which is defined within strata of unique values of X. Since U and X are not necessarily independent, Δ_{T_X} and Δ_{T_U} may be related. Thus, blocking on the observed confounders may have an effect on the unobserved confounders.

4.2. Average treatment effect on the treated

For some purposes, we might consider the quantity of interest to be the treatment effect averaged over only the treated units. For example, a medical researcher may wish to learn the effect

of a drug on those who receive or would receive the treatment and no others. In our motivating example, we may be interested in the effect of receiving breast conservation, for the women who choose that therapy. For this, common practice is to define the sample or population *average treatment effect on the treated*, which are respectively

$$\text{SATT} \equiv \frac{1}{n/2} \sum_{i \in \{I_i=1, T_i=1\}} \text{TE}_i$$

and

$$\text{PATT} \equiv \frac{1}{N^*} \sum_{i \in \{T_i=1\}} \text{TE}_i,$$

where $N^* = \Sigma_{i=1}^{N} T_i$ is the number of treated units in the population. (The definition of SATT assumes, as we do throughout, that half the units receive the treatment and half do not.)

An analogous version of our PATE estimation error decomposition in equation (4) also holds for the estimation error for PATT, $\Delta' \equiv \text{PATT} - D$, which is equal to

$$\Delta' = \Delta'_{\text{S}_X} + \Delta'_{\text{S}_U} + \Delta'_{\text{T}_X} + \Delta'_{\text{T}_U}, \tag{10}$$

where

$$\begin{aligned}
\Delta'_{\text{S}_X} &= \frac{N^* - n/2}{N^*} \int \{g_1(X) - g_0(X)\} \, \text{d}\{\tilde{F}(X|T=1, I=0) - \tilde{F}(X|T=1, I=1)\}, \\
\Delta'_{\text{S}_U} &= \frac{N^* - n/2}{N^*} \int \{h_1(U) - h_0(U)\} \, \text{d}\{\tilde{F}(U|T=1, I=0) - \tilde{F}(U|T=1, I=1)\}, \\
\Delta'_{\text{T}_X} &= \int g_0(X) \, \text{d}\{\tilde{F}(X|T=0, I=1) - \tilde{F}(X|T=1, I=1)\}, \\
\Delta'_{\text{T}_U} &= \int h_0(U) \, \text{d}\{\tilde{F}(U|T=0, I=1) - \tilde{F}(U|T=1, I=1)\}.
\end{aligned}$$

Only the terms that are involved in $Y_i(0)$ enter treatment imbalance Δ'_{T_X} and Δ'_{T_U}. This is because SATT restricts itself to the treated units in sample for which $T_i = 1$ and thus the terms involving $Y_i(1)$ in SATT and D are identical. This means that terms involving $g_1(X_i)$ and $h_1(U_i)$ cancel on taking $\Delta_{\text{T}} = \text{SATT} - D$ and decomposing into Δ'_{T_X} and Δ'_{T_U}.

The sample selection error is given by

$$\Delta'_{\text{S}} \equiv \text{PATT} - \text{SATT} = \frac{N^* - n/2}{N^*} (\text{NATT} - \text{SATT}) = \Delta'_{\text{S}_X} + \Delta'_{\text{S}_U},$$

where

$$\text{NATT} \equiv \sum_{i \in \{I_i=0, T_i=1\}} \frac{\text{TE}_i}{N^* - n/2}$$

is the non-sample average treatment effect. As a result, all the intuition that we develop for PATE and SATE applies also to PATT and SATT, and this decomposition, as well.

Since almost any implementation of matching would affect Δ_{S} in estimating PATE and SATE, applications of matching typically change the goal to PATT or SATT. For example, if matching is implemented by selectively dropping only control units and the quantity of interest is changed to SATT, then researchers avoid the sample selection error completely, i.e. $\Delta'_{\text{S}} = 0$. PATT or SATT could be used for randomized experiments, but if the treated group is randomly selected these quantities will not differ systematically from PATE and SATE respectively.

5. Reducing estimation error

We now attempt to clarify how the specific features of common research designs that are used in a variety of disciplines can help to reduce estimation error. The decomposition that we offer in Section 3 provides a guide for demonstrating how each contributes to reducing different components of the error. We begin with the specific features of these designs and common statistical assumptions and then discuss the research designs themselves.

5.1. Design features

We summarize the effects of different design features in Table 1, which shows the effect of each design in reducing specific components of estimation error. For example, randomly sampling units from a population, which is normally considered the *sine qua non* of survey research, works to reduce sample selection error on average across experiments, i.e. $E(\Delta_{S_X}) = E(\Delta_{S_U}) = 0$, but not necessarily in any one sample. Only by changing the quantity of interest from PATE to SATE, or equivalently by taking a census of the population, is the sample selection component exactly eliminated in sample ($\Delta_{S_X} = \Delta_{S_U} = 0$). Weighting can eliminate the observed component of estimation error but cannot affect the unobserved component except in as much as it is related to the (observed) variables from which the weights are built.

Randomly assigning the values of the treatment variable (as in the randomized breast cancer study), which is normally considered the *sine qua non* of experimental research, reduces the components of estimation error arising from observed and unobserved variables on average, but not exactly in sample, i.e. $E(\Delta_{T_X}) = E(\Delta_{T_U}) = 0$. For example, if the randomized breast cancer experiment could have been conducted many times we would expect no differences between the women in the two treatment groups on average. However, in any one study, including the one

Table 1. Effects on the components of estimation error of various choices of design and statistical assumptions†

	Sample selection estimation error		*Treatment imbalance estimation error*	
	Observed Δ_{S_X}	*Unobserved* Δ_{S_U}	*Observed* Δ_{T_X}	*Unobserved* Δ_{T_U}
Design choice				
Random sampling	$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$		
Focus on SATE rather than PATE	$= 0$	$= 0$		
Weighting for non-random sampling	$= 0$	$=?$		
Large sample size	$\to ?$	$\to ?$	$\to ?$	$\to ?$
Random treatment assignment			$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$
Complete blocking			$= 0$	$=?$
Exact matching			$= 0$	$=?$
Assumption				
No selection bias	$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$		
Ignorability				$\overset{\text{avg}}{=} 0$
No omitted variables				$= 0$

†For column Q, '$\to A$' (where A in the table is either a fixed but unknown point, denoted '?', or 0) denotes $E(Q) = A$ and $\lim_{n\to\infty}\{\text{var}(Q)\} = 0$, whereas '$=^{\text{avg}} A$' indicates only $E(Q) = A$. No entry means no systematic effect, and '=?' indicates an effect of indeterminate size and direction. Matching is normally designed to estimate PATT or SATT and so this row in the table should be read as affecting components in equation (10) rather than equation (4).

that is actually being conducted, differences may remain between the women who receive breast conservation and mastectomy that form a component of estimation error.

Complete blocking (i.e. before randomization) eliminates imbalance on observed variables in sample, i.e. $\Delta_{T_X} = 0$, but its only effect on unobserved confounders is on the portion that is correlated with X, which is eliminated or reduced as well. When estimating the superpopulation average treatment effect, adding blocking to random treatment will always reduce estimation variance of the causal effect compared with randomization alone. If PATE is the estimand, then the same relationship also holds unless n is small. This is true no matter how badly the blocking variables are chosen. (This result, which to our knowledge has not appeared before in the literature, is given in Appendix A; related results are discussed in Cochran and Cox (1957), Greevy *et al.* (2004) and Imai (2007).) However, despite this gain in efficiency, a blocked experiment has fewer degrees of freedom and so can have lower power in small samples; simulations indicate that this is not an issue except in very small data sets (Imai *et al.*, 2007), and so blocking is almost always preferable when feasible. Appendix A also formalizes the common recommendation in the experimental design literature that researchers increase the variation of the outcome variables across blocks relative to that within blocks.

Lichter *et al.* (1992) blocked on three variables. If it had been feasible to block on other relevant variables, such as psychological status or other clinical indicators of disease, efficiency could have been improved. Of course, because patients cannot wait for another patient who matches them on background characteristics to arrive at the hospital before they are randomized to treatment, additional blocking may not have been feasible.

Exact matching in observational research has the same logical effect as blocking in experimental research, but it also comes with four weaknesses that blocking does not have. First, to avoid selection bias even with a random sample from the known population, the quantity of interest must typically be changed from PATE to PATT or SATT. With PATE, we would probably make $\Delta_{S_X} \neq 0$ while trying to make $\Delta_{T_X} = 0$; in contrast, by switching to PATT or SATT, matching researchers can make $\Delta'_{T_X} = 0$ while not affecting Δ'_S. Second, by definition, random treatment assignment following matching is impossible. Third, exact matching is dependent on the already collected data happening to contain sufficiently good matches. With blocking, we are not dependent on any existing set of treatment assignments because the blocks are established before randomization.

Finally, matching (or parametric adjustment) in the worst case scenario, such as on only a subset of highly correlated covariates that are uncorrelated with T but related to post-treatment variables, can increase bias compared with an unadjusted difference in means (Pearl, 2000). Although observationalists typically argue that this exception for matching does not affect their own research because they have sufficient prior theoretical knowledge to choose covariates appropriately, the possibility always exists. Adding matching to an existing parametric adjustment procedure almost always reduces model dependence, bias, variance and mean-square error (Ho *et al.*, 2007), but a parametric adjustment and matching taken together (like parametric analysis on its own) can in this worst case scenario increase bias and variance compared with an unadjusted difference in means.

This worst case scenario with matching and parametric analysis cannot occur with blocking followed by random treatment assignment, even when blocking on irrelevant covariates or on only a subset of relevant covariates. This benefit of blocking may seem especially surprising to observationalists. However, the inefficiency and bias in procedures for observational data can be seen, by analogy, as a result of needing to estimate the coefficients from an incorrectly specified parametric model. In contrast, blocking is equivalent to parametric adjustment where the model specification *and* the exact numerical values of the coefficients on the potential con-

founders are known and so can be adjusted for exactly, even if all covariates are not available. Thus, except in very small samples, blocking on pretreatment variables followed by random treatment assignment cannot be worse than randomization alone. Blocking on variables related to the outcome is of course more effective in increasing statistical efficiency than blocking on irrelevant variables, and so it pays to choose the variables to block carefully. But choosing not to block on a relevant pretreatment variable before randomization, that is feasible to use, is not justified.

When the sample size is large, the variance of each of the four components of estimation error becomes small. If n becomes large when the expected value of one of these components is 0, then the value of that component will become smaller and at the limit will approach 0 even in sample.

5.2. Assumptions

Experimentalists and observationalists often make assumptions about unobserved processes on the basis of prior evidence or theory. At worst, when the question is judged to be sufficiently important but no better evidence exists, these assumptions are sometimes based on no more than wishful thinking for lack of anything better to do. Either way, we need to understand these assumptions precisely, and what their consequences are for the components of estimation error.

The second portion of Table 1 lists three assumptions that are commonly used in the same way and for some of the same purposes as design features in the rest of Table 1. For example, the assumption of no selection bias that is made in numerous studies is that $E(\Delta_S) = 0$, not necessarily that $\Delta_S = 0$ in the observed sample. We could of course strengthen this assumption (to $\Delta_S = 0$) but this level of optimism is rarely justified or made in the literature.

The assumption of ignorability, which is most often made in statistics, implies that the component of estimation error due to unobserved variables is 0 in expectation ($E(\Delta_{T_U}) = 0$). In contrast, the assumption of no omitted variables (or no omitted variable bias), which is typically made in classical econometrics and many of the social sciences, is that U is either uncorrelated with X or has no causal effect on Y, conditional on X; the result is that $\Delta_{T_U} = 0$ exactly in sample. Assumptions need not be made about imbalance in observables since they can be checked directly, but the various types of parametric models and non-parametric adjustment procedures are routinely used to try to reduce Δ_{T_X} (or Δ'_{T_X}) further.

5.3. Major research designs

The major research designs are each combinations of the features and assumptions that were described above. Table 2 summarizes how a particular design affects each of the four components of the estimation error.

We begin with what we call the *ideal experiment*, which involves selecting a large number of units randomly from a well-defined population of interest, measuring and blocking on all known confounders X and then randomly assigning values of the treatment variable T. In this situation, researchers can claim that

(a) $\Delta_{S_X} \approx 0$ and $\Delta_{S_U} \approx 0$ because random sampling ensures that $E(\Delta_{S_X}) = E(\Delta_{S_U}) = 0$ and a large n makes the variances of Δ_{S_X} and Δ_{S_U} small while yielding $(N-n)/N \approx 0$ and SATE $\approx$ NATE,
(b) $\Delta_{T_X} = 0$ because of blocking and
(c) $\Delta_{T_U} \approx 0$ because random assignment implies that $E(\Delta_{T_U}) = 0$ and the large n makes the variance of Δ_{T_U} across repeated treatment assignments small also.

If the confounders in X include all confounders rather merely than all confounders that we happen to know, then $\Delta_{T_U} = 0$.

Table 2. Components of bias when estimating PATE†

Design choice	*Sample selection estimation error*		*Treatment imbalance estimation error*	
	Observed Δ_{S_X}	*Unobserved* Δ_{S_U}	*Observed* Δ_{T_X}	*Unobserved* Δ_{T_U}
Ideal experiment	$\to 0$	$\to 0$	$=0$	$\to 0$
Randomized clinical trials (limited or no blocking)	$\neq 0$	$\neq 0$	$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$
Randomized clinical trials (complete blocking)	$\neq 0$	$\neq 0$	$=0$	$\overset{\text{avg}}{=} 0$
Social science field experiment (limited or no blocking)	$\neq 0$	$\neq 0$	$\to 0$	$\to 0$
Survey experiment (limited or no blocking)	$\to 0$	$\to 0$	$\to 0$	$\to 0$
Observational study (representative data set, well matched)	≈ 0	≈ 0	≈ 0	$\neq 0$
Observational study (unrepresentative but partially correctable data, well matched)	≈ 0	$\neq 0$	≈ 0	$\neq 0$
Observational study (unrepresentative data set, well matched)	$\neq 0$	$\neq 0$	≈ 0	$\neq 0$

†For column Q, '$\to 0$' denotes $E(Q) = 0$ and $\lim_{n\to\infty}\{\text{var}(Q)\} = 0$, whereas '$\overset{\text{avg}}{=} 0$' indicates $E(Q) = 0$ for a design with a small n and so asymptotic limits are not relevant. Quantities in the columns marked Δ_{S_X} and Δ_{S_U} can be set to 0 if the quantity of interest is changed from PATE to SATE. Matching is normally designed to estimate PATT or SATT and so designs using it should be read as affecting components in equation (10) rather than equation (4).

Of course, for numerous logistical reasons, ideal experiments are rarely run in practice, and many other research designs are used, depending on the constraints that are imposed by the research situation. For example, in the most common form of *randomized clinical trials* in medicine, n is small, the sample is not drawn randomly and not from a known population of interest, treatment is randomly assigned and blocking is only sometimes used. The randomized breast cancer study is one such example, as it was carried out by using 237 non-randomly selected women who agreed to be in the trial and who were randomly assigned to treatment with some blocking.

In these trials, researchers must admit that $\Delta_S \neq 0$, although they sometimes sidestep the problem by switching their quantity of interest from PATE to SATE, and inferring to PATE only after their results have been replicated in a different setting, perhaps by different research teams. Researchers then are left basing a claim that $\Delta_S \approx 0$ on the hope or argument that their subjects are sufficiently similar to subjects everywhere ('a kidney is a kidney is a kidney is a kidney ...') and so NATE $\approx$ SATE; this claim is somewhat more plausible if estimates from replications in diverse settings are relatively constant, but as seems to be recognized the generalizations wind up depending on qualitative arguments rather than statistical science. As with partially correctable observational studies, randomized clinical trials sometimes select patients according to some known characteristics X and some unknown; in this situation, Δ_{S_X} can equal 0 if a weighted difference in means is used instead of D, but even in this situation Δ_{S_U} is not 0 exactly, in the limit, or on average.

Randomized clinical trials that block on all the information in X benefit directly because $\Delta_{T_X} = 0$. However, medical researchers often block on only a few variables and so $\Delta_{T_X} \neq 0$

and of course $\Delta_{\mathrm{T}U} \neq 0$. Nevertheless, random assignment means that on average these error components vanish, i.e. $E(\Delta_{\mathrm{T}X}) = E(\Delta_{\mathrm{T}U}) = 0$. Since n is small in most of these (including most academic works as well as most phase I and II clinical trials), these expectations alone are not so comforting, but, since the practice in this field is for many researchers to replicate roughly the same experiments, the concept of $\Delta_{\mathrm{T}X}$ and $\Delta_{\mathrm{T}U}$ vanishing on average across many repeated experiments is plausible.

In *social science field experiments*, researchers typically have large non-random convenience samples, such as from one city, non-governmental organization or company to which they were able to gain access and permission to run the experiment. They may or may not use blocking, but they can randomly assign the values of the treatment variable. For example, Gerber and Green (2000) conducted a voter mobilization experiment containing many one- and two-voter households in New Haven. In these settings, since $\Delta_{\mathrm{S}} \neq 0$ and replication of the experiment is not common, often the best that researchers can do with regard to sample selection error is to settle for estimating SATE rather than PATE. If they use complete blocking before random assignment, $\Delta_{\mathrm{T}X} = 0$ and not otherwise. However, random assignment with a large n means that $E(\Delta_{\mathrm{T}X}) = E(\Delta_{\mathrm{T}U}) = 0$ and the variances of both $\Delta_{\mathrm{T}X}$ and $\Delta_{\mathrm{T}U}$ drop as n increases.

A related research design involves *survey experiments*, where a large number of randomly selected respondents from a known population of interest are randomly assigned to treatment and control groups (the treatment in such studies often being different questionnaire wordings). This design is also sometimes used in public policy experiments when the population of interest is known. One example is the Job Corps evaluation, in which all applicants to the programme were randomly assigned to the treatment group and were therefore allowed to enrol in Job Corps at that time, or to the control group, where they were not allowed to enrol at that time (Schochet *et al.*, 2003). If the sample was properly drawn, $E(\Delta_{\mathrm{S}X}) = E(\Delta_{\mathrm{S}U}) = 0$ with a small variance tending towards 0. Unfortunately, random sampling of survey respondents is becoming increasingly difficult with the rise in cell phones and unit non-response. Blocking is rarely used in these experiments unless respondents' characteristics are collected before the experiment, and so $\Delta_{\mathrm{T}X} \neq 0$, but $\Delta_{\mathrm{T}X}$ and $\Delta_{\mathrm{T}U}$ both equal 0 in expectation and have small variances.

Finally, purely *observational studies* typically have large samples that are often randomly selected, but blocking and random assignment are infeasible. The last three rows of Table 2 include a summary of results for three general categories of observational studies. The first includes data that are representative of a fixed population, such as from a random sample. The second is not a random sample but includes enough information to correct for unrepresentativeness, such as via weighting. The third is based on a convenience sample with no known relationship to the population of interest. All three data types in Table 2 are assumed to contain data that make high quality matching possible.

As an example of the three types of observational studies, the General Accounting Office breast cancer researchers were interested in comparing breast cancer treatments among women who would not necessarily choose to have their treatment selected randomly. To study that question, nearly all women with breast cancer in five states and four metropolitan areas were included, but the women chose which treatment to receive (US General Accounting Office, 1994). In these studies, Δ_{S} is 0 or reasonably close to it exactly or in expectation. When the population differs from the sample, SATE is a sufficiently interesting quantity on its own that its difference from PATE becomes a definitional matter of minor importance. Studies that select on the basis of variables that are known in part, and corrected via weighting, imputation or some other procedure, can eliminate or reduce $\Delta_{\mathrm{S}X}$ but of course cannot affect $\Delta_{\mathrm{S}U}$, except in as much as X and U are related. Much of the work in observational studies goes into collecting the best pretreatment covariates, and adjusting for them after the data have been collected.

If adjustment is done well, $\Delta_{T_X} \approx 0$, but unfortunately in general $\Delta_{T_U} \neq 0$, and the absence of random assignment means that these studies cannot avoid error due to U even on average or as n grows. The hope of these researchers is that enough is known from 'theory', prior observational studies or qualitative evidence ('clinical information') that an assumption of ignorability or no omitted variable bias is sufficiently close for reasonably accurate inferences.

5.4. What is the best design?

If an ideal experimental design is infeasible, which of the remaining research designs is best? This question is not directly material, since medical researchers cannot randomly select subjects to administer medical procedures and those conducting observational studies of, say, the US Congressional elections cannot randomly assign incumbency status to candidates for public office. However, none of these procedures reduces all four components of estimation error to 0 with certainty.

From this perspective, the Achilles heel of observational studies is error due to imbalance in unobserved variables, whereas in experimental studies it is a small n and the lack of random selection. The estimation error in either can overwhelm all the good that these research designs otherwise achieve, but both approaches have ways of attacking their biggest weaknesses. Neither is better; both are adapted as well as possible to the constraints of their subjects and research situation. Experimentalists may envy the large, randomly selected samples in observational studies, and observationalists may envy the ability of experimentalists to assign treatments randomly, but the good of each approach comes also with a different set of constraints that cause other difficulties.

6. Fallacies in experimental research

Numerous experimental researchers across many fields make two mistakes that are easy to understand and correct with reference to our decomposition of estimation error.

First, experimentalists often fail to block at all, whereas any observed covariates should be fully blocked if feasible. The common practice of rerandomizing, when the first set of random draws for treatment assignments is unsatisfactory, can be thought of as an inefficient form of blocking. To see this, note that rerandomizing is equivalent to rejection sampling, where sampling from a known unrestricted distribution and discarding any samples that do not meet desired restrictions are equivalent to sampling directly from the restricted population.

Blocking of course is not always feasible, such as when patients in a medical experiment trickle in over time and treatment decisions need to be made for each quickly (as may have been so in Lichter *et al.* (1992)), or when important pretreatment covariates cannot be measured until after treatment. However, when feasible, blocking on potentially confounding covariates should always be used. As Box *et al.* (1978), page 103, wrote 'block what you can and randomize what you cannot'. Randomization is remarkable because it can eliminate imbalance on all covariates in expectation, even if those covariates are unobserved. But randomization without blocking is incapable of achieving what blocking can, which is to eliminate one component of estimation error entirely, setting $\Delta_{T_X} = 0$, rather than merely ensuring that $E(\Delta_{T_X}) = 0$. Since individual researchers care about obtaining the right answer in their experiment, rather than on average over their career or on average across different researchers in the scientific community, failing to block on an observed covariate can be a huge missed opportunity. Greevy *et al.* (2004) pointed out that algorithms have been developed to make blocking on many covariates considerably

easier than it once was, and that blocking even on irrelevant variables introduces no inferential problems, although it may reduce statistical power or efficiency relative to better chosen blocking covariates.

Second, experimenters who block on some or all available covariates and then randomize sometimes evaluate the balance of the treated and control groups by conducting various hypothesis tests, such as the difference in means. Senn (1994), page 1716, explained that this 'common procedure' is 'philosophically unsound, of no practical value, and potentially misleading'. He wrote:

> '1. over all randomizations the groups are balanced; 2. for a particular randomization they are unbalanced. No 'significant imbalance' can cause 1 to be untrue and no lack of significant balance can make 2 untrue. The only reason to employ such a test must be to examine the process of randomization itself. Thus, a significant result should lead to the decision that the treatment groups have not been randomized, and hence either that the trialist has ... dishonestly manipulated the allocation or that some incompetence ... has occurred.'

Any other purpose for conducting such a test is fallacious. Inappropriate randomization may be more often an issue in social science field experiments than in medical research, as the social scientist often conducts and implements random assignment only through a third party such as a government, firm or other organization (Imai, 2005).

These points are easy to understand by using our decomposition, since under random assignment $E(\Delta_{\mathrm{T}X}) = E(\Delta_{\mathrm{T}U}) = 0$, but for unblocked randomization $\Delta_{\mathrm{T}X} \neq 0$ (and of course $\Delta_{\mathrm{T}U} \neq 0$ under random assignment with or without blocking). Hypothesis tests are used to evaluate expectations, which we know are 0 owing to randomization, but are not needed to evaluate the components of estimation error, which can be calculated directly, in sample, and without any need for averaging over random sampling from a superpopulation or repeated experiments. Moreover, even if the population from which X comes is sampled from a superpopulation, $\Delta_{\mathrm{T}X}$ and not its expectation is the relevant component of estimation error, and the difference in the empirical cumulative distribution function between the treated and control groups is a directly observable feature of the sample. So hypothesis tests in this circumstance have no relevant role. This point is also central for a related fallacy that arises in matching, to which we now turn, and for which results that we give are also relevant for experimenters.

7. The balance test fallacy in matching studies

7.1. Matching

From the perspective of our decomposition, the only purpose of matching and blocking is to reduce imbalance in the observables $\Delta'_{\mathrm{T}X}$, and in any portion of imbalance in the unobservables $\Delta'_{\mathrm{T}U|X}$ for which U and X are related. Although blocking is easy to apply whenever the variables to block on are observed and treatment assignment is under the control of the investigator, matching requires sometimes difficult searching to find the best matches in the available data (Rosenbaum, 2002; Rubin, 2006). Matching also operates by deleting (or duplicating) observations and so, to keep the quantity of interest fixed during this process, researchers typically focus on PATT or SATT and try to keep the treated group fixed.

Matching is not a method of estimation, and so any application of it must be followed by a simple difference in means of the outcome or some other method. In the best case, the data exactly match and so satisfy equation (9) so $\Delta'_{\mathrm{T}X} = 0$, without losing too many observations in the process. In this best case of exact matching, T and X are unrelated in the matched sample, and no further adjustments for X are necessary, and so the PATT or SATT can be estimated by the simple difference in means, D. When imbalance $\Delta'_{\mathrm{T}X}$ is not eliminated, further adjustment for

X after matching may be necessary, such as via the same parametric methods as are commonly applied when matching is not applied. Since methodological work on matching is growing fast, the list of available matching algorithms from which to choose is also growing (Ho *et al.*, 2007).

Choosing the most appropriate algorithm for a given problem involves assessing how well equation (9) holds in the matched samples. Ideally that would involve comparing the (joint) empirical distributions of all covariates X between the matched treated and control groups. However, when X is high dimensional, this is generally infeasible and thus lower dimensional measures of balance are used instead. Standard practice in observational studies is for researchers to evaluate an implication of equation (9) for the chosen matching algorithm by conducting t-tests for the difference in means for each variable in X between the matched treated and control groups, thus seemingly addressing imbalance in at least one important aspect of a high dimensional relationship. Other hypothesis tests, such as χ^2-, F- and Kolmogorov–Smirnov tests, are also sometimes used for each covariate, but the same problems as those which we describe below still apply, and so for expository purposes we focus on the most commonly used t-test.

7.2. The balance test fallacy

The practice of using hypothesis tests to evaluate balance is widespread and includes a large volume of otherwise high quality work in economics (Smith and Todd, 2005), political science (Imai, 2005), sociology (Lundquist and Smith, 2005), psychology (Haviland and Nagin, 2005), education (Crosnoe, 2005), management science (Villalonga, 2004), medicine (Mangano *et al.*, 2006), public health (Novak *et al.*, 2006) and statistics (Lu *et al.*, 2001). Tables of t and other test statistics and/or their p-values are used as a justification in these and other references for the adequacy of the chosen matching method, and statistically insignificant t-tests are used as a stopping rule for maximizing balance in the search for the appropriate matched sample from which to draw inferences. Although we do not trace the exact consequences of this practice in the aforementioned studies, this approach is problematic for at least four reasons.

First, as an illustration, consider a data set on the 'School dropout demonstration assistance program' which sought to reduce drop-out rates by a series of school 'restructuring' initiatives, including curriculum reform and expanding teacher training (Stuart and Rubin, 2007). The design is observational, with a school with the restructuring effort compared with a control school. We use a subset of these data that includes 428 students from a treated school and 434 from a control school. The outcome variable Y is a test score (on a scale from 0 to 100), and X includes a variety of variables but we focus here only on the baseline mathematics test score. Matching analysis begins with the full data set and then selectively deletes students until equation (9) is best satisfied without losing too many observations. Suppose instead that we choose a matching algorithm that chooses observations from the control group to discard *randomly*, rather than (as usual) to maximize balance, i.e. we literally draw observations from the control group with equal probability and discard them from the data. Clearly, this algorithm would not affect expected balance between the treated and control group, or the bias in the ultimate analysis that satisfying equation (9) is meant to improve. In other words, on average across different randomly generated deletions, $\Delta'_{\mathrm{T}X}$ would not drop. Yet, we can show that randomly deleting observations seems to do wonders according to the t-test. To do this, we create a sequence of matching solutions that randomly drop different numbers of control observations (with results averaged over 5000 draws) and plot the average results in Fig. 1(a) (we discuss Fig. 1(b) later). The horizontal axis in Fig. 1(a) reports the number of control units that are randomly dropped, whereas the vertical axis gives the size of the t-test. We have shaded in the area below a t-test of 2, which is the region in which results are conventionally referred to as 'statistically

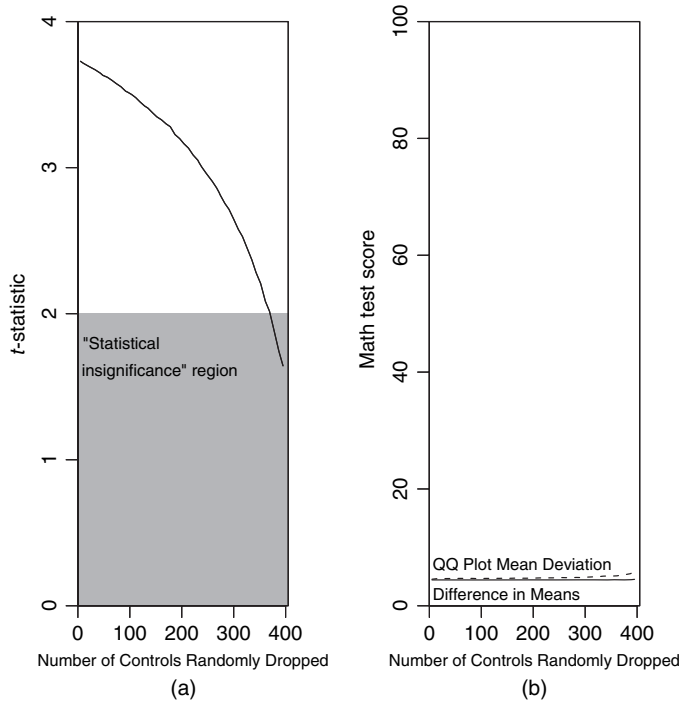

Fig. 1. Dangers in relying on t-statistics as a measure of balance (average value of a measure of balance when a given number of control units are randomly dropped from the data set (out of a total of 434)): with larger numbers of control units dropped (i.e. smaller numbers of control units in the resulting sample), the value of the t-statistic becomes closer to 0, falsely indicating improvements in balance, even though true balance does not vary systematically across the data sets (and efficiency declines); the difference in means and quantile–quantile plot mean deviation, which are given in (b), correctly indicate no change in bias as observations are randomly dropped

insignificant'. The curve on the plot clearly shows that, according to the t-test, randomly dropping more control units does an 'excellent' job at achieving balance, reducing the statistic from 3.7 to 1.6 in Fig. 1(a). This of course makes no sense at all.

Second, the problem in Fig. 1 can be seen by recognizing that dropping observations can influence not only balance but also statistical power, and unfortunately the t-test, like most statistical tests, is a function of both. The more observations that are dropped, the less power the tests have to detect imbalance in observed covariates. Formally, let n_{mt} and n_{mc} be the sample sizes for the matched treated and matched control groups, and define $r_{\mathrm{m}} = n_{\mathrm{mt}}/n_{\mathrm{m}}$ where $n_{\mathrm{m}} = n_{\mathrm{mt}} + n_{\mathrm{mc}}$. Then, write the two-sample t-test statistic with unknown and unequal variances as

$$\frac{\sqrt{n_{\mathrm{m}}}(\bar{X}_{\mathrm{mt}} - \bar{X}_{\mathrm{mc}})}{\sqrt{\{s^2_{\mathrm{mt}}/r_{\mathrm{m}} + s^2_{\mathrm{mc}}/(1 - r_{\mathrm{m}})\}}}$$

where $\bar{X}_{\mathrm{mt}} = \Sigma_{i=1}^{n_{\mathrm{m}}} T_i X_i / n_{\mathrm{mt}}$ and $\bar{X}_{\mathrm{mc}} = \Sigma_{i=1}^{n_{\mathrm{m}}} (1 - T_i) X_i / n_{\mathrm{mc}}$ are the sample means, and

$$s^2_{\mathrm{mt}} = \frac{\sum_{i=1}^{n_{\mathrm{m}}} T_i (X_i - \bar{X}_{\mathrm{mt}})^2}{n_{\mathrm{mt}} - 1}$$

and

$$s^2_{\text{mc}} = \frac{\sum_{i=1}^{n_{\text{m}}} (1 - T_i)(X_i - \bar{X}_{\text{mt}})^2}{n_{\text{mc}} - 1}$$

represent the sample variances of the matched treated and control groups respectively. Hence, the difference in sample means as a measure of balance is distorted in the t-test by three factors:

(a) the total number of remaining observations n_{m},
(b) the ratio of remaining treated units to the total number of remaining observations r_{m} and
(c) the sample variance of X for the remaining treated and control units, s^2_{mt} and s^2_{mc}.

Since the values of (this and other) hypothesis tests are affected by factors other than balance, they cannot even be counted on to be monotone functions of balance. The t-test can indicate that balance is becoming better whereas the actual balance is growing worse, staying the same or improving. Although we choose the most commonly used t-test for illustration, the same problem applies to many other test statistics that are used in applied research. For example, the same simulation applied to the Kolmogorov–Smirnov test shows that its p-value monotonically increases as we randomly drop more control units. This is because a smaller sample size typically produces less statistical power and hence a larger p-value.

Third, from a theoretical perspective, balance is a characteristic of the sample, not some hypothetical population, and so, strictly speaking, hypothesis tests are irrelevant in this context (Ho *et al.*, 2007). Whether the quantity of interest is SATT, PATT or SPATT, balance affected by matching affects only Δ'_{T_X}. Virtually all methods of adjustment condition on the observed values of X, and so X can be dropped in these analyses only when equation (9) is satisfied in sample, not in some population from which the data are hypothetically or actually drawn. For the same reason that randomized blocks or paired matching are preferable to classical randomization in experimental design—i.e. the imbalance in the variables that defines the blocks can be set to zero in sample, without having to hope that the sample size is sufficiently large for the advantages of randomization to kick in (see also Greevy *et al.* (2004), page 264)—matching on all observed differences in X is preferable whenever feasible and other goals such as variance reduction are not harmed. The goal of reducing estimation error is reducing Δ'_{T_X} and not merely $E(\Delta'_{\text{T}_X})$, and so imbalance with respect to observed pretreatment covariates—the difference between $\tilde{F}(X|T=1, I=1)$ and $\tilde{F}(X|T=0, I=1)$—should be minimized without limit where possible, so long as we are not unduly compromising other goals in the process (such as efficiency).

Finally, we offer a simple model that conveys why matching contains no threshold below which the level of imbalance is always acceptable. To see this, consider data that are generated by the classical regression model, $E(Y|T, X) = \theta + T\beta + X\gamma$ (Goldberger, 1991), a special case of the model in equations (3). Then the regression of Y on a constant and T (without X) gives a difference in means, the (conditional) bias of which as an estimate of β is $E(\hat{\beta} - \beta|T, X) = G\gamma$, where G contains vectors of coefficients from regressions of each of the variables in X on a constant and T. Using matching to eliminate bias under this simplified data generation process involves dropping or repeating observations so that G is as close to a matrix of 0s as possible. But what happens to bias if G is smaller than it was before matching but still not 0? The answer is that the bias is reduced, but without knowledge of γ—which researchers eschew estimating to avoid inadvertently introducing selection error by choosing matching solutions that stack the deck for their favoured hypotheses—it could be that a non-zero portion of G, when multiplied by its corresponding elements of γ, will generate arbitrarily large bias. This also shows that no

measure of balance which is a function of X alone can be guaranteed to be a monotone function of bias without special assumptions (Rubin and Stuart, 2006), and so proper measures of imbalance should always be minimized without limit, subject to efficiency constraints. Thus, whether or not some hypothesis tests indicate that G is not significantly different from 0 is immaterial: the smaller G is the better, either above or below the t-test threshold of statistical significance, since G (i.e. balance) is a characteristic of the observed data.

An argument that is related to our last point has been made in the context of randomized experiments, where researchers have shown that even small (and statistically insignificant) differences in important (or what they call in this literature 'prognostic') covariates can result in large differences in the results of the experiment (Senn, 1994; Pocock *et al.*, 2002). However, the problem that we describe in this section with statistical power and stopping rules being a function of the remaining sample size does not arise in randomized experiments.

Researchers analysing data from randomized experiments that did not block on all observed covariates can check balance, but they do not need hypothesis tests to do so. The issue of balance is entirely in sample and involves no inference to populations or superpopulations. Thus, everything that is needed to check balance and to determine it directly is available (Cochran, 1965). Issues of sample selection and sampling bias arise only in Δ_{S}; in contrast, Δ_{T} always involves just the sample at hand, whether your perspective is sample based, population based or superpopulation based. If the samples are not balanced, then researchers may wish to settle for Δ_{T} being close to 0 in expectation or they can adjust. Adjustment will improve balance and thus reduce $\Delta_{\mathrm{T}X}$, but if not done properly can be at the expense of estimation variance or bias. Normally, however, even matching on irrelevant covariates will only slightly increase the variance (Rubin and Thomas, 1996; Ho *et al.*, 2007).

7.3. Better alternatives

In any study where all observed covariates were not fully blocked ahead of time, balance should be checked routinely by comparing observed covariate differences between the treated and control groups. Any statistic that is used to evaluate balance should have two key features:

(a) it should be a characteristic of the sample and not of some hypothetical population and
(b) the sample size should not affect the value of the statistic.

If matching is used, the difference between the groups should be minimized without limit. A difference in means is a fine way to start. Cochran (1968) suggested a rule of thumb that a mean difference should not differ by more than a quarter of a standard deviation, though we emphasize that imbalance should be minimized without limit. Other options include higher order moments than the mean, non-parametric density plots and propensity score summary statistics (e.g. Austin and Mamdani (2006) and Rubin (2001)).

A more general approach is quantile–quantile (or 'QQ'-) plots that directly compare the empirical distribution of two variables, although statistics that are based on QQ-plots can be sensitive to small features of the data. Fig. 1(b) plots for comparison the difference in means and a QQ-plot summary statistic, the average distance between the empirical quantile distributions of the treated and control groups calculated over the same samples as for Fig. 1(a). (Formally, this measure can be defined as $(1/n)\Sigma_{i=1}^{n}|\tilde{q}_{X_{\mathrm{mt}}}(i/n) - \tilde{q}_{X_{\mathrm{mc}}}(i/n)|$ where $\tilde{q}_{X_{\mathrm{mt}}}$ and $\tilde{q}_{X_{\mathrm{mc}}}$ are the empirical quantile functions of a covariate X for the matched treated and matched control groups respectively and $n = \min(n_{\mathrm{mt}}, n_{\mathrm{mc}})$.) Unlike the t-test, the level of balance does not change for either statistic as more units are randomly dropped. These statistics are by no means perfect, but they and the many other possibilities do not have the flaw that we show hypothesis

tests have when used as a stopping rule for assessing balance. As is widely recognized, we also ultimately need better ways of comparing two multidimensional empirical distributions, but these should be sample quantities, not hypothesis tests.

Although these results indicate that future researchers should not use hypothesis tests as a balance stopping rule, a reasonable question is how to interpret the considerable volume of published literature that does so without reporting better balance statistics. One interpretation would be that published tables which report small p-values or large t-tests should cause readers to worry about balance, whereas the reverse would not suggest any level of comfort. In studies with small numbers of observations and thus larger p-values, low levels of imbalance relative to the unobserved importance of the covariates might be acceptable if the bias induced is swamped by the uncertainty of the ultimate quantity of interest at the analysis stage; however, because importance is unobserved, the threshold 'low level' is not defined, and so p-value cut-offs (e.g. significance at the 0.05-level) are not of use for this purpose. Best practice should be to minimize imbalance for all covariates, by using measures like those described above, and then to adjust parametrically for any remaining differences (Ho *et al.*, 2007).

8. Concluding remarks

Random selection and random assignment—which enable researchers to avoid some statistical assumptions—along with matching and blocking—which adjust non-parametrically for heterogeneity and potential confounders—are among the most practically useful ideas in the history of statistical science. At times, they are also among the most misunderstood. We have tried to lay out some of the key issues in this paper so that they will be more transparent to all, and so that future researchers from both experimental and observational research traditions will be able to avoid the fallacies of causal inference to which many have previously fallen prey.

Of course, our decomposition and analysis describe the basic contributions of each approach and do not attempt to control for the many sophisticated data problems that inevitably arise in a wide range of statistical research in the real world. For example, even in the best experimental work, some information goes missing, randomly selected subjects sometimes refuse to participate, some subjects do not comply with treatment assignments, random numbers do not always become assigned as planned or must be assigned at a more aggregated level than desired and outcomes are not always measured correctly or recorded appropriately. To account for these problems, when they cannot be fixed through better data collection, more sophisticated methods become necessary. But, throughout that more advanced work, the more basic issues that have been discussed here should remain at the forefront.

Acknowledgements

Our thanks go to Alberto Abadie, Neal Beck, Jack Buckley, Alexis Diamond, Felix Elwert, Andrew Gelman, Ben Hansen, Guido Imbens, Paul Rosenbaum, Don Rubin, Jas Sekhon, Chris Winship and the participants of the 'Northeast political methodology program' conference for many helpful comments, the National Institutes of Aging (P01 AG17625-01), the National Institute of Mental Health and the National Institute of Drug Abuse (MH066247), the National Science Foundation (SES-0318275, IIS-9874747 and SES-0550873) and the Princeton University Committee on Research in the Humanities and Social Sciences for research support. We also thank the Joint Editor of the journal for detailed comments which significantly improved the presentation of the results that are given in this paper.

Appendix A: Efficiency of adding blocking to random treatment

We first show that, if the estimand is SPATE, then blocking always improves classical randomization in terms of statistical efficiency. Suppose that blocking is done on the variable X whose support is $\mathcal{X}$. Then, the variances of D under classical randomization and blocking are given by

$$\mathrm{var}^{\mathrm{C}}(D) = \frac{2}{n}[\mathrm{var}\{Y(1)\} + \mathrm{var}\{Y(0)\}],$$

$$\mathrm{var}^{\mathrm{B}}(D) = \frac{2}{n}\sum_{x\in\mathcal{X}} w_x[\mathrm{var}_x\{Y(1)\} + \mathrm{var}_x\{Y(0)\}],$$

where $\mathrm{var}(\cdot)$ represents the (super)population variance, $\mathrm{var}_x(\cdot)$ represents the conditional (super)population variance with covariate value $X_i = x$ and w_x is the known (super)population weight for the units with $X_i = x$. Then, if we define the within-block mean as $\overline{Y(t)}_x \equiv E\{Y(t)|X = x\}$ for $t = 0, 1$, we have

$$\begin{aligned}\mathrm{var}\{Y(t)\} &= E[\mathrm{var}_x\{Y(t)\}] + \mathrm{var}\{\overline{Y(t)}_x\} \\ &\geqslant E[\mathrm{var}_x\{Y(t)\}] = \sum_{x\in\mathcal{X}} w_x \,\mathrm{var}_x\{Y(t)\}.\end{aligned}$$

Thus, it follows that the variance under blocking is smaller than or equal to the variance under classical randomization, i.e. $\mathrm{var}^{\mathrm{C}}(D) \geqslant \mathrm{var}^{\mathrm{B}}(D)$.

Next, we consider the case where the estimand is PATE. In this case, the variances of D under complete randomization and blocking are given by

$$\mathrm{var}^{\mathrm{C}}(D) = \frac{1}{n}[2\,\mathrm{var}\{Y(1)\} + 2\,\mathrm{var}\{Y(0)\} - \mathrm{var}(\mathrm{TE})], \tag{11}$$

$$\mathrm{var}^{\mathrm{B}}(D) = \frac{1}{n}\sum_{x\in\mathcal{X}} w_x[2\,\mathrm{var}_x\{Y(1)\} + 2\,\mathrm{var}_x\{Y(0)\} - \mathrm{var}_x(\mathrm{TE})], \tag{12}$$

where $\mathrm{var}(\cdot)$ ($\mathrm{var}_x(\cdot)$) now represents the finite (conditional) population variance, and $w_x = n_x/n$ with n_x being the number of observations with $X_i = x$. Note that the third term in each of the variance expressions cannot be estimated from the data. (However, if the treatment effect is constant across units, i.e. $Y_i(1) - Y_i(0)$ for all i, this term will be 0.)

Now, for any variable δ and finite sample size n, the standard analysis-of-variance formula implies that

$$(n-1)\,\mathrm{var}(\delta) = \sum_{x\in\mathcal{X}}\{(n_x - 1)\,\mathrm{var}_x(\delta) + n_x(\bar{\delta}_x - \bar{\delta})^2\},$$

where $\bar{\delta}_x = \Sigma_{i\in\{X_i=x\}}\delta_i/n_x$, and $\bar{\delta} = \Sigma_{i=1}^n \delta_i/n$. Then,

$$\mathrm{var}(\delta) = \sum_{x\in\mathcal{X}}\left(w_x - \frac{1-w_x}{n-1}\right)\mathrm{var}_x(\delta) + \left(w_x + \frac{w_x}{n-1}\right)(\bar{\delta}_x - \bar{\delta})^2.$$

Applying this result and after some algebra, the difference between equations (11) and (12) can be written as

$$\mathrm{var}^{\mathrm{C}}(D) - \mathrm{var}^{\mathrm{B}}(D) = \Theta_{\mathrm{B}} - \Theta_{\mathrm{W}}, \tag{13}$$

where the two components Θ_{B} and Θ_{W} are closely related to the between-block variation and the within-block variation of the potential outcomes respectively. They are defined as

$$\Theta_{\mathrm{B}} = \frac{m-1}{m(n-1)}\,\mathrm{var}_w\left\{\overline{Y(1)}_x + \overline{Y(0)}_x\right\},$$

$$\Theta_{\mathrm{W}} = \frac{1}{n(n-1)}\sum_{x\in\mathcal{X}}(1-w_x)\,\mathrm{var}_x\{Y(1) + Y(0)\},$$

where m is the number of blocks, and

$$\mathrm{var}_w(\delta) = \frac{m \sum_{x \in \mathcal{X}} w_x (\delta_x - \bar{\delta}_x)^2}{(m-1) \sum_{x \in \mathcal{X}} w_x}$$

is the weighted variance between blocks.

Equation (13) gives the exact expression for the gain in efficiency due to blocking. If we assume that m stays constant while n grows, the first term, which is positive and $o(1)$, dominates the second term, which is negative but $o(n^{-1})$. Hence, unless the sample size is small, blocking improves efficiency. Moreover, applying the central limit theorem, the asymptotic variances under classical randomization and blocking are given by $n\,\mathrm{var}^{\mathrm{C}}(D)$ and $n\,\mathrm{var}^{\mathrm{B}}(D)$ respectively. Then, the difference between these two asymptotic variances equals

$$n\Theta_{\mathrm{B}} = \frac{m-1}{m}\,\mathrm{var}_w\{\overline{Y(1)}_x + \overline{Y(0)}_x\},$$

which is always positive. Thus, blocking is asymptotically more efficient than classical randomization when PATE is the estimand. Results similar to those given in this appendix can also be derived for the matched pair design (see Imai (2007)).

References

Austin, P. C. and Mamdani, M. M. (2006) A comparison of propensity score methods: a case-study estimating the effectiveness of post-ami statin use. *Statist. Med.*, **25**, 2084–2106.

Box, G. E., Hunter, W. G. and Hunter, J. S. (1978) *Statistics for Experimenters*. New York: Wiley-Interscience.

Cochran, W. G. (1965) The planning of observational studies of human populations (with discussion). *J. R. Statist. Soc.* A, **128**, 234–265.

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

Cochran, W. and Cox, G. (1957) *Experimental Designs*. New York: Wiley.

Crosnoe, R. (2005) Double disadvantage or signs of resilience?: the elementary school contexts of children from mexican immigrant families. *Am. Educ. Res. J.*, **42**, 269–303.

Fisher, R. A. (1935) *The Design of Experiments*. London: Oliver and Boyd.

Gerber, A. S. and Green, D. P. (2000) The effects of canvassing, telephone calls, and direct mail on voter turnout: a field experiment. *Am. Polit. Sci. Rev.*, **94**, 653–663.

Goldberger, A. (1991) *A Course in Econometrics*. Cambridge: Harvard University Press.

Greevy, R., Lu, B., Silver, J. H. and Rosenbaum, P. (2004) Optimal multivariate matching before randomization. *Biostatistics*, **5**, 263–275.

Haviland, A. M. and Nagin, D. S. (2005) Causal inferences with group based trajectory models. *Psychometrika*, **70**, 557–578.

Heckman, J. J., Ichimura, H., Smith, J. and Todd, P. (1998) Characterizing selection bias using experimental data. *Econometrika*, **66**, 1017–1098.

Ho, D., Imai, K., King, G. and Stuart, E. (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.*, **15**, 199–236.

Imai, K. (2005) Do get-out-the-vote calls reduce turnout?: the importance of statistical methods for field experiments. *Am. Polit. Sci. Rev.*, **99**, 283–300.

Imai, K. (2007) Variance identification and efficiency analysis in experiments under the matched-pair design. *Technical Report*. Department of Politics, Princeton University, Princeton.

Imai, K., King, G. and Nall, C. (2007) The essential role of pair-matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Technical Report*. Department of Politics, Princeton University, Princeton.

Imbens, G. W. (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev. Econ. Statist.*, **86**, 4–29.

King, G. and Zeng, L. (2006) The dangers of extreme counterfactuals. *Polit. Anal.*, **14**, 131–159.

Lichter, A. S., Lippman, Jr, M. E., Danforth, Jr, D. N., d'Angelo, T., Steinberg, S. M., deMoss, E., MacDonald, H. D., Reichert, C. M., Merino, M., Swain, S. M., Cowan, K., Gerber, L. H., Bader, J. L., Findlay, P. A., Schain, W., Gorrell, C. R., Straus, K., Rosenberg, S. A. and Glatstein, E. (1992) Mastectomy versus breast-conserving therapy in the treatment of stage I and II carcinoma of the breast: a randomized trial at the National Cancer Institute. *J. Clin. Oncol.*, **10**, 976–983.

Lu, B., Zanuto, E., Hornik, R. and Rosenbaum, P. R. (2001) Matching with doses in an observational study of a media campaign against drug abuse. *J. Am. Statist. Ass.*, **96**, 1245–1253.

Lundquist, J. H. and Smith, H. L. (2005) Family formation among women in the U.S. military: evidence from the NLSY. *J. Marrge Fam.*, **67**, 1–13.
Mangano, D. T., Tudor, J. C. and Dietzel, C. (2006) The risk associated with aprotinin in cardiac surgery. *New Engl. J. Med.*, **354**, 353–365.
Martin, D. C., Diehr, P., Perrin, E. B. and Koepsell, T. D. (1993) The effect of matching on the power of randomized community intervention studies. *Statist. Med.*, **12**, 329–338.
Novak, S. P., Reardon, S. F., Raudenbush, S. W. and Buka, S. L. (2006) Retail tobacco outlet density and youth cigarette smoking: a propensity-modeling approach. *Am. J. Publ. Hlth*, **96**, 670–676.
Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.
Pocock, S. J., Assmann, S. E., Enos, L. E. and Kasen, L. E. (2002) Subgroup analysis covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statist. Med.*, **21**, 2917–2930.
Rosenbaum, P. R. (2002) *Observational Studies*, 2nd edn. New York: Springer.
Rubin, D. B. (1973) Matching to remove bias in observational studies. *Biometrics*, **29**, 159–184.
Rubin, D. B. (1997) Estimating causal effects from large data sets using propensity scores. *Ann. Intern. Med.*, **127**, 757–763.
Rubin, D. B. (2001) Using propensity scores to help design observational studies: application to the tobacco litigation. *Hlth Serv. Outcms Res. Methodol.*, **2**, 169–188.
Rubin, D. B. (2006) *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press.
Rubin, D. B. and Stuart, E. A. (2006) Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *Ann. Statist.*, **34**, 1814–1826.
Rubin, D. B. and Thomas, N. (1996) Matching using estimated propensity scores, relating theory to practice. *Biometrics*, **52**, 249–264.
Schochet, P., McConnell, S. and Burghardt, J. (2003) National Job Corps study: findings using administrative earnings records data: final report. *Technical Report*. Mathematica Policy Research, Princeton.
Senn, S. (1994) Testing for baseline balance in clinical trials. *Statist. Med.*, **13**, 1715–1726.
Smith, J. A. and Todd, P. E. (2005) Does matching overcome lalonde's critique of nonexperimental estimators? *J. Econometr.*, **125**, 305–353.
Stuart, E. A. and Rubin, D. B. (2007) Matching with multiple control groups with adjustment for group differences. *J. Educ. Behav. Statist.*, to be published.
US General Accounting Office (1994) Breast conservation versus mastectomy: patient survival in day-to-day medical practice and randomized studies: report to the chairman, subcommitee on human resources and intergovernmental relations, committee on government operations, house of representatives. *Technical Report GAO-PEMD-95-9*. US General Accounting Office, Washington DC.
Villalonga, B. (2004) Does diversification cause the 'diversification discount'? *Finan. Mangmnt*, **33**, 5–27.