ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Modern Algorithms for Matching in Observational Studies

Paul R. Rosenbaum

Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA; email: rosenbaum@wharton.upenn.edu

Annu. Rev. Stat. Appl. 2020. 7:143–76

First published as a Review in Advance on August 16, 2019

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031219-041058

Keywords

assignment algorithm, causal inference, design sensitivity, integer programming, fine balance, Mahalanobis distance, near-fine balance, network optimization, optimal matching, principal unobserved covariate, propensity score, refined balance, sensitivity analysis

Abstract

Using a small example as an illustration, this article reviews multivariate matching from the perspective of a working scientist who wishes to make effective use of available methods. The several goals of multivariate matching are discussed. Matching tools are reviewed, including propensity scores, covariate distances, fine balance, and related methods such as near-fine and refined balance, exact and near-exact matching, tactics addressing missing covariate values, the entire number, and checks of covariate balance. Matching structures are described, such as matching with a variable number of controls, full matching, subset matching and risk-set matching. Software packages in R are described. A brief review is given of the theory underlying propensity scores and the associated sensitivity analysis concerning an unobserved covariate omitted from the propensity score.

1. INTRODUCTION: OBSERVATIONAL STUDIES, CAUSAL EFFECTS, IMPORTANCE OF DESIGN

Multivariate matching: one form of adjustment for observed covariates **x** in the design of an observational study

1.1. What Are Observational Studies of Treatment Effects?

Cochran (1965, p. 234) defined an observational study as an empirical investigation in which

> the objective is to elucidate cause-and-effect relationships [. . .and. . .] it is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover, or to assign subjects at random to different procedures. . . . Examples are the studies of the relationship between smoking and health, studies of factors that affect the probability of injuries in motor accidents, studies of the differences in behavior of school children under permissive and authoritarian regimes, and studies of the effects of new social programmes such as replacing slum housing by public housing.

A covariate is an attribute or quantity describing an individual prior to assignment to treatment or control. Absent random assignment, treated and control individuals may differ in terms of covariates, so direct comparison of the outcomes of treated individuals and controls may compare individuals who are not comparable—that is, a direct comparison may be biased as an estimate of the effect caused by the treatment. A covariate, **x**, is observed if its value is recorded; otherwise a covariate, u, is unobserved. From the data, we can see whether treated and control groups are comparable in terms of observed covariates, **x**, and we can often remove by adjustments the differences we can see, for instance by matching for these covariates. Unlike randomized experiments, most if not all observational studies face critical debate centered on the possibility that adjustments for observed covariates are inadequate to render comparable the treated and control groups, based on speculation that the groups differ in terms of unobserved covariates, u. Critical debate of this kind is part of virtually every observational study, not a failure of a particular study, but studies vary widely in their abilities to inform and address the inevitable debate. This review focuses on the first step, using multivariate matching to adjust for observed covariates, but the fate of an observational study is largely determined by whether its design and analysis adequately address potential bias from unobserved covariates (Rosenbaum 2010, 2015a, 2017b). Nonetheless, matching often facilitates that second step (see Section 5.4).

The line between observed covariates and unobserved covariates is practically and sharply defined by what is observed. If a covariate is observed with measurement error, then the fallible but observed value is an observed covariate in **x**, and the difference between the observed and true values is an unobserved covariate in u. If values of a covariate are sometimes missing, then the observed covariate is either its observed value or a blank, indicating that it is missing; then, another complementary unobserved covariate is either a blank, if the covariate was observed, or it is the unobserved missing value of the covariate. We may realistically ask that matching balance the observed values of observed covariates together with the observed pattern of missing data—for instance, that people in certain occupations more often decline to respond to a question about their incomes—but we cannot realistically ask matching to balance the missing values (Rosenbaum & Rubin 1984, appendix). Missing values for observed covariates occur in several ways in the example and are discussed further in Section 4.5.

1.2. Observational Studies Should Be Designed to Resemble Simple Experiments

In his paper of 1965, Cochran wrote, "The planner of an observational study should always ask himself the question: 'How would the study be conducted if it were possible to do it by controlled

experimentation?'" (Cochran 1965, p. 236). An observational study seeks to answer a question that might have been answered by an experiment, typically an experiment that cannot be conducted for ethical or practical reasons. The question is the same, the intended answer is the same, and it is a question about the world, not a question about a particular statistical model: How would a certain experiment turn out? This experimental question is difficult to answer in an observational study, because treatments were not randomly assigned to individuals.

Randomization: in experiments and clinical trials, assigning treatments by the independent flips of a fair coin

Effective design: a well-designed observational study does more to address the inevitable critical discussion of possible bias from unmeasured covariates

Among the basic tools of experimental design are blocking for observed covariates, counterbalancing observed covariates, and randomized treatment assignment to prevent bias from unobserved covariates. Blocking or pairing before randomization puts together individuals who are similar in terms of important observed covariates; for a modern method, readers are directed to Greevy et al. (2004). Sometimes individuals are, of necessity, different: Two locations in a farmer's field must be different locations. Where blocking to make individuals the same is impossible, counterbalancing, as in a Latin square, is used to prevent systematic patterns in the way treated and control groups differ; for instance, a Latin square on a farmer's field prevents assignment of treatment to all of the northern locations in the field. Finally, random assignment within blocks subject to counterbalancing constraints prevents bias from unmeasured covariates and provides the "reasoned basis for inference" about the effects caused by treatments, to use Fisher's (1935, chapter 2) phrase. An important aspect of Fisher's theory of inference in experiments is that randomization inferences require no modeling assumptions, no assumptions about sampling a population, and no assumption that randomization has succeeded in balancing unobserved covariates; rather, the inferences account for the imbalances in unobserved covariates that randomization may by chance produce.

Developing these thoughts, Rubin (2007, pp. 20, 26) wrote:

> Observational studies can and should be designed to approximate randomized experiments as closely as possible. In particular, observational studies should be designed using only background information to create subgroups of similar treated and control units, where 'similar' here refers to their distributions of background variables [i.e., covariates]. Of great importance, this activity should be conducted without any access to any outcome data, thereby assuring the objectivity of the design. . . Of course, objectivity is not the same as finding truth, but I believe that it is generally a necessary ingredient if we are to find truth.

Designing observational studies to resemble simple experiments has positive and negative aspects. One positive aspect emphasizes matching or blocking for observed covariates, either making the covariates similar within matched sets or counterbalancing covariates across different sets when they cannot be made similar within sets. Another positive aspect seeks situations—so-called natural experiments—in which haphazard and irrelevant factors, rather than careful, purposeful decisions, play a large role in deciding treatment assignments, thereby taking a small step in the direction of random assignment (Angrist & Krueger 1999, Meyer 1995, Sekhon 2009, Vandenbroucke 2004). There are negative aspects as well—things the investigator should not do. The investigator should not obscure the sources of uncertainty present in an observational study that would have been absent in a randomized experiment. The manner in which an observational study fails to resemble an experiment should be transparent, open to view, and open to responsible critical discussion.

1.3. Goals of Matching in Observational Studies

Matching has several goals, including the following.

- **Effective design:** Prior to collecting outcomes, in the design of an experiment, care is taken to structure the relationships between treatments and observed covariates, in part

to minimize confounding and aid transparency (see, e.g., Wu & Hamada 2011). In an observational study, matching is completed without access to outcomes, so it is part of the study's design (see Section 1.2). Just as one compares experimental designs before picking a satisfactory design, so too one compares several matched designs for an observational study, selecting a satisfactory design. Because outcomes are not available during this process, the search for a good design neither biases analyses of outcomes nor requires corrections for multiple inference.

- **Framing one primary analysis:** A randomized clinical trial includes a primary analysis that is described in the trial's protocol prior to collection of outcomes. A primary analysis does not preclude secondary and exploratory analyses; rather, it distinguishes such analyses. In parallel, a matched observational study has a primary analysis built into its design, typically consisting of a matched comparison of treated and control groups in terms of a primary outcome. Although John Tukey made many contributions to exploratory data analysis and to honesty in testing multiple hypotheses, he was also an advocate for focused, confirmatory analyses. Tukey (1980, p. 24) wrote:

> Important questions can demand the most careful planning for confirmatory analysis. . . .Preplan the main analysis (having even two main analyses may be too many!) . . .I see no real alternative, in most truly confirmatory studies, to having a single main question—in which a question is specified by ALL of design, collection, monitoring, and analysis.

- **Facilitating exploratory analyses:** Colin Mallows remarked, "The most robust method I know is to look at the data." You cannot look at observational data until you have adjusted for observed covariates; otherwise, you may be comparing infants to the elderly and princes to paupers. If you adjust using a model, then you end up looking at the model, not the data, because only the model is adjusted for covariates. In contrast, matching permits exploratory analysis and display of data adjusted for observed covariates. Matching does not preclude additional model-based adjustments of a matched sample when these are needed (Rubin 1979).
- **Framing critical discussion:** Observational studies may be greeted, and perhaps typically are greeted, with genuine skepticism or credible challenge. Critical discussion often raises the possibility that treated and control groups are not comparable, despite efforts to make them comparable. A matched study is simple in form, often transparent, so critical discussion is sometimes enlightening. A neutral audience may feel confident that it understands a matched observational study, hence also confident in judging, and perhaps rejecting, critical comments. In contrast, a neutral audience impressed, perhaps even awed, by elaborate methodology may be shaken by critical comments if it has a shaky understanding of that elaborate methodology. Addressing critical discussion has technical aspects (Rosenbaum 1991b, 2015a), but these technical aspects are likely to be more compelling to a neutral audience that feels confident in its understanding of the underlying observational study. As discussed in Section 5.4, decisions taken during the design of an observational study affect its ability to address critical discussion; that is, they affect the design sensitivity.

2. MOTIVATING EXAMPLE: ANTIDEPRESSANT MEDICATION AND BONE DENSITY IN ADOLESCENTS

2.1. Do Selective Serotonin Reuptake Inhibitors Reduce Bone Density?

There is concern that an important class of antidepressant medications, selective serotonin reuptake inhibitors (SSRIs), may have the side effect of reducing bone density. Feuer et al. (2015)

examined this possibility in adolescents using publicly available data from three National Health and Nutrition Examination Surveys (NHANES), 2005–2010, that obtained bone density measures for adolescents. In some of their analyses, they compared the total femur bone density of children aged 12 to 20 who had been receiving an SSRI for at least 180 days to that of other children not receiving an SSRI. This comparison will be used here to provide a tangible illustration of matching concepts and methods. In contrast, the reader should consult Feuer et al. (2015) and the references there for discussion of the effects of SSRIs on bone density. See the Appendix for details about use of the NHANES data in this example.

The observed covariates considered here are age (8 to 20 years); gender; black race; Hispanic ethnicity; family income recorded as a multiple of the poverty level and capped at five times the poverty level; serum cotinine in ng/mL, which is a marker of recent exposure to tobacco; and body mass index (BMI). As will be seen in detail later, when comparing the treated group of 49 SSRI users to the entire pool of 6,435 potential controls, the treated group had a much higher percentage of girls, a much lower percentage of blacks and Hispanics, higher family incomes, and a higher percentage with exposure to tobacco, but comparable BMIs. Depression in adolescents is sometimes associated with eating disorders, such as anorexia or bulimia, and diet could affect bone density. Smoking may affect appetite and bone density. Income, race, and ethnicity are often associated with various aspects of health, although the mechanism by which that association is produced is often opaque.

2.2. Before Matching, Treated and Control Groups Are Not Comparable

Figure 1*a* shows the distribution of age by gender in treated and control groups. Treated girls as a group are older than control girls, but treated boys are younger than control boys. The bias in

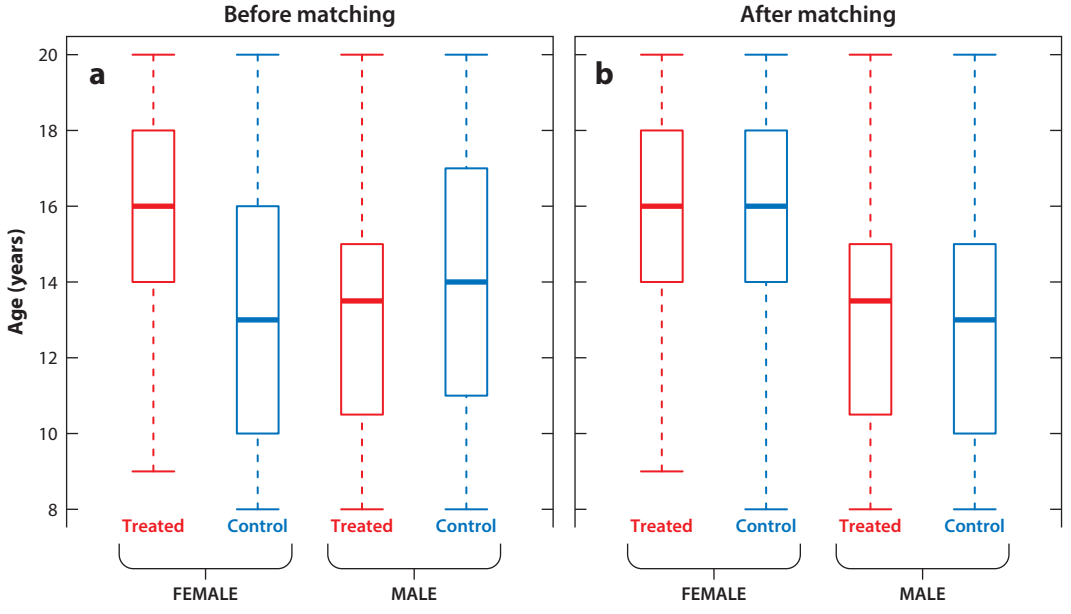

Figure 1

Age and gender for selective serotonin reuptake inhibitor users (treated) and nonusers (control), before and after matching. Before matching treated females are older than control females, while treated males are younger than control males, but this is corrected by matching 10 controls to each treated child.

age for girls is different from the bias in age for boys. Depression in adolescence, or its treatment, or both, may be different for boys and girls. So the control girls need to be adjusted to be older, while the control boys need to be adjusted to be younger.

Figure 1*a* shows that the lower quartile of age for control girls is 10 years old, but almost all of the treated girls are much older, so many controls look nothing like the treated children. Should control girls under 10 play a large role in estimating the effect of SSRIs on the bone density of much older girls, especially when there is an abundance of older girls in NHANES who can serve as controls? This pattern is greatly understated in **Figure 1*a*** because it focuses on just two covariates. Later, this pattern will become clearer, and also more dramatic, in **Figure 3** (discussed in the following section), where several covariates are considered at once.

Some investigators adjust for covariates using what is known as covariance adjustment, that is, some form of regression adjustment, say, linear least squares regression, logit regression, or the proportional hazards model, simply placing the individual covariates in the model along with a binary indicator of treated or control. Perhaps a complex modeling effort might succeed if skillfully executed, if it correctly modeled the interactions among the covariates and between covariates and treatment, but simply placing the covariates in the model one by one would not be adequate here. Inspection of the formula for least squares covariance adjustment for age shows that it makes a single adjustment for the overall mean of age; however, it cannot adjust the control girls to be older and the control boys to be younger. Similarly, the fixed adult standard for BMI does not apply in childhood, where the normal range of BMI varies with age; so, a single covariance adjustment for BMI is not appropriate for children of widely varied ages. We would like to compare treated and control children with similar BMI at the same age. In simulations, Rubin (1979) found that covariance adjustment unaided by matching can increase, rather than decrease, the bias from covariates if the model is not quite correct. Rubin suggests that covariance adjustment may serve as a supplement to, but not a substitute for, matching in the design.

Caliper: a caliper on the propensity score forbids matching of two individuals whose propensity scores are very different

Propensity score: the conditional probability of treatment given the observed covariates

Covariate distance: numerical measure indicating how similar two individuals are in terms of observed covariates **x**

Near-fine balance: occurs when the distributions are as close to fine balance as the data will allow

Fine balance: a nominal covariate has the same distribution in matched treated and control groups; does not refer to who is matched to whom

3. EXAMINING A MATCHED COMPARISON

3.1. Checking to See Whether Covariates Are Balanced in Treated and Matched Control Groups

An important aspect of matching is that a scientific audience can examine a matched design and can see that matching has successfully balanced observed covariates **x**, knowing nothing about how the matched design was constructed except the key element that it was built without access to the outcomes. Scientific experts can focus on the science, not the matching algorithm. This match has 49 adolescents treated with SSRIs, each matched to 10 untreated controls, so there are 490 controls in total. Let us first examine a matched comparison for the example in Section 2.1, then turn to its construction in Section 4. A modern matched sample is built using several mutually reinforcing technical tools, but in the words of the proverb, the proof of the pudding is in the eating, not in the stirring: The match can be assessed before or without getting into the technical details of its construction. As discussed in Section 4, the match was built using a caliper on the propensity score, minimizing the total covariate distance within the caliper, requiring an exact match for gender and a near-exact match for a missing BMI, with a near-fine balance constraint for 24 discrete categories built from the interaction of age categories × gender × cotinine categories; however, the scientific reader does not need to get into any of that to assess whether the match has produced groups comparable in terms of observed covariates.

Figure 1*a* shows that, prior to matching, the treated girls were older than the control girls, while the treated boys were younger than the control boys. **Figure 1*b*** shows that matching has corrected this, so treated and control girls have similar distributions of age, and treated and control

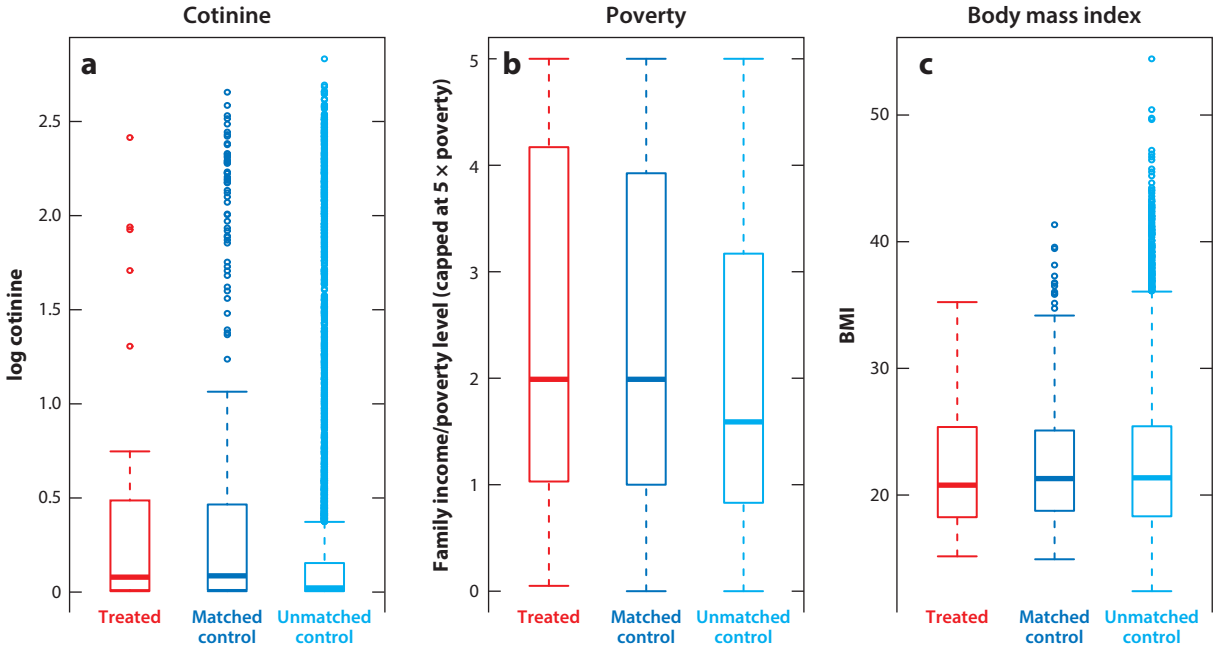

Figure 2

Cotinine, poverty, and body mass index (BMI) for 49 selective serotonin reuptake inhibitor users, their 490 = 10 × 49 matched controls, and the remaining 5,945 unmatched potential controls. Unmatched potential controls were poorer and had less cotinine in their blood, but this is corrected by matching.

boys have similar distributions of age. **Figures 1*b***, **2**, and **3**, and **Tables 1** and **2** display covariate balance, the marginal distribution of covariates in matched groups, ignoring for the moment who is matched to whom.

Figure 2 looks at three additional covariates. Each panel of **Figure 2** shows the distribution of one covariate in the treated group, the matched control group, and the potential controls not selected for the matched comparison. We hope to see similar distributions of a covariate for the treated group and for matched controls.

Cotinine in the blood (ng/mL) is a biomarker for recent exposure to tobacco. High levels of cotinine are indicative of tobacco use, such as smoking cigarettes, and medium levels are consistent with an environment in which other people are using tobacco. Because the distribution of cotinine is extremely skewed, **Figure 2** displays $\log_{10}(1 + \text{cotinine})$. Notably, the treated group has higher levels of cotinine than the unmatched controls, but this does not occur in the matched sample.

In NHANES, family income is recorded as the ratio of family income to the poverty level and is capped at five times poverty to preserve confidentiality. In **Figure 2**, family incomes for SSRI users are somewhat higher than for the unmatched controls, but matching has corrected this.

In adults, BMI is a common measure of obesity, with values above 25 called overweight and values above 30 called obese. In children, these norms for BMI no longer apply, and the norms for BMI vary with age. In **Figure 2**, the distribution of BMI is similar for the three groups.

The propensity score is the single covariate defined as the conditional probability of treatment given observed covariates **x** (see Section 6). The propensity score can be estimated from the observed data without using outcome information, perhaps using a model, such as a logit model, to predict treatment from observed covariates **x**. Such a model was fit using age, the poverty measure

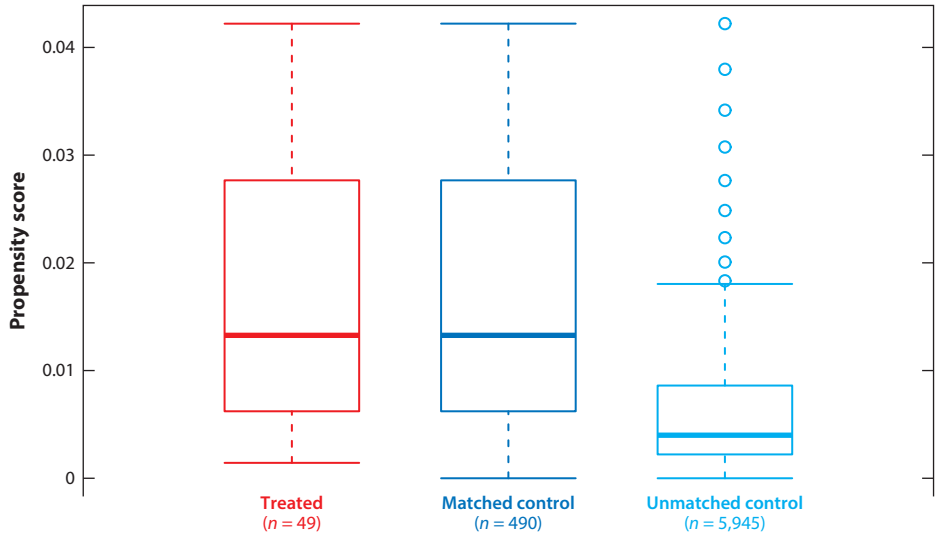

Figure 3

Estimated propensity score for 49 selective serotonin reuptake inhibitor users, their 490 = 10 × 49 matched controls, and the remaining 5,945 unmatched potential controls. The propensity score was estimated from three binary indicators for the following: female, black, and Hispanic; age; and poverty ratio and an indicator for a missing poverty ratio. The unmatched controls are quite different from the rest in terms of their propensity scores.

in **Figure 2**, and four binary indicators for female, black, Hispanic, and a missing value of income. **Figure 1*a*** shows this model is not quite correct—there is at least an interaction between age and gender—but possible misspecification of the propensity score is addressed by several other matching methods in Section 4. In general, the common, recommended joint use of several matching methods, together with checking covariate balance prior to accepting a matched design, means that a matched design is not dependent on the success of any single method. The use of multiple matching methods plus diagnostic checks makes a matched sample multiply robust—robust to failures of individual methods.

Figure 3 depicts the distribution of estimated propensity scores. The values vary considerably but rarely exceed 4%. Notably, the distribution of the propensity score for unmatched controls exhibits limited overlap with the distribution in the treated group, but the matched controls are similar to the treated group. Taking account of several covariates at once in the propensity score, the unmatched controls are seen to be very different from the treated group.

Table 1 provides means and percentages for several observed covariates. When a covariate has missing values, **Table 1** describes both the observed values and the percentages of missing values. Compared with unmatched controls, in addition to the patterns seen in **Figures 1–3**, the treated group has more females, fewer blacks, and fewer Hispanics.

In **Table 1**, the pattern of missing data is also quite different. The pattern of missing data is observed, but the missing values themselves are not observed. Missing family incomes and cotinine values are much more common among unmatched potential controls than among treated children. The pattern of missing BMIs was also different, but this observation requires extended discussion.

Missing BMIs were extremely rare: There were only 13 missing BMIs among all 6,484 children, or 13/6,484 = 0.2%; however, 2 of the 13 missing BMIs were among the 49 treated children, or 2/49 = 4.0%, an enormous, 20-fold, relative difference. Because the 6,435 potential controls

Table 1 Covariate means or percentages for 49 selective serotonin reuptake inhibitor users, 490 = 49 × 10 matched controls, and 5,945 unmatched controls

	Treated	Control	Unmatched
Female %	67.3	67.3	46.3
Age (mean)	14.9	14.7	13.5
Black %	14.3	15.1	28.1
Hispanic %	14.3	14.9	39.9
BMI (mean)	22.3	22.4	22.5
BMI missing %	4.1	2.2	0.0
Family income/poverty (mean)	2.4	2.4	2.1
Family income/poverty missing %	0.0	0.6	7.0
$\log_{10}$ (1 + cotinine)	0.4	0.4	0.3
Cotinine missing	2.0	2.7	11.1
Propensity score (mean, as a %)	1.7	1.7	0.7

Abbreviation: BMI, body mass index.

contain only 13 − 2 = 11 children with missing BMIs, it is not possible to match 10-to-1 and yet balance the indicator of missing BMI. Among adolescents, depression is sometimes associated with eating disorders, leading us to wonder about these missing BMIs. Is the pattern of missing BMIs a reason to worry about the study's conclusions? The matching method in Section 4 uses a technique called near-exact matching to force all 11 potential controls with missing BMIs to be matched to the two treated children with missing BMIs, although 9 other potential controls were matched to these two treated children to maintain the 10-to-1 matching ratio. As will be seen in a moment, this method of matching for missing BMIs will aid us in thinking about whether to be worried about the very different pattern of missing BMIs in treated and control groups. We hope that the missing BMIs are a minor matter, but it would be nice to see something in the observable data that clinches the matter, perhaps in a graph.

Near-exact matching: maximizes the number of pairs that are exactly matched for a covariate but tolerates inexact pairs when they cannot be avoided

Mostly, we have been checking covariate balance one covariate at a time. In contrast, **Figures 1** and **3** each look at more than one covariate, and these figures suggest that care is needed to balance the joint distributions of several covariates. **Table 2** continues the examination of the joint distribution of several covariates in the matched sample, specifically age, gender, and cotinine, including the pattern of missing cotinine values. Consider the upper left corner of

Table 2 Counts illustrating near-fine balance for the interaction of age, gender, and cotinine

		Female				Male			
		Cotinine, ng/mL				Cotinine, ng/mL			
Age	Group	<2	2–50	≥50	Missing	<2	2–50	≥50	Missing
8–11	Treated	2	1	0	0	4	1	0	0
	Control	20	10	0	0	40	9	0	2
12–16	Treated	10	5	0	0	6	1	1	0
	Control	100	49	0	1	60	10	10	0
17–20	Treated	12	0	3	0	2	0	0	1
	Control	120	0	30	0	18	0	1	10

The total counts are 49 treated children and 490 = 10 × 49 matched controls. In each cell, the control count would equal ten times the treated count if fine balance were feasible, but there are small, unavoidable deviations from the desired 10-to-1 ratio. Near-fine balance is as close as possible to the desired 10-to-1 ratio.

Table 2 for female children aged 8 to 11 with cotinine value <2 ng/mL. There are 2 such treated children and 20 such controls, just what we wanted in a 10-to-1 match. In contrast, for female children aged 12 to 16 with cotinine values between 2 and 50, the ratio is slightly off—5 treated and 49 controls, rather than the desired 5 treated and 50 controls. **Table 2** would exhibit a pattern known as fine balance if every cell had the desired 10-to-1 ratio. Fine balance does not make reference to who is matched to whom, just to the frequencies in the treated and matched control groups as a whole. **Table 2** is close to fine balance, but there are a few small deviations from fine balance. In fact, **Table 2** exhibits what is known as near-fine balance, meaning that it is as close to fine balance as the data will permit; that is, the total absolute deviation from fine balance has been minimized. Actually, the deviation from fine balance was minimized subject to a few constraints; for instance, precedence was given to the requirement that missing BMIs be matched exactly. Although there are small deviations from fine balance in **Table 2**, the balance is better than we expect from complete randomization. Pimentel et al. (2015a) develop and illustrate a general method comparing covariate balance in a matched sample to covariate balance in a completely randomized experiment built from the same data. Essentially, that method compares a table like **Table 2** to 10,000 analogous tables from a completely randomized experiment with the same marginal totals.

The match we have been examining seems satisfactory in terms of covariate balance, as seen in **Figures 1–3** and **Tables 1** and **2**, so we accept this design and turn to the next step of examining outcomes. The construction of this matched design went through several iterations, as discussed in Section 4, each improvement removing a problem evident from the previous iteration in figures or tables similar to **Figures 1–3** and **Tables 1** and **2**. All iterations were conducted without examining outcomes, and once we accept a design and examine outcomes we cannot go back and revise the design.

3.2. A Primary Analysis of Outcomes in Matched Groups

Figure 4 depicts the primary analysis of total femur bone mineral density in treated and control groups. Bone density is somewhat lower in the treated group. To the eye, the marginal distributions look shifted with similar dispersion and shape. Bone density tends to be lower among children receiving SSRIs than among matched controls.

The treated boxplot in **Figure 4*a*** describes 49 children, while the control boxplot describes 490 children. Is the control group more prone to extreme bone densities? Certainly, the most extreme bone densities in **Figure 4*a*** are all in the control boxplot, but that is not a good guide when one boxplot describes ten times as many children as the other. We expect the maximum and minimum of 490 children to be more extreme than the maximum and minimum of 49 children, even if the two groups were drawn from the same population with no effect of SSRIs. **Figure 4*b*** does not suffer from this limitation: It is a quantile-quantile plot of the distributions of bone densities in the two groups, and it takes account of the differing sample sizes. Although one cannot be certain with 49 treated children, **Figure 4*b*** does not provide a strong indication of more extreme bone densities in the control group. Because all but one of the points in **Figure 4*b*** fall below the line of equality, the distribution of bone densities looks stochastically smaller in the treated group.

An M-test is a robust test based on the quantity equated to zero in the definition of Huber's (1981) M-estimates. Maritz (1979) developed randomization inference based on M-tests in matched pairs, and there is a straightforward extension to matched sets with multiple controls (Rosenbaum 2007).

The difference in **Figure 4** would not easily be attributed to chance if it were to occur in a 10-to-1 matched randomized experiment. Using the randomization distribution of an M-test, the one-sided p-value testing no effect of SSRIs is 0.00020, and the two-sided p-value is twice that.

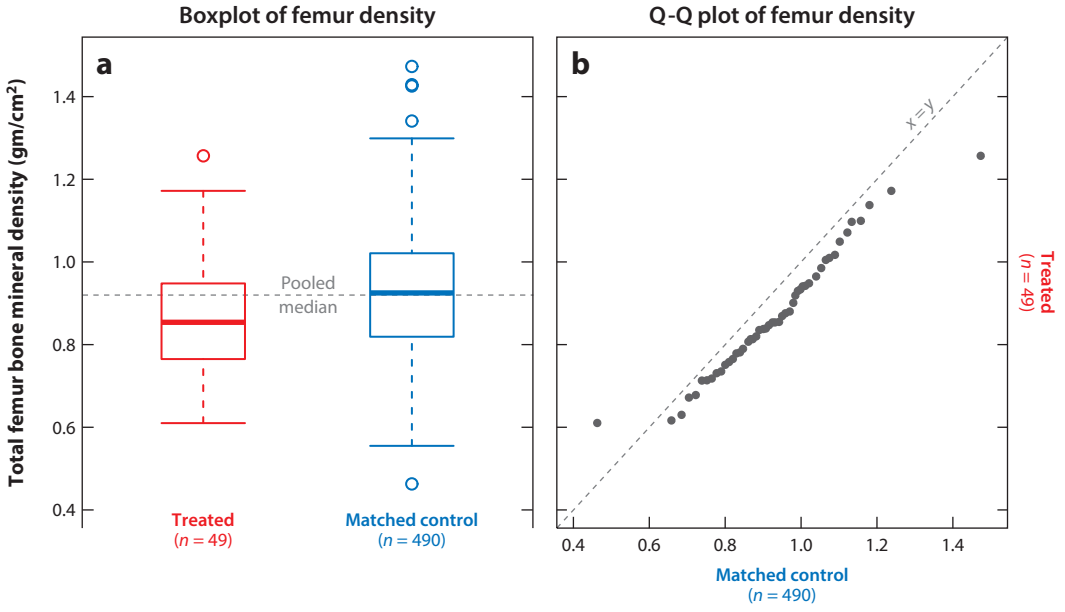

Figure 4

Total femur bone mineral density for 49 selective serotonin reuptake inhibitor users and their 490 = 10 × 49 matched controls. (*a*) Boxplot of femur density. (*b*) Q-Q plot of femur density. Abbreviation: Q-Q, quantile-quantile.

So, **Figure 4** would have been convincing evidence that SSRIs reduce bone density had **Figure 4** been produced by a randomized experiment. **Figure 4** is not, however, from a randomized experiment. Is it useful to have a very small *p*-value computed from an assumption we know to be false? Could small departures from a randomized experiment produce a much larger *p*-value? How far would treatment assignment in **Figure 4** need to depart from a randomized experiment to render plausible the null hypothesis of no effect of SSRIs on bone density? **Figure 4** is matched for several observed covariates, but it is easy to think of covariates that were not controlled by matching. What would such an unobserved covariate have to be like if failure to match for it is to explain away, as noncausal, the difference seen in **Figure 4**?

Figure 4 is not from a randomized experiment, but small departures from randomization could not easily explain the pattern in **Figure 4**: An unobserved covariate would have to increase the odds of a lower bone density by a factor of five and increase the odds of treatment with SSRIs by a factor of three to produce a one-sided *p*-value of 0.050. Despite the small sample size in the treated group, the comparison in **Figure 4** is not sensitive to small biases from unmeasured covariates. Of course, a sufficiently large bias can explain away, as noncausal, any association in any observational study—after all, association, no matter how strong, does not logically entail causation. Indeed, though insensitive to small biases in treatment assignment, the comparison in **Figure 4** is far more sensitive to unmeasured bias than, say, the studies of smoking as a cause of lung cancer (Rosenbaum 2002, section 4.3.2).

Sensitivity analysis: determines how much bias from an unmeasured covariate would need to be present to change a study's conclusion

The calculations in the previous paragraph are a sensitivity analysis for an M-test of no treatment effect (Rosenbaum 2007, Rosenbaum & Silber 2009). Briefly, in a matched randomized experiment, the 11 children in a 10-to-1 matched set would each have probability 1/11 of being randomly assigned to treatment rather than to control. The sensitivity analysis allows that probability to depart form 1/11, with the magnitude of departure controlled by a sensitivity

Stability analysis: determines whether a minor change in the analysis could change a study's conclusion

parameter, Γ, but with the pattern of departure left unspecified, left to do its worst. Specifically, before matching, two subjects with the same observed covariates might differ in their odds of treatment by at most a factor of $\Gamma \geq 1$. For $\Gamma = 1$, this produces the randomization inference, the *p*-value of 0.00020 mentioned above. For $\Gamma = 2$, you and I might look the same in terms of observed covariates, but because we are not the same in other unobserved ways, you might be twice as likely as I to receive the treatment. For $\Gamma > 1$, there is no longer a single *p*-value; rather, there is an interval of possible *p*-values depending upon the specific unknown pattern of biases of magnitude at most Γ. If we have rejected the hypothesis of no effect in a randomization test, as we did with *p*-value 0.00020 at $\Gamma = 1$, then it is natural to focus on the upper endpoint of the interval of *p*-values for $\Gamma > 1$. By doing this, we ask: What magnitude of bias, Γ, would need to be present to produce a *p*-value that would accept the hypothesis of no effect? In **Figure 4**, it turns out that the maximum possible *p*-value testing no effect just equals the conventional 0.050 level at $\Gamma = 2$, a moderately large but not enormous bias. For instance, one of the studies of heavy smoking and lung cancer becomes sensitive to bias at $\Gamma = 6$ rather than $\Gamma = 2$, so that study is insensitive to much larger unmeasured biases (Rosenbaum 2002, section 4.3.2).

There are various aids to interpreting the parameter Γ; see, for instance, Rosenbaum (2017b, table 9.1). In particular, the single parameter Γ may be interpreted or amplified in terms of two parameters, where (*a*) Λ limits the association between an unobserved covariate and the treatment, here SSRIs; (*b*) Δ limits the association between the same unobserved covariate and the outcome, here bone density; and (*c*) $\Gamma = (\Lambda\Delta + 1)/(\Lambda + \Delta)$ (see Rosenbaum & Silber 2009). The claim in the previous paragraph that an unobserved covariate would have to increase the odds of a lower bone density by a factor of five and increase the odds of treatment with SSRIs by a factor of three to produce a one-sided *p*-value of 0.050 is deduced from $\Gamma = 2 = (5 \times 3 + 1)/(5 + 3) = (\Lambda\Delta + 1)/(\Lambda + \Delta)$. [The reported calculations were produced by the `senm` and `amplify` functions in the `sensitivitymult` package in R with default settings, where the *p*-value of 0.050 is produced at $\Gamma = 2$ and amplifies to $(\Lambda, \Delta) = (3, 5)$. Additionally, `senmCI` produces sensitivity analyses for point estimates and confidence intervals.]

3.3. Secondary and Exploratory Analyses

Figure 5 returns to the issue that missing BMIs, though very rare, were much more common among SSRI users. We hope that this is a minor issue, and hope to put a minor issue to rest. In **Figure 5**, each matched set produces one treated-minus-control difference, and these differences are plotted. Each difference is the value for the one treated child in a matched set minus a typical value for the ten controls in the set. The typical value for the ten controls is Huber's M-estimate, as implemented in the `huber` function in the `MASS` package in R. Structured in this way, we may compare two boxplots, one with all 49 differences, the other with the 47 differences for 47 matched sets with complete data on BMI. **Figure 5** shows these two boxplots, plus the two omitted differences as asterisks. The two boxplots look similar, and the omitted points are close to the center of both boxplots, so it is difficult to imagine any way that the two missing BMIs seriously distort inferences that use robust methods like M-statistics. It is convenient that matching placed all of the missing BMIs in two matched sets, so boxplots with and without missing BMIs could easily be compared.

Figure 5 is a stability analysis. A stability analysis varies a minor analytical decision to check that it produces only minor changes occur in the conclusions. No explicit statistical assumption is involved in a stability analysis. In contrast, a sensitivity analysis is an explicit mathematical calculation: It varies an assumption underlying a statistical procedure to determine what magnitude of departure from that assumption would be needed to alter the conclusion. The

Figure 5

Treated-minus-control differences in femur bone mineral density for all 49 matched sets and for the 47 matched sets with complete BMI data. The two asterisks at the right show the two differences for the two matched sets with some missing BMI data. The gray dashed horizontal line is at zero difference. Abbreviation: BMI, body mass index.

discussion in Section 3.2 of **Figure 4** included a sensitivity analysis that relaxed the naive assumption that treatments were assigned at random within matched sets. As noted in Section 3.2, every observational study is sensitive to sufficiently large biases from nonrandom treatment assignment, so the sensitivity analysis answers "how much," not "whether": How much bias would need to be present to change the conclusions? Section 3.2 concluded that no small bias, no matter what its form, could alter the conclusion, and it defined "small" precisely. People who are confused about the role of assumptions in statistical inference exhibit a parallel confusion about the distinction between a stability analysis and a sensitivity analysis, often mislabeling one as the other. If you do not understand an assumption, then you cannot relax it.

The 10-to-1 match was exact for gender: Boys were matched to boys, girls to girls. There were 16 matched sets containing boys and 33 sets containing girls, with $I = 49 = 16 + 33$ sets in total. Gender would be an effect modifier if the treatment effect were different for boys than for girls. In general, an effect modifier is an interaction between a covariate and a treatment. **Figure 6** compares the sets of boys and the sets of girls. To the eye, the effect of SSRIs looks larger for boys than for girls, but do remember that the boys are typically younger, and there are only 16 treated boys. As seen in **Figure 1**, the available data do not sharply distinguish age and gender: We know little about the effect of SSRIs on younger girls or older boys because we have very few of them. We may ask two questions about effect modification: (*a*) Are we confident that it exists? (*b*) Whether or not we are confident that it exists, are we confident that it alters the degree of sensitivity to unmeasured biases? Despite the visual impression, in **Figure 6** we cannot be confident that there is effect modification: Applying Wilcoxon's two-sample test to compare the 16 differences for boys and the 33 differences for girls, we obtain a two-sided p-value of 0.12.

Effect modification: present if the magnitude of the treatment effect varies with the level of an observed covariate

When there is effect modification, the treatment effect is larger in certain subgroups, and generally when sample sizes are very large, larger effects are insensitive to larger biases (Rosenbaum

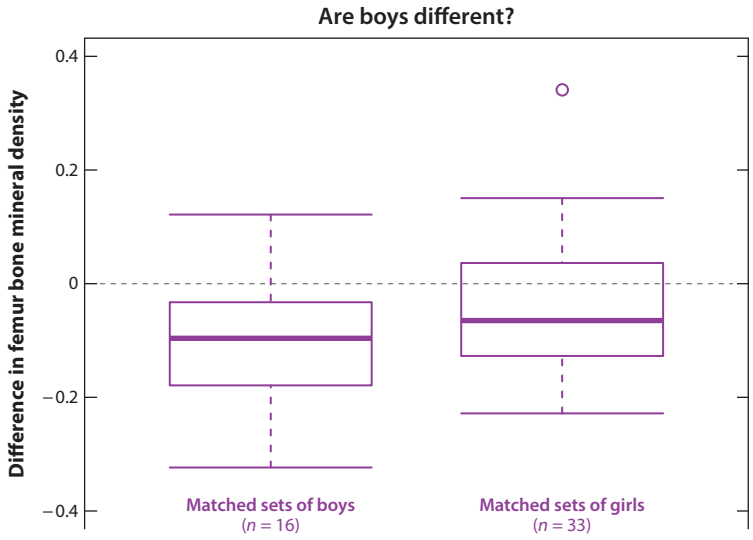

Figure 6

Treated-minus-control differences in femur bone mineral density for 16 matched sets of boys and 33 matched sets of girls. The gray dashed horizontal line is at zero difference. Recall that the boys are younger.

2010, part 3). Here, the sample sizes are not large. In **Figure 6**, the method of Lee et al. (2018) does three sensitivity analyses, the analysis for all $I = 49$ sets from Section 3.2, a parallel analysis confined to 16 sets of boys, and a parallel analysis confined to 33 sets of girls, correcting for performing three correlated tests. That analysis finds that the results for the 16 sets of boys are insensitive to larger biases than the analysis in Section 3.2 of all $I = 49$ sets. Specifically, to explain away the ostensible effect for boys, an unobserved covariate would have to increase the odds of a lower bone density by more than a factor of eleven and increase the odds of treatment with SSRIs by a factor of three to produce a one-sided p-value of 0.050. [These calculations use the `submax` and `amplify` functions in the `submax` package in R with default settings, where the p-value of 0.050 is produced at $\Gamma = 2.45$ and amplifies to $(\Lambda, \Delta) = (3, 11.5)$.]

3.4. Summary of the Examination of a Matched Comparison

Transparency means making evidence evident. It means having warranted confidence about some topics, warranted concern about others, and no confusion between topics that warrant confidence and topics that warrant concern. It means accurately informing the critical debate that follows an observational study, not garbling it. My hope is that the matched comparison in **Figures 1–6** and **Tables 1** and **2** strikes you as transparent in this sense. Transparency is a key goal in matched comparisons.

We saw the following: (*a*) The matched groups look comparable in terms of measured covariates, including certain aspects of the joint distributions of several covariates at once in **Figures 1** and **3** and **Table 2**. (*b*) Total femur bone density looks lower in SSRI users, and this difference is not readily explained by small biases from nonrandom treatment assignment, though, as always in any observational study, it could be explained by sufficiently large biases. Bias from nonrandom treatment assignment starts to become an alternative explanation at about $\Gamma = 2$, far from a trivially small bias, but much smaller than the bias, $\Gamma = 6$, needed to explain away the association between heavy smoking and lung cancer. (*c*) The issue of missing BMIs looks to be a minor,

negligible matter in the stability analysis in **Figure 5**. (*d*) The insensitivity to unmeasured biases is greater for boys, keeping in mind that the boys were younger and we have little data about younger girls receiving SSRIs.

3.5. How Many Controls?

In Section 2.1, there are $I = 49$ treated children. One source of uncertainty in Section 2.1 is that pooling three NHANES surveys yields only 49 treated children. But there are plenty of potential controls. How many controls should be matched to each treated child? The example matched ten controls to each treated child. Is that too many? It is too few? How is this issue decided?

Consider a matched study in which each of I treated individuals is matched to $\kappa \geq 1$ controls, making I treated individuals and κI controls. In Section 2.1, matched pairs means $\kappa = 1$ with $I = 49$ treated children and $I = 49$ controls. Matching with $\kappa = 2$ means 49 matched sets, each with one treated child and two controls, making $I = 49$ treated children and $\kappa I = 2 \times 49 = 98$ controls, and so on. The spectrum of possibilities extends from pairs, $\kappa = 1$, to having infinitely many controls for each treated individual, far more controls than are available in NHANES. How does the stability of an estimated treatment effect change as we move along this spectrum? Stability here refers to sampling variability, the uncertainty from a limited sample size, which we measure by the variance of the estimator. In observational studies, sampling variably is only one source of uncertainty; it is typically not the major source of uncertainty, but it is the most manageable source, and so we should make a sensible choice about managing it before trying to address the major sources of uncertainty, namely biases from observed and unobserved covariates.

Under a simple, familiar, conventional Gaussian model for matched sets, the variance of the estimator is proportional to $1 + \frac{1}{\kappa}$, where the omitted constant of proportionality does not depend on κ, but depends on all sorts of other things: the sample size I, the variance of errors, and so on (see, for instance, Rosenbaum 2010, section 8.7). Holding all those other things fixed for the moment, varying just the number κ of controls, we see the effects of κ completely captured by the simple formula $1 + \frac{1}{\kappa}$.

For matched pairs, $1 + \frac{1}{\kappa} = 1 + 1 = 2$, but with infinitely many controls, $\kappa = \infty$, the constant becomes $1 + \frac{1}{\kappa} = 1 + 0 = 1$. Adding controls, $\kappa \to \infty$, drives out sampling uncertainty among controls but leaves the treated group untouched, so the instability of our estimator drops in half, but it does not drop to zero. Even if we had infinitely many controls in Section 2.1, we still have only $I = 49$ treated children, and nothing we do with the controls—no model or machine learning algorithm—is going to change that. If we match with $\kappa = 2$ controls per treated, then $1 + \frac{1}{\kappa} = 1 + \frac{1}{2} = 1.5$, and we have traveled half the distance from 2 for pairs to 1 for infinitely many controls. With $\kappa = 4$ controls, $1 + \frac{1}{\kappa} = 1 + \frac{1}{4} = 1.25$, so we have cut the distance to infinitely many controls in half yet again. It is not that $\kappa = 4$ controls is wonderful—after all, we still have only $I = 49$ treated children in Section 2.1—it simply that driving up the number of controls, $\kappa \to \infty$, rapidly becomes less relevant to the study's real problems and uncertainties once $\kappa \geq 4$. Economists use the phrase "diminishing returns," and the formula $1 + \frac{1}{\kappa}$ exhibits diminishing returns with a vengeance: The return to increasing κ from 1 to 2 equals the return from increasing κ from 2 to ∞. With $\kappa = 10$ controls, $1 + \frac{1}{\kappa} = 1 + \frac{1}{10} = 1.1$, so almost all of the sampling variability comes from the I treated subjects, not from the κI controls.

The model in this section is a consequential oversimplification in the following sense. The model assumes, falsely, that the quality of matched sets can be held fixed as κ increases, but that is far from true. If we pick the closest $I = 49$ controls for matched pairs, those pairs will be much closer on covariates than if we take the nearest $4 \times 49 = 296$ or $10 \times 49 = 490$ controls. Worst of all, if we irresponsibly use all of $C = 6{,}435$ controls in NHANES, then the bone density of girls

aged 8 will play an important role in judging the effects of SSRIs on girls whose typical age is 16, even though not one girl aged 8 received SSRIs. In Section 2.1, we could match each of $I = 49$ treated children to $\kappa = 131 \approx 131.32 = 6{,}435/49 = C/I$ controls, but the match quality would be terrible.

Distance matrix: a table indicating the covariate distance between each treated individual and each potential control

The example used $\kappa = 10$ controls per treated child, mostly because $\kappa = 10$ yielded matches of good quality, but this is only slightly better in terms of sampling variability than $\kappa = 5$ controls, $1 + \frac{1}{5} = 1.2 > 1.1 = 1 + \frac{1}{10}$. Even the small gain, from 1.2 for $\kappa = 5$ to 1.1 for $\kappa = 10$, is larger than the gain from 1.1 for $\kappa = 10$ to 1.008 for $\kappa = 131$ using all controls. If the quality of the 1-to-10 match had been poor, there would have been little loss in using a 1-to-5 match instead.

Several matching structures enhance bias reduction by matching with a variable number of controls rather than a fixed number, κ (see Section 5.1). One treated individual has one control, another has four controls, depending upon the number of similar controls available. However, using variable numbers of controls rather than a fixed number makes sampling variability worse by requiring weights in analysis. The best case for sampling variability occurs when matching with a fixed number κ of controls, and, even in this most optimistic case, the gains are small once $\kappa \geq 5$.

4. CONSTRUCTING A MATCHED COMPARISON

4.1. Matching as an Optimization Problem

Modern methods find a match by solving an optimization problem subject to various constraints. These optimization problems can be described and solved in various ways, but I will describe them in the simplest way until Section 5.5. The simplest description starts with a table or matrix with I rows, one for each of I treated individuals, and C columns, one for each potential control. In Section 2.1, $I = 49$ and $C = 6{,}435$, making a fairly small table of size $49 \times 6{,}435$.

Table 3 shows a portion of a $49 \times 6{,}435$ distance matrix, specifically the first four rows and ten columns, for the first four treated children and the first ten potential controls. The NHANES identifiers for these $14 = 4 + 10$ children are given in the Appendix. A distance in **Table 3** is small if two children look similar in terms of observed covariates. For example, the distance between treated child 4 and control 5 is the smallest in this table, with value 2.0. These two children are both girls, neither is black or Hispanic, their ages are 12 and 11, and they have BMIs of 21 and 18, but their cotinine values are quite different. For comparison, treated child 4 is much further from control 6, with a distance of 25.1: Unlike treated child 4, who has just been described, control 6 is a Hispanic male aged 16.

Obviously, in a much larger table, the smallest distances are much smaller: In the $49 \times 6{,}435$ distance matrix, the smallest distance is not 2.0 but 0.03. For treated child 1, in the $49 \times 6{,}435$

Table 3 Robust Mahalanobis distances between the first 4 treated children and the first 10 potential controls

Treated	Potential control									
	1	2	3	4	5	6	7	8	9	10
1	11.3	24.5	25.8	13.6	18.7	29.6	17.3	22.4	5.7	13.4
2	5.7	14.1	2.7	13.4	9.7	16.7	9.0	15.8	7.1	9.8
3	11.3	7.6	22.3	15.3	17.7	17.9	17.9	17.7	20.7	15.7
4	11.3	23.4	18.2	17.8	2.0	25.1	3.4	6.0	9.3	12.9

A small distance means that a treated child resembles a control in terms of observed covariates.

distance matrix, the closest control child is at distance 0.073, not distance 5.69 in **Table 3**. For the first four treated children, the best pair match in the 4 × 6,435 distance matrix is vastly better than the best pair match in **Table 3**. **Table 3** is simply a tolerably small illustration of matching concepts. The computation of the distances in **Table 3** is discussed in Section 4.5.

Optimal matching is a well-solved problem, but not a trivial one (Bertsekas 1981); that is, large problems can be quickly solved using existing but nontrivial algorithms. **Table 3** illustrates why, though small, the problem is not trivial. Suppose that we wanted to match each of the 4 treated children to two controls so that no control is used twice. We cannot go through the table row by row, picking the two closest controls for each treated child. If we did this, treated child 1 would want controls 1 and 9, while treated child 2 would want controls 1 and 3, so there is a conflict about which treated child gets control 1. Obviously, we could develop an elaborate statistical technique that uses control 1 twice and corrects the analysis for double use of control 1, but that is inefficient because control 1 does not become two controls by virtue of being used twice; Rosenbaum (2017a, section 1.2) provides a small numerical illustration of the inefficiency of so-called matching with replacement. In **Table 3**, control 10 is almost as good as control 1 for treated child 1, and vastly better close swaps are available in the 49 × 6,435 distance matrix.

The 10-to-1 match from the 49 × 6,435 distance matrix makes 49 matched sets, each with ten distances between a treated child and a control, or 490 = 49 × 10 distances in total. The optimal match will minimize the total of these 490 distances over all possible 10-to-1 matches built from the 49 × 6,435 distance matrix (Rosenbaum 1989). Hansen's (2007) `pairmatch` function in his `optmatch` package in R may be used to find this optimal match using the `RELAX IV` Fortran code of Bertsekas & Tseng (1988). This optimal matching problem is not solved by a greedy or nearest-available match that picks the smallest distance in the 49 × 6,435 distance matrix, removes that column, picks the second smallest distance in the reduced matrix, and so on. The total distance for greedy matching can be much worse than for optimal matching: The ratio of the total distances may be arbitrarily large.

The optimal matching or optimal assignment problem is a standard combinatorial optimization problem for which several quick solutions are known (Korte & Vygen 2012, chapter 11). Details aside, perhaps the most intuitive solution is the auction algorithm of Bertsekas (1981). The problem is that two or more treated individuals may want the same control. How is competition for the same control to be decided? Bertsekas literally holds an auction, selling controls to the highest bidder, with prices that adjust as competition emerges for the same control. As prices rise, a treated subject may settle for the second closest control at a much lower price than the closest control. In the auction algorithm, some key concepts in optimization, such as duality and complementary slackness, acquire a familiar economic form. Bertsekas (1990, 2001) provides an attractive, informal explanation of the auction algorithm.

Optimal matching: finds the closest pairing of individuals subject to certain requirements or constraints on covariate balance

`RELAX IV`: Fortran code solving the minimum cost flow problem, available in R using the function callrelax

Auction algorithm: sells controls to treated individuals in an auction with variable prices

Exact matching: requiring identical values of an observed covariate for matched individuals, e.g., boys matched to boys, girls matched to girls

4.2. Forbidding Certain Matches: Exact Matching and Calipers

We often wish to avoid matching certain controls to certain treated children. The match in Section 3 imposed two requirements of this kind. First, it required treated and control children to have an absolute difference in their propensity scores of at most 0.02. A requirement of this kind is called a caliper (Cochran & Rubin 1973). As seen in **Figure 3**, a caliper of 0.02 is not a particularly tight caliper, but it eliminates the worst matches on the propensity score. In **Table 3**, we replace a distance by ∞ if individuals differ on the propensity score by more than 0.02. Also, the match in Section 3 required boys to be matched to boys, girls to girls, so in **Table 3** we replace a distance in row i and column j by ∞ if treated child i and control j have different genders. The resulting distance matrix is **Table 4**.

Table 4 Distance matrix with ∞ for mismatches for gender and violations of the 0.02 caliper on the propensity score

Treated	Potential control									
	1	2	3	4	5	6	7	8	9	10
1	∞	∞	∞	∞	∞	∞	∞	∞	5.7	∞
2	5.7	∞	2.7	13.4	9.7	∞	9.0	15.8	7.1	9.8
3	∞	7.6	∞	∞	∞	17.9	∞	∞	∞	∞
4	11.3	∞	18.2	17.8	2.0	∞	3.4	6.0	9.3	12.9

If we cannot match while avoiding the ∞s, then matching is said to be infeasible. In **Table 4**, matching in pairs, 1-to-1, is feasible, but matching 2-to-1 is infeasible because treated child 1 has only one potential control, namely control 9. Using the 49 × 6,435 distance matrix, matching 10-to-1 was feasible in Section 3, meaning that boys were matched to boys, girls to girls, and the caliper on the propensity score was never violated.

A distance matrix may contain many ∞s, and when this is so it is wise to determine the location of the ∞s first, computing the distances only in positions without ∞s.

Exact matching for gender is helpful in that it simplifies looking for effect modification by gender in **Figure 6**. Exact matching is not essential, however. The method of Lee et al. (2018), illustrated in Section 3.3, does not require exact matching: It uses all individuals who happen to be exactly matched. For instance, it might use five controls in one matched set and eight in another.

4.3. Near-Exact Matching

In Section 3, we wanted to match exactly for the indicator of a missing BMI, but there were too few controls to do this. In the terminology of Section 4.2, 10-to-1 exact matching for this indicator is infeasible.

Near-exact matching entails matching exactly as often as is possible but tolerating a mismatch when one is unavoidable (Rosenbaum 2010, chapter 9). In **Table 1**, all of the controls with missing BMIs are included in the matched control group, although that was only 11 controls, not the desired 20 = 2 × 10 controls. Moreover, every one of these 11 controls was matched to one of the two treated children with a missing BMI, permitting us to construct **Figure 5** in which two matched sets with some missing BMIs were viewed separately.

In near-exact matching, we do not alter the infinite distances in **Table 4**, but if a finite distance corresponds to mismatch for missing BMI, then we add to that finite distance a large penalty. In Section 3, the penalty was $p = 100{,}000$ (see **Table 5**, where treated child 4 has a missing BMI, but none of these 10 control children has a missing BMI). In the actual 49 × 6,435 distance matrix,

Table 5 Treated × control distance matrix penalized for near-exact matching for the indicator of missing BMI

	Potential control									
	1	2	3	4	5	6	7	8	9	10
1	∞	∞	∞	∞	∞	∞	∞	∞	5.7	∞
2	5.7	∞	2.7	13.4	9.7	∞	9.0	15.8	7.1	9.8
3	∞	7.6	∞	∞	∞	17.9	∞	∞	∞	∞
4	11.3+p	∞	18.2+p	17.8+p	2.0+p	∞	3.4+p	6.0+p	9.3+p	12.9+p

Treated child 4 had a missing BMI, but none of these ten controls had a missing BMI, so the penalty is imposed throughout row 4. In the example, the penalty is $p = 100{,}000$. Abbreviation: BMI, body mass index.

there are only 11 of 6,435 columns in which treated child 4 has a distance that is neither ∞ nor penalized by the addition of $p = 100{,}000$.

For large enough p, a minimum distance match will avoid the penalized distances whenever possible, but in Section 3 it will be forced to accept 9 penalized distances. Notice that, because $p = 100{,}000$ is added to the original distances, if forced to incur the penalty p, the minimum distance match would still prefer to match treated child 4 to control 5 rather than to control 3.

A nominal covariate, such as gender, has two or more unordered categories. Near-exact matching can be used with $M \geq 1$ nominal covariates by adding a large penalty of p for a mismatch on covariate 1, another penalty of p for a mismatch on covariate 2, . . . , yet another penalty of p for a mismatch on covariate M. For large enough p, this will minimize the total number of mismatches on the M nominal covariates, and among designs that do this, it will minimize the total of the within-pair covariate distances. This tactic can be useful when studying effect modification, as in Section 3.3, Hsu et al. (2015), and Lee et al. (2018, appendix).

Use exact and near-exact matching sparingly. Exact or near-exact matching may be helpful for one or a few nominal covariates, as in **Figures 5** and **6**, but it is not possible when the number, M, of nominal covariates is large (see Rosenbaum 2017b, table 5.6). Consider using fine balance or refined balance instead (see Section 4.4). When an important nominal covariate has many levels, near-exact matching may usefully be combined with fine balance (Zubizarreta et al. 2011).

For large enough p, penalized distances give total priority to the covariates chosen for near-exact matching, so other covariates may be poorly matched. Small penalties—not $p = 100{,}000$, but perhaps $p = 1$ or $p = 2$—are sometimes used informally but successfully to give more emphasis to a few problematic covariates. If a balance table such as **Table 1** exhibits an unacceptable imbalance for a particular nominal covariate, then a small penalty may fix this. There are many informal but practical variations on this theme of changing the distances to emphasize problematic covariates.

4.4. Fine Balance and Related Techniques

In the discussion in Section 3.1 of **Table 2**, the concepts of fine balance and near-fine balance were introduced. In **Table 2**, fine balance would mean a 10-to-1 ratio of control-to-treated counts in every one of the $24 = 3 \times 2 \times 4$ cells—that is, in every age $\times$ gender $\times$ cotinine cell. Fine balance is not feasible in Section 3.1 given the other requirements, such as exact matching for gender and the caliper on the propensity score. Instead, **Table 2** exhibits near-fine balance: Each cell is as close as possible to the desired 10-to-1 ratio; that is, more precisely, the total count of deviations is as small as possible. Fine balance and near-fine balance do not refer to who is matched to whom; rather, they describe the distribution of individuals over the $24 = 3 \times 2 \times 4$ cells in the treated and control groups. Exact matching for gender implied fine balance for gender alone. In contrast, fine balance for the six age $\times$ gender categories finds older control girls and younger control boys, as seen in **Figure 1**, but it does not imply children are exactly matched for these six categories. Nonetheless, the covariate distances try to closely match individuals for age.

Fine balance and its relatives can be implemented in several ways (Pimentel et al. 2015a, Rosenbaum 1989, Rosenbaum et al. 2007, Yang et al. 2012, Zubizarreta 2012), although some of these variations are of more interest to programmers than to scientists. The goal is a minimum distance match subject to the constraint of fine balance. The discussion that follows gives the gist of the idea, leaving practical implementation to Section 4.6. Fine balance is like sculpting: We are interested in the sculpture left behind, but the art is in what you take away. We remove controls in such a way that a finely balanced sample remains. **Table 6** imagines that we wish to finely balance "black race" in a 2-to-1 match in **Table 5**, ignoring for a moment the infeasibility of matching treated child 1 to two controls in this small table. Notice that treated children 2 and 3

Table 6 Distance matrix augmented to finely balance black race (indicated by b)

	Potential control									
	1b	**2b**	**3b**	**4b**	**5**	**6**	**7**	**8**	**9**	**10**
1	∞	∞	∞	∞	∞	∞	∞	∞	5.7	∞
2b	5.7	∞	2.7	13.4	9.7	∞	9.0	15.8	7.1	9.8
3b	∞	7.6	∞	∞	∞	17.9	∞	∞	∞	∞
4	11.3+p	∞	18.2+p	17.8+p	2.0+p	∞	3.4+p	6.0+p	9.3+p	12.9+p
5	∞	∞	∞	∞	0	0	0	0	0	0

An auxiliary row 5 is added to the distance matrix. Were it feasible, a minimum distance 2-to-1 match would pair two nonblack controls to auxiliary 5, and then this matched set would be deleted, leaving behind a finely balanced match. Note that the addition of row 5 does not require blacks to be matched with blacks; rather, it corrects the imbalance in the frequency of blacks.

are black, as are controls 1, 2, 3, and 4. In a 2-to-1 match, fine balance for black race would mean that controls 1, 2, 3, and 4 are included as controls, though they may or may not be matched to treated children 2 and 3. **Table 6** adds a fifth auxiliary row at infinite distance from all the black controls and at zero distance from the six nonblack controls. Were minimum distance matching feasible in **Table 5**, a minimum distance match in **Table 6** would match two nonblack controls to the auxiliary row. Discarding the matched set for the auxiliary row leaves behind a 2-to-1 match finely balanced for black race; moreover, it would minimize the total distance subject to the constraint that the match is finely balanced.

Table 2 viewed its $24 = 3 \times 2 \times 4$ cells as one nominal variable with 24 categories. That is reasonable when fine balance is feasible. However, if fine balance is infeasible, if as in **Table 2** we must tolerate small deviations from fine balance, then we may prefer certain deviations to others. For instance, if we must tolerate an imbalance for cotinine, we might prefer to retain balance for age. Pimentel et al. (2015a) introduce the concept of refined balance in which the 24 categories are given a hierarchical structure, say, gender first, age second, cotinine third. In refined balance, gender is balanced as closely as possible, age as closely as possible subject to the requirement that gender is maximally balanced, and cotinine is balanced as closely as possible subject to the requirement that gender and age are maximally balanced. The nice thing about refined balance is that adding many less important covariates to the bottom of the hierarchy does not degrade the balance achieved for the most important covariates at the top of the hierarchy. In an example, Pimentel et al. (2015a) balanced a hierarchically structured nominal covariate with 2.8 million levels.

Table 2 attempts to finely balance a joint distribution of three covariates with $24 = 3 \times 2 \times 4$ levels. Zubizarreta (2012) developed a method for finely balancing the marginal distributions of several individual variables without balancing their joint distributions. He later extended this idea to strength K balance, in which all of the joint distributions of K of M covariates are balanced (Hsu et al. 2015). For example, strength 2 balance of gender, age, and cotinine would balance the joint distribution of gender and age, of gender and cotinine, and of age and cotinine, but might not balance the three-way joint distribution. Strength K balance and refined balance are two strategies for implementing fine balance with a factorial array of many covariates.

4.5. Robust Mahalanobis Distances

Rosenbaum & Rubin (1985a) suggested matching to minimize the Mahalanobis distance within calipers defined by the propensity score. As discussed in Section 6, matching for one covariate, the propensity score, tends to balance all of the covariates used to build that score, but two individuals with the same propensity score may differ in important ways. Use of the Mahalanobis distance

inside propensity score calipers tries to balance covariates and also pair similar individuals. Also, as mentioned in Section 3.1, use of the Mahalanobis distance in addition to the propensity score is one of several layers of protection against the failure of a single matching technique. Perhaps mistakenly, we ignored the age-by-gender interaction in **Figure 1*a*** when building the propensity score, but both near-fine balance in Section 4.4 and the Mahalanobis distance paid attention to that interaction, and **Figure 1*b*** shows that the interaction is balanced in the matched sample.

The Mahalanobis distance has a property, affine invariance, which means that certain changes to the data do not change the distance. A convenient practical consequence occurs when a covariate has missing data, as is true of income and cotinine in **Figure 2**. If a covariate, say income, has missing data, then do two things: Replace the missing incomes by an arbitrary number, say the mean or the median income, and include a binary variable indicating missing income, so income is represented by two covariates in the distance. Then the Mahalanobis distance is unchanged by changing the arbitrary number substituted for the missing incomes, and when income is missing, the distance prefers to pair people with missing incomes. As noted previously, such a tactic tends to balance the observed pattern of missing data, as seen in **Table 1**, but of course it cannot be expected to balance the missing values themselves. The propensity scores obtained from a linear logit model also have this property of affine invariance when a covariate and its missing indicator are both included in the model (Rosenbaum & Rubin 1984, appendix).

The Mahalanobis distance is linked to the multivariate Normal distribution, and it can do some odd things with data that are not Normal. Take a second look at **Figure 2**. An outlier or long tails in one covariate can inflate the sample variance for that covariate, leading the Mahalanobis distance to pay little attention to the covariate. Binary covariates are common, but fair coin flips have much larger variances than rare binary traits, so the Mahalanobis distance pays much more attention to mismatches for rare binary covariates than for binary covariates that divide the population in half. The Mahalanobis distance is much more concerned to pair US residents who live in Wyoming, much less concerned to pair men to men and women to women. Small adjustments remove both oddities at a small price (Rosenbaum 2010, chapter 8).

The covariate distance in **Table 3** is one robust version of the Mahalanobis distance (Rosenbaum 2010, chapter 8). Before computing the Mahalanobis distance, covariates are replaced by their ranks, with average ranks for ties. Ranks eliminate concerns about outliers and long tails. Ties reduce the variance of ranks, but the covariance matrix of the ranks is rescaled so that every covariate has its untied variance, the same value for every covariate. Covariates, like rare binary covariates, do not become more important as they become rarer and hence more heavily tied. Alas, this particular robust distance is no longer affinely invariant. In **Table 3**, medians and indicators were used in the distance for missing income and cotinine.

`optmatch`: R package useful when matching with variable numbers of controls or full matching

`rcbalance`: R package useful when matching with refined balance

`bigmatch`: R package useful in large matching problems

`nbpMatching`: R package used for matching without treated and control groups

`designmatch`: R package that provides a greatly enlarged toolkit for matching

`DiPs`: R package useful when one or two stubborn, recalcitrant covariates refuse to be balanced

Full matching: can match everyone, but weights are needed to describe the population

4.6. Software in R

For matching, I recommend six software packages in R, namely `optmatch`, `rcbalance`, `bigmatch`, `nbpMatching`, `designmatch`, and `DiPs`. The first three use the auction algorithm of Bertsekas (1981) and the Fortran code of Bertsekas & Tseng (1988); they all require you to load the `optmatch` package and accept its academic license, but no special installation is required, so they install and execute as easily as other R packages. In contrast, `designmatch` runs best with a commercial solver, either gurobi or Cplex; these are available for free to academics but require a special installation.

Hansen's `optmatch` is the earliest of these packages (Hansen & Klopfer 2006). It excels when matching with a variable number of controls or with full matching. Use of `optmatch` is discussed in Hansen (2007) and Rosenbaum (2010, chapter 13). The `DOS` package in R contains data sets from Rosenbaum (2010) that may be used to learn about and compare the five matching packages.

Subset matching: uses only the part of the treated population where controls commonly occur

Pimentel's `rcbalance` package implements refined covariate balance (Pimentel et al. 2015a), as well as fine balance and near-fine balance. Use of the package is illustrated in Pimentel (2016). The companion package `rcbsubset` implements a form of subset matching.

Yu's `bigmatch` package uses various techniques to match large data sets, hundreds of thousands of people, in a single optimization (Yu et al. 2019), although it can be used with smaller data sets as well. The package uses a version of Glover's (1967) algorithm to find an optimal caliper for the propensity score, and it can apply near-fine balance on a very large scale.

The `nbpMatching` package (Lu et al. 2011) uses the algorithm and Fortran code of Derigs (1988) to implement so-called nonbipartite (i.e., not two parts) matching. Instead of matching treated individuals to controls, `nbpMatching` divides one population into pairs so that the total distance within pairs is minimized. Nonbipartite matching has a variety of applications: (*a*) in optimal matching before randomization in experiments (Greevy et al. 2004), (*b*) in strengthening an instrumental variable (Baiocchi et al. 2010, 2014), and (*c*) in testing the fit of the Gaussian linear model (Pimentel et al. 2017).

Zubizarreta's (2012) `designmatch` differs from the others in using mixed integer programming rather than network optimization techniques. The relevant network optimization techniques can be implemented to run in polynomial time, that is, rather quickly, whereas some integer programs are very difficult to solve. The `designmatch` package offers a wide variety of new matching tactics without a firm promise that they can be implemented with reasonable computational effort; yet, the package typically performs competitively without such a promise. For instance, `designmatch` can finely balance many one-dimensional marginal distributions without balancing their interactions, or it can balance their two-dimensional joint distributions without balancing higher-way joint distributions. It can balance means, variances, covariances, empirical distribution functions and other attributes of distributions. Also, `designmatch` implements cardinality matching (Zubizarreta et al. 2014a).

Yu's `DiPs` package contains a flexible, basic function, `match`, that implements many modern matching techniques in a simple form. The package offers several additional devices using directional penalties and Lagrangians that can improve the balance for a few covariates that are somewhat out-of-balance in an existing match.

The `exteriorMatch` package constructs the exterior match that is used to compare two overlapping matched control groups (Rosenbaum & Silber 2013). For sensitivity analysis in matched observational studies, consider the `senm` and `senmCI` functions in the `sensitivitymult` package (Rosenbaum 2015b); these are illustrated in a shinyapp (**https://rosenbap.shinyapps.io/learnsenShiny/**).

5. BRIEF DISCUSSION OF ADDITIONAL PRACTICAL TOPICS

The current section briefly discusses a number of additional practical topics with references to the literature. Some of these topics are reviewed in greater detail by Stuart (2010) and Rosenbaum (2010, part II; 2017b, chapter 11).

5.1. Variable Controls and Full Matching

The focus has been on pair matching or matching with a fixed number of controls, but full matching and matching with a variable number of controls have some advantages and disadvantages (Austin & Stuart 2015, Hansen & Klopfer 2006, Ming & Rosenbaum 2000, Pimentel et al. 2015b, Rosenbaum 1991a). Today in the United States, if you were matching smokers to nonsmoking controls, you would find that few people with a college degree are smokers. As a result, a smoker

with a college degree would have available many nonsmokers as potential controls, while smokers who did not complete high school would have fewer controls available. Matching with a variable number of controls might give five controls to the smoker with a college degree but only one or two controls to the smoker who did not finish high school. Yoon's entire number can guide decisions about how many controls to match to particular individuals, and this is discussed further Section 6.5.

Full matching takes this a step further, creating matched sets that contain either one treated individual and one or more controls, or else one control and one or more treated individuals. Full matching has certain optimal properties and can match everyone (Rosenbaum 1991a). Hansen's (2007) `optmatch` package implements full matching in R. Hansen (2004) presents an interesting application of full matching.

Entire number: theoretical estimate of the number of controls available per treated individual at each value of the observed covariate, $\mathbf{x}$

Multilevel matching: dynamic programming solution that matches clusters and units within clusters

Risk set matching: matching used for time-dependent covariates and treatments

5.2. Subset Matching

Treated and potential control groups sometimes exhibit too little overlap on covariates to permit matching of the entire treated group, and various forms of subset matching have been proposed. If only a subset of the treated population is matched, then the question addressed by the study has been changed (Rosenbaum & Rubin 1985b). Nonetheless, the revised question may be interesting, and it often focuses on the border region between two established treatments, perhaps a narrow zone in which both treatments are commonly used. The conclusions of a study of such a border region might shift the location of the border region, so that over a period of years, a sequence of studies might gradually eliminate one of the two established treatments by showing it remains inferior as the border shifts.

Crump et al. (2009) suggested defining the treated population to exclude extreme propensity scores (see Section 6). A population defined by an interval of values of the propensity score may be difficult to interpret—it may, for instance, lack a clinical meaning to physicians because it fails to pick out a natural group of patients—so several authors have suggested methods to trim the population in an interpretable way (Fogarty et al. 2016, Traskin & Small 2011). For instance, medical studies commonly limit the population under study to a rectangle of some dimension defined by intervals of several covariates, say age 20–65 with stage 2–3 breast cancer, and Fogarty et al. (2016) and Traskin & Small (2011) used algorithms to find such a rectangle in which matching is feasible. In contrast, Rosenbaum (2012) used an optimal matching algorithm that has the option of discarding, at a price, a treated individual who is very difficult to match. Zubizarreta et al. (2014a) proposed maximum cardinality matching that finds the largest matched sample exhibiting good covariate balance. Pimentel's R package `rcbsubset` and Zubizarreta's R package `designmatch` implement versions of subset matching.

5.3. Beyond Treatment Versus Control Comparisons

Matching can strengthen quasi-experimental devices. For instance, matching can be used with multiple control groups (Daniel et al. 2008, Karmakar et al. 2019, Lu & Rosenbaum 2004, Pimentel et al. 2016, Rosenbaum & Silber 2013, Stuart & Rubin 2008). It can also be used to ensure local estimation of treatment effects in discontinuity designs (Keele et al. 2015, Gelman & Imbens 2019). Matching may be used to strengthen a weak instrumental variable (Baiocchi et al. 2010, 2012; Keele & Morgan 2016; Ertefaie et al. 2018). Multilevel matching, for instance, of schools to schools, and students to students within matched schools, was introduced by Zubizarreta & Keele (2017).

Risk set matching is used to compare treatments that may be given at various times, perhaps a treatment, such as incarceration or surgery or joining a gang, that may occur in response to

a particular development (Apel & Sweeten 2010, Haviland et al. 2008, Li et al. 2001, Lu 2005, Nieuwbeerta et al. 2009, Zubizarreta et al. 2014b). Daniel et al. (2013) provide a general discussion of treatments given at various times and subject to confounding that changes with time.

5.4. Matching and Sensitivity to Unmeasured Biases

The design of an observational study affects its sensitivity to unmeasured biases (Rosenbaum 2004, 2010; Zubizarreta et al. 2013). For the example in Section 2.1, an analysis of sensitivity to unmeasured biases is conducted in Section 3.2. The findings of such a sensitivity analysis are affected by the study's design in predictable ways, so that better decisions made during study design result in reporting insensitivity to larger biases when the data are analyzed. For instance, Section 2.1 restricted attention to children with at least six months of exposure to SSRIs, which would omit a child with five days of exposure. Omitting negligible exposures increases the design sensitivity, the limiting sensitivity to bias as the sample size increases (see Rosenbaum 2004, table 3, and Rosenbaum 2010, section 17.3). Balancing all covariates but matching closely for covariates that strongly predict the outcome also improves the design sensitivity (Rosenbaum 2005, Zubizarreta et al. 2014a), as does the use of $\kappa \geq 2$ controls per treated subject (Rosenbaum 2013). Clustered treatment assignment and multilevel matching also affect sensitivity to unmeasured biases (Hansen et al. 2014).

5.5. Omitted Technical Topics

This review of matching techniques has emphasized practical rather than technical topics. The technical literature on matching in observational studies includes the following two topics not discussed here.

First, a technical article typically contains a proof of one kind or another saying that a particular algorithm or tactic optimizes a certain function subject to certain constraints, i.e., that a well-defined optimization problem has been solved. For instance, I made such an assertion in Section 4.4 about adding row 5 to **Table 6**, but of course one needs a general statement with a proof (Rosenbaum et al. 2007). Alternatively, the article may prove that a particular algorithm provides what may be a suboptimal solution, but one that is only slightly worse than the optimal solution (Crama & Spieksma 1992, Karmakar et al. 2019, Vazirani 2010, Williamson & Shmoys 2011). Typically, one tolerates such an approximation algorithm when there is a proof that no algorithm can guarantee a quick optimal solution to large problems.

Second, a technical article may prove that, in the worst case, a particular optimization or approximation algorithm runs in time proportional to some power of the size of the problem. Results of this kind typically draw upon the extensive literature in computer science and operations research (Bang-Jensen & Gutin 2009, Bertsekas 1998, Burkard et al. 2009, Korte & Vygen 2012). For instance, minimum distance pair matching with $C \geq I$ potential controls can be implemented to run in time that is at most $O(C^3)$, although this may decline to $O\{C^2 \log(C)\}$ if the number of finite distances in the distance matrix is at most ρI for some constant $\rho \geq 1$. For comparison, sorting C numbers can be implemented to run in $O\{C \log(C)\}$ steps. Results of this kind may guide the design of algorithms for large matching problems.

Optimal matching has been described in terms of optimal pairing of the rows and columns of a matrix of distances, the so-called optimal assignment problem. Instead, it may be described in terms of minimizing the cost of a flow in a network (Bertsekas 1998, Rosenbaum 1989). At a high level of abstraction, the two formulations are equivalent, but complicated tasks are often easier to visualize, understand, and program with the aid of a network. Alternatively, matching

may be viewed as an integer program with special features that make it computationally tractable (see Korte & Vygen 2012, section 5.4, and Zubizarreta 2012). In R, the `callrelax` function of Pimentel's `rcbalance` package provides direct access to the Fortran subroutine of Bertsekas & Tseng (1988) that solves the minimum cost flow problem in a network. For more information about using gurobi to solve integer programs and other optimization problems inside R, readers are directed to **http://www.gurobi.com/products/modeling-languages/r**.

Bertsekas (1998) provides an introduction to network optimization algorithms. Attractive, concise introductions to integer programming are given by Bertsimas & Tsitsiklis (1997, chapters 7, 10, and 11) and Wolsey (1998).

Principal unobserved covariate: the scalar unobserved covariate that would permit causal inference were it observed

6. SOME THEORY

6.1. Motivation

The goal in Section 6 is to sketch a few relevant aspects of the theory of multivariate matching. Must we match exactly for observed covariates? Or does it suffice to merely balance observed covariates? Must the pairs each be exactly the same in terms of observed covariates, or is it sufficient that the treated and control groups have the same distributions of observed covariates? Clearly, in one sense, it suffices neither to match exactly nor to balance observed covariates, because bias from unmeasured covariates is possible. So we rephrase the initial question: If it did suffice to match exactly for observed covariates, would it also suffice to merely balance them? Hopefully, the answer is yes, because it is not possible to match exactly for many observed covariates at once, and happily, the answer is indeed yes. In contrast to matching exactly, balancing many observed covariates is often quite practical (see Section 6.3).

It is easy to think of many possible unobserved covariates. Does this mean that addressing the issue of unobserved covariates entails thinking about many covariates at once? At least conceptually, the answer turns out to be no: There is only one scalar unobserved covariate u that can do any harm—the principal unobserved covariate—and it always satisfies $0 \leq u \leq 1$. The principal unobserved covariate can do quite a bit of harm but, for what it is worth, that harm is one-dimensional (see Section 6.3).

Suppose one of the many observed covariates is innocuous: In some suitable sense, it predicts treatment assignment but not outcomes once you take account of the other observed covariates. Suppose that, because you cannot imagine that you are mistaken about the innocuous nature of this covariate, you decide to exclude it from matching, letting it exhibit a substantial imbalance between the treated and control groups. Perhaps you hope that this supposedly innocuous covariate will determine many treatment assignments, thereby attenuating bias from unmeasured covariates. Several investigators have examined this possibility (Brooks & Ohsfeldt 2013, Sanni et al. 2014), but in this case qualitative and quantitative findings clash (see Section 6.4). For instance, this issue arises when deciding whether or not to adjust for the provider of health care (Walker 2013, Zubizarreta et al. 2012).

6.2. Causal Effects as Comparisons of Potential Outcomes Under Competing Treatments

Write $Z = 1$ for a child in Section 2.1 who is treated with an SSRI and $Z = 0$ for a control child, so 49 children have $Z = 1$ and 6,435 potential controls have $Z = 0$. Each of the $49 + 6,435 = 6,484$ children in Section 2.1 has two potential femur bone densities, r_T if treated with an SSRI, or r_C under the control condition without SSRI treatment, but we see r_T only for the 49 treated children with $Z = 1$, and we see r_C for the 6,435 potential controls with $Z = 0$. The causal effect of SSRIs

on bone density is a comparison of r_T and r_C, such as $r_T - r_C$, but this is not observed for any child (see Neyman 1990 and Rubin 1974).

Balancing property of propensity scores: observed covariates $\mathbf{x}$ are conditionally independent of treatment Z given the propensity score

Write $R = Z\,r_T + (1 - Z)\,r_C$ for the observed bone density for a child, that is, $R = r_T$ for a treated child with $Z = 1$ or $R = r_C$ for a control child with $Z = 0$; then, we observe $(R, Z, \mathbf{x})$ for every child, where $\mathbf{x}$ is the observed covariate. A simple, though often inadequate, formalism views the 6,484 children as a random sample independently drawn from an infinite population, and this simple view is adopted in the following brief discussion.

6.3. Propensity Scores, Ignorable Treatment Assignment, and the Principal Unobserved Covariate

This section contains a brief discussion of propensity scores, mostly based on Rosenbaum & Rubin (1983). For the conditional probability of treatment, $Z = 1$, given the observed covariates, $\mathbf{x}$, write $\lambda = \lambda(\mathbf{x}) = \Pr(Z = 1|\mathbf{x})$, so that λ is a random variable as it is a function of $\mathbf{x}$, which is, itself, a random variable. Then $\lambda = \lambda(\mathbf{x})$ is called the propensity score. Because $\lambda(\mathbf{x})$ refers only to Z and $\mathbf{x}$, which are part of the observed data $(R, Z, \mathbf{x})$, we may estimate $\lambda(\mathbf{x})$ from observed data, perhaps using a model, such as a logit model predicting Z from $\mathbf{x}$.

Following Dawid (1979), write $A \perp\!\!\!\perp B \,|\, C$ for A is conditionally independent of B given C. Then we have the following covariate balancing property of propensity scores,

$$\mathbf{x} \perp\!\!\!\perp Z \,\big|\, \lambda, \tag{1.}$$

or equivalently,

$$\Pr(\mathbf{x} \mid Z = 1, \lambda) = \Pr(\mathbf{x} \mid Z = 0, \lambda), \tag{2.}$$

which says that treated and control subjects with the same propensity score, λ, have the same distribution of observed covariates, $\mathbf{x}$. In a randomized experiment in which treatments are assigned by independent flips of a fair coin, $\lambda = \lambda(\mathbf{x}) = 1/2$ for all $\mathbf{x}$, so Equation 2 becomes $\Pr(\mathbf{x} \mid Z = 1) = \Pr(\mathbf{x} \mid Z = 0)$; however, in general in observational studies, $\lambda = \lambda(\mathbf{x})$ does vary with $\mathbf{x}$, but Equation 2 says that systematic bias in the observed covariate $\mathbf{x}$ is captured by a single or scalar covariate, λ. [The proof of Equation 2 is immediate (Rosenbaum & Rubin 1983, theorem 1): Trivially, $\lambda = \Pr(Z = 1|\mathbf{x}) = \Pr(Z = 1|\mathbf{x}, \lambda)$ because $\lambda = \lambda(\mathbf{x})$ is a function of $\mathbf{x}$, and $\Pr(Z = 1|\lambda) = \mathrm{E}\{\Pr(Z = 1|\mathbf{x}, \lambda) \mid \lambda\} = \mathrm{E}\{\Pr(Z = 1|\mathbf{x}) \mid \lambda\} = \mathrm{E}(\lambda|\lambda) = \lambda$, so $\Pr(Z = 1|\mathbf{x}, \lambda) = \Pr(Z = 1|\lambda)$, proving Equation 2.]

Alas, the propensity score, $\lambda(\mathbf{x})$, is not quite the quantity that we need. The propensity score balances observed covariates, as in Equation 2, but balancing observed covariates is not generally sufficient to estimate the effects caused by treatments. Write $\zeta = \Pr(Z = 1|\mathbf{x}, r_T, r_C)$. We cannot estimate $\zeta = \Pr(Z = 1|\mathbf{x}, r_T, r_C)$ from observable data because we always observe R but never observe (r_T, r_C). In parallel with Equation 2, we have:

$$(\mathbf{x}, r_T, r_C) \perp\!\!\!\perp Z \,\big|\, \zeta$$

so that

$$\Pr(r_T, r_C \mid Z = 1, \zeta) = \Pr(r_T, r_C \mid Z = 0, \zeta) = \Pr(r_T, r_C \mid \zeta) \tag{3.}$$

and

$$\Pr(r_T, r_C \mid Z = 1, \mathbf{x}, \zeta) = \Pr(r_T, r_C \mid Z = 0, \mathbf{x}, \zeta) = \Pr(r_T, r_C \mid \mathbf{x}, \zeta).$$

(The proof of Equation 3 is the same as the proof of Equation 2, with $(\mathbf{x}, r_T, r_C)$ in place of $\mathbf{x}$ and ζ in place of λ.) If somehow we could sample a value of ζ at random, and sample one treated subject, $Z = 1$, and one control, $Z = 0$, with this value of ζ, then the matched-pair difference in their responses would be an unbiased estimate of the average treatment effect, $\mathrm{E}(r_T - r_C)$, because Equation 3 implies $\mathrm{E}(R \mid Z = 1, \zeta) = \mathrm{E}(r_T \mid Z = 1, \zeta) = \mathrm{E}(r_T \mid \zeta)$ and $\mathrm{E}(R \mid Z = 0, \zeta) = \mathrm{E}(r_C \mid Z = 0, \zeta) = \mathrm{E}(r_C \mid \zeta)$, and trivially $\mathrm{E}(r_T - r_C) = \mathrm{E}\{\mathrm{E}(r_T \mid \zeta) - \mathrm{E}(r_C \mid \zeta) \mid \zeta\}$. Indeed, by the same argument, we obtain an unbiased estimate of $\mathrm{E}(r_T - r_C)$ by a treated-minus-control pair difference in outcomes matching for both ζ and any function $\mathbf{b}(\mathbf{x})$ of the observed covariates, $\mathbf{x}$.

Treatment assignment Z is said to be ignorable given observed covariates $\mathbf{x}$ if $(r_T, r_C) \perp\!\!\!\perp Z \mid \mathbf{x}$ and $0 < \lambda(\mathbf{x}) < 1$, or equivalently, if

$$0 < \zeta = \Pr(Z = 1 \mid \mathbf{x}, r_T, r_C) = \Pr(Z = 1 \mid \mathbf{x}) = \lambda < 1. \qquad 4.$$

For instance, treatment assignment would be ignorable given $\mathbf{x}$ if treatments were assigned by independently flipping a fair coin, in which case $\Pr(Z = 1 \mid \mathbf{x}, r_T, r_C) = \Pr(Z = 1 \mid \mathbf{x}) = 1/2$. More generally, treatment assignment would be ignorable given $\mathbf{x}$ if treatments were independently assigned with coins whose probabilities of success were a function of $\mathbf{x}$ alone and were neither zero nor one. If treatment assignment were ignorable given $\mathbf{x}$, so $\zeta = \lambda$ in Equation 4, then using Equation 3 as above shows that matching for the propensity score $\lambda = \lambda(\mathbf{x})$ alone or matching for the propensity score plus any function $\mathbf{b}(\mathbf{x})$ of $\mathbf{x}$ yields an unbiased estimate of $\mathrm{E}(r_T - r_C)$.

In brief, bias from observed covariates $\mathbf{x}$ acts through a unidimensional summary, namely the propensity score, $\lambda = \lambda(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})$, while bias from observed and unobserved covariates jointly act through another unidimensional summary, namely $\zeta = \Pr(Z = 1 \mid \mathbf{x}, r_T, r_C)$. There is no bias from unobserved covariates if treatment assignment is ignorable given $\mathbf{x}$, that is, if $0 < \lambda = \zeta < 1$, and then adjustments for $\mathbf{x}$ or λ or both suffice to estimate treatment effects. If $\lambda \neq \zeta$, then only by fortunate and unlikely coincidence will adjustments for $\mathbf{x}$ or λ yield an unbiased estimate of $\mathrm{E}(r_T - r_C)$. There is always a single unobserved covariate u with $0 \leq u \leq 1$ such that $(r_T, r_C) \perp\!\!\!\perp Z \mid (\mathbf{x}, u)$, namely $u = \zeta$, as a consequence of Equation 3. Frangakis & Rubin (2002) refer to the potential outcomes, (r_T, r_C), as the principal stratification, and consistent with that terminology, one might refer to $u = \zeta = \Pr(Z = 1 \mid \mathbf{x}, r_T, r_C)$ as the principal unobserved covariate, the only one that matters. The condition $(r_T, r_C) \perp\!\!\!\perp Z \mid (\mathbf{x}, u)$ with $0 \leq u \leq 1$ gives structure to one form of sensitivity analysis in observational studies, entertaining the possibility that ignorability fails to hold, $\lambda \neq \zeta$, because the propensity score λ omits an unobserved covariate u (see Rosenbaum 1987b and Rosenbaum 2002, section 4). This is the sensitivity analysis illustrated in Section 3.2.

6.4. Attenuation of Unmeasured Biases

Does leaving an innocuous observed covariate unmatched attenuate bias from an unmeasured covariate? The current section briefly sketches results in Pimentel et al. (2016).

Partition the observed covariate $\mathbf{x}$ into $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$. Suppose throughout the following discussion that treatment assignment would have been ignorable given $(\mathbf{x}_1, \mathbf{x}_2, u)$ had the covariate u been measured, so the failure to measure u is the source of our problems. Following Heller et al. (2010), say that $\mathbf{x}_2$ is innocuous given $(\mathbf{x}_1, u)$ if $(r_T, r_C) \perp\!\!\!\perp (Z, \mathbf{x}_2) \mid (\mathbf{x}_1, u)$. If $\mathbf{x}_2$ were innocuous in this sense, then the problem remains that u is not measured, but $\mathbf{x}_2$ is not part of that problem, because treatment assignment is ignorable given $(\mathbf{x}_1, u)$. Would it be wise to leave $\mathbf{x}_2$ unmatched? Would doing this attenuate or reduce the bias from u?

Pimentel et al. (2016, definition 3) call $\mathbf{x}_2$ a prod to receive treatment if $\mathbf{x}_2 \perp\!\!\!\perp u \mid \mathbf{x}_1$ and $\text{var}\{\Pr(Z=1 \mid \mathbf{x}_1, \mathbf{x}_2, u) \mid \mathbf{x}_1, u\} > 0$. If $\mathbf{x}_2$ is a prod, then it provides no new information about u, but it induces some variation in treatment assignment Z. Pimentel et al. (2016, proposition 1) show that if $\mathbf{x}_2$ is a prod and is innocuous, then leaving $\mathbf{x}_2$ unmatched attenuates bias; essentially, it reduces the relevant value of Γ in a sensitivity analysis. This is the qualitative fact: Under strong assumptions that are difficult if not impossible to check, bias is attenuated.

Quantitatively, how large is the attenuation? In a simple model with a Gaussian x_2, if the bias from u is enormous, there can be meaningful attenuation, but the bias that remains from u is still enormous (e.g., Γ attenuates from 10 to 9.07), whereas if the bias from u is moderate, the attenuation is trivially small (e.g., Γ attenuates from 1.5 to 1.47); this calculation and others can be found in Pimentel et al. (2016, table 1). Slightly larger attenuation is possible by forcing separation on $\mathbf{x}_2$, rather than merely leaving it unmatched.

These and related results lead Pimentel et al. (2016) to argue against leaving $\mathbf{x}_2$ unmatched: Too little attenuation is produced by assumptions that are too strong and too speculative. Instead, if the claim that $\mathbf{x}_2$ is an innocuous prod seems plausible, they argue that two control groups should be constructed, one matched for $(\mathbf{x}_1, \mathbf{x}_2)$, the other for $\mathbf{x}_1$ alone perhaps forcing separation on $\mathbf{x}_2$. With two control groups formed in this way, the investigator can examine and report the situation from both perspectives, without endorsing strong speculative assumptions. Because the presence of two control groups entails multiple analyses, appropriate methods are needed to control the family-wise error rate in sensitivity analyses (Pimentel et al. 2016, section 6).

6.5. Some Extensions

Commonly, we do not sample a value λ of the propensity score at random, finding a treated and control subject with this λ. Rather, as in the example in Section 2.1, we take the entire treated group, $Z = 1$, and match its members to controls with similar λ. A parallel argument shows that if treatment assignment were ignorable given $\mathbf{x}$, then matched pairs would provide an unbiased estimate of the average effect of the treatment on the treated group, $\mathrm{E}(r_T - r_C \mid Z = 1)$ (see Rosenbaum & Rubin 1985b).

Yoon calls the quantity $\eta(\mathbf{x}) = \{1 - \lambda(\mathbf{x})\}/\lambda(\mathbf{x})$ the "entire number," and he observes that we expect to see $\eta(\mathbf{x})$ controls for each treated individual when the observed covariate equals $\mathbf{x}$. Estimating $\eta(\mathbf{x})$ can provide an additional perspective on the question in Sections 3.5 and 5.1, namely, how many controls to match to a treated individual. The entire number, $\eta(\mathbf{x})$, is particularly relevant in full matching or in matching with a variable number of controls, as $\eta(\mathbf{x})$ acts as an upper limit on the number of controls available at $\mathbf{x}$. Pimentel et al. (2015b) shows the use of the entire number in an extension of the concept of fine balance to matching with a variable number of controls.

Although the propensity score, $\lambda(\mathbf{x})$, can be estimated from observed data, we do not have direct access to the true propensity score, raising questions about inference. If $\lambda(\mathbf{x})$ followed a linear logit model with unknown parameters, then these unknown parameters may be eliminated by conditioning, with exact inferences about treatment effects based on the resulting known permutation distribution of treatment assignments; moreover, these exact inferences permit straightforward asymptotic approximations (Rosenbaum 1984). Viewed from a different perspective, an estimated propensity score resembles poststratification of a random sample in surveys; that is, the inverse-probability weighted estimate often performs somewhat better with estimated propensity scores than with the true propensity score (Rosenbaum 1987a). All of this presumes

treatment assignment is ignorable. Slight overfitting of a correctly specified propensity score has the harmless, perhaps slightly beneficial, effect of slightly overbalancing covariates in Equation 2, producing slightly better balance than coin flips.

APPENDIX: DETAILS OF THE NHANES EXAMPLE

The example is from NHANES 2005–2010, merging three surveys. The SSRIs are d03157=paroxetine, d00236=fluoxetine, d00880=sertraline, d03804=fluvoxamine, d04332=citalopram, and d04812=escitalopram, and the analysis is restricted to children with ages 8 to 20, as in Feuer et al. (2015, p. 29). As in some, but not all, analyses in Feuer et al. (2015), the analysis is restricted to children with either no current exposure to an SSRI or current exposure extending back for at least 180 days; so, children with current but brief exposures are excluded. To be included, an individual had to have had dual energy X-ray absorptiometry for the femur and had to have a nonmissing value for the duration of exposure to SSRIs. For children with $\geq$180 days of use of SSRIs, there are 49 treated children in the matched comparison, rather than 42 in Feuer et al. (2015, table 3), in part because missing covariates did not lead to exclusion. In **Table 3** and later distance tables, the 10 potential controls in the columns have NHANES identifiers 1 = 31128, 2 = 31129, 3 = 31133, 4 = 31137, 5 = 31140, 6 = 31141, 7 = 31145, 8 = 31146, 9 = 31148, and 10 = 31157, and the 4 treated children in the rows have 1 = 31430, 2 = 31641, 3 = 32390, and 4 = 32652.

SUMMARY POINTS

1. Matching is done without access to outcomes, so it is an aspect of the design of an observational study. Design ends and analysis begins when outcomes are examined for the first time. Complete and write the design section of an empirical paper before examining outcomes.
2. In parallel with a randomized clinical trial, a matched observational study should plan and define a primary analysis defined prior to the examination of outcomes.
3. Modern matching algorithms use a tool kit that includes propensity scores, minimum distance matching, near-exact matching, fine balance, and related techniques.
4. Use of several matching tools produces desired properties, such as balance for many covariates, close pairing for key covariates, and the ability to examine a subset of pairs exactly matched for a possible effect modifier. Use of several matching tools provides some robustness to the failure of any one tool used alone.
5. The balance achieved by a matching should be closely examined before accepting a matched design, before examining outcomes. Modern matching techniques provide many ways to improve a matched design that has not yet achieved adequate control for covariates.
6. Improved design of an observational study can reduce its sensitivity to unmeasured biases, as demonstrated in its sensitivity analysis, and as anticipated from theory by its design sensitivity. Improved design can better inform and address the inevitable debate that follows every consequential observational study.

FUTURE ISSUES

1. Currently, matching employs fast network optimization algorithms to build optimal comparisons of a treated and a control group. More complex designs for observational studies cannot be optimized by polynomial-time algorithms, but approximation algorithms are just beginning to provide near optimal designs in polynomial time (Karmakar et al. 2019).
2. Administrative databases are becoming ever larger yet also ever more accessible. These databases include Medicare claims data (Silber et al. 2014, 2016) and other similar health records, national data on unemployment compensation (Card et al. 2007, Lalive et al. 2006), and credit card data (Gross & Souleles 2002). Matching methods for very large databases are just beginning to be developed (Yu et al. 2019).

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Angrist JD, Krueger AB. 1999. Empirical strategies in labor economics. In *Handbook of Labor Economics*, Vol. 3, ed. O Ashenfelter, D Card, pp. 1237–366. Amsterdam: Elsevier

Apel RJ, Sweeten G. 2010. Propensity score matching in criminology and criminal justice. In *Handbook of Quantitative Criminology*, ed. AR Piquero, D Weisburd, pp. 543–62. New York: Springer

Austin PC, Stuart EA. 2015. Optimal full matching for survival outcomes: a method that merits more widespread use. *Stat. Med.* 34:3949–67

Baiocchi M, Cheng J, Small DS. 2014. Instrumental variable methods for causal inference. *Stat. Med.* 33:2297–340

Baiocchi M, Small DS, Lorch S, Rosenbaum PR. 2010. Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Am. Stat. Assoc.* 105:1285–96

Baiocchi M, Small DS, Yang L, Polsky D, Groeneveld PW. 2012. Near/far matching: a study design approach to instrumental variables. *Health Serv. Outcomes Res. Method.* 12:237–53

Bang-Jensen J, Gutin G. 2009. *Digraphs: Theory, Algorithms and Applications*. New York: Springer

Bertsekas DP. 1981. A new algorithm for the assignment problem. *Math. Prog.* 21:152–71

Bertsekas DP. 1990. The auction algorithm for assignment and other network flow problems: a tutorial. *Interfaces* 20:133–49

Bertsekas DP. 1998. *Network Optimization*. Belmont, MA: Athena Sci.

Bertsekas DP. 2001. Auction algorithms. In *Encyclopedia of Optimization*, ed. CA Floudas, PM Pardalos, pp. 73–77. New York: Springer

Bertsekas DP, Tseng P. 1988. The relax codes for linear minimum cost network flow problems. *Ann. Oper. Res.* 13:125–90

Bertsimas D, Tsitsiklis JN. 1997. *Introduction to Linear Optimization*. Belmont, MA: Athena Sci.

Brooks JM, Ohsfeldt RL. 2013. Squeezing the balloon: propensity scores and unmeasured covariate balance. *Health Serv. Res.* 48:3078–94

Burkard R, Dellamico M, Martello S. 2009. *Assignment Problems*. Philadelphia: SIAM

Card D, Chetty R, Weber A. 2007. The spike at benefit exhaustion: leaving the unemployment system or starting a new job? *Am. Econ. Rev.* 97:113–18

Cochran WG. 1965. The planning of observational studies of human populations. *J. R. Stat. Soc. A* 128:234–66

Cochran WG, Rubin DB. 1973. Controlling bias in observational studies: a review. *Sankhyā* 35:417–46
Crama Y, Spieksma FCR. 1992. Approximation algorithms for three-dimensional assignment problems with triangle inequalities. *Eur. J. Oper. Res.* 60:273–79
Crump RK, Hotz J, Imbens GW, Mitnik OA. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96:187–99
Daniel RM, Cousens S, De Stavola B, Kenward M, Sterne J. 2013. Methods for dealing with time-dependent confounding. *Stat. Med.* 32:1584–618
Daniel SR, Armstrong K, Silber JH, Rosenbaum PR. 2008. An algorithm for optimal tapered matching, with application to disparities in survival. *J. Comput. Graph. Stat.* 17:914–24
Dawid AP. 1979. Conditional independence in statistical theory. *J. R. Stat. Soc. B* 41:1–31
Derigs U. 1988. Solving non-bipartite matching problems via shortest path techniques. *Ann. Oper. Res.* 13:225–61
Ertefaie A, Small DS, Rosenbaum PR. 2018. Quantitative evaluation of the trade-off of strengthened instruments and sample size in observational studies. *J. Am. Stat. Assoc.* 113:1122–34
Feuer AJ, Demmer RT, Thai A, Vogiatzi MG. 2015. Use of selective serotonin reuptake inhibitors and bone mass in adolescents. *Bone* 78:28–33
Fisher RA. 1935. *Design of Experiments*. Edinburgh: Oliver and Boyd
Fogarty CB, Mikkelsen ME, Gaieski DF, Small DS. 2016. Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *J. Am. Stat. Assoc.* 111:447–58
Frangakis CE, Rubin DB. 2002. Principal stratification in causal inference. *Biometrics* 58:21–29
Gelman A, Imbens G. 2019. Why high-order polynomials should not be used in regression discontinuity designs. *J. Bus. Econ. Stat.* 37:447–56
Glover F. 1967. Maximum matching in a convex bipartite graph. *Naval Res. Logist.* 14:313–16
Greevy R, Lu B, Silber JH, Rosenbaum PR. 2004. Optimal multivariate matching before randomization. *Biostatistics* 5:263–75
Gross DB, Souleles NS. 2002. Do liquidity constraints and interest rates matter for consumer behavior? Evidence from credit card data. *Q. J. Econ.* 117:149–85
Hansen BB. 2004. Full matching in an observational study of coaching for the SAT. *J. Am. Stat. Assoc.* 99:609–18
Hansen BB. 2007. Flexible, optimal matching for observational studies. *R News* 7:18–24
Hansen BB, Klopfer SO. 2006. Optimal full matching and related designs via network flows. *J. Comput. Graph. Stat.* 15:609–27
Hansen BB, Rosenbaum PR, Small DS. 2014. Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *J. Am. Stat. Assoc.* 109:133–44
Haviland A, Nagin DS, Rosenbaum PR, Tremblay RE. 2008. Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data. *Dev. Psychol.* 44:422–36
Heller R, Rosenbaum PR, Small DS. 2010. Using the cross-match test to appraise covariate balance in matched pairs. *Am. Stat.* 64:2990–309
Hsu JY, Zubizarreta JR, Small DS, Rosenbaum PR. 2015. Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika* 102:767–82
Huber P. 1981. *Robust Statistics*. New York: Wiley
Karmakar B, Small DS, Rosenbaum PR. 2019. Using approximation algorithms to build evidence factors and related designs for observational studies. *J. Comput. Graph. Stat.* 28:698–709
Keele L, Morgan JW. 2016. Strengthening instruments through matching and weak instrument tests. *Ann. Appl. Stat.* 10:1086–106
Keele L, Titiunik R, Zubizarreta JR. 2015. Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *J. R. Stat. Soc. A* 178:223–39
Korte B, Vygen J. 2012. *Combinatorial Optimization*. New York: Springer
Lalive R, Van Ours J, Zweimuller J. 2015. How changes in financial incentives affect the duration of unemployment. *Rev. Econ. Stud.* 73:1009–38

Lee K, Small DS, Rosenbaum PR. 2018. A powerful approach to the study of moderate effect modification in observational studies. *Biometrics* 74:1161–70
Li YP, Propert KJ, Rosenbaum PR. 2001. Balanced risk set matching. *J. Am. Stat. Assoc.* 96:870–82
Lu B. 2005. Propensity score matching with time-dependent covariates. *Biometrics* 61:721–28
Lu B, Greevy R, Xu X, Beck C. 2011. Optimal nonbipartite matching and its statistical applications. *Am. Stat.* 65:21–30
Lu B, Rosenbaum PR. 2004. Optimal pair matching with two control groups. *J. Comput. Graph. Stat.* 13:422–34
Maritz JS. 1979. A note on exact robust confidence intervals for location. *Biometrika* 66:163–70
Meyer BD. 1995. Natural and quasi-experiments in economics. *J. Bus. Econ. Stat.* 13:151–61
Ming K, Rosenbaum PR. 2000. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* 56:118–24
Neyman J. 1990. On the application of probability theory to agricultural experiments. *Stat. Sci.* 5:465–72
Nieuwbeerta P, Nagin DS, Blokland AAJ. 2009. Assessing the impact of first-time imprisonment on offenders subsequent criminal career development: a matched samples comparison. *J. Quant. Criminol.* 25:227–57
Pimentel SD. 2016. Large, sparse optimal matching with R package rcbalance. *Obs. Stud.* 2:4–23
Pimentel SD, Kelz RR, Silber JH, Rosenbaum PR. 2015a. Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *J. Am. Stat. Assoc.* 110:515–27
Pimentel SD, Small DS, Rosenbaum PR. 2016. Constructed second control groups and attenuation of unmeasured biases. *J. Am. Stat. Assoc.* 111:1157–67
Pimentel SD, Small DS, Rosenbaum PR. 2017. An exact test of fit for the Gaussian linear model using optimal nonbipartite matching. *Technometrics* 59:330–37
Pimentel SD, Yoon F, Keele L. 2015b. Variable-ratio matching with fine balance in a study of the peer health exchange. *Stat. Med.* 34:4070–82
Rosenbaum PR. 1984. Conditional permutation tests and the propensity score in observational studies. *J. Am. Stat. Assoc.* 79:565–74
Rosenbaum PR. 1987a. Model-based direct adjustment. *J. Am. Stat. Assoc.* 82:387–94
Rosenbaum PR. 1987b. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* 74:13–26
Rosenbaum PR. 1989. Optimal matching for observational studies. *J. Am. Stat. Assoc.* 84:1024–32
Rosenbaum PR. 1991a. A characterization of optimal designs for observational studies. *J. R. Stat. Soc. B* 53:597–610
Rosenbaum PR. 1991b. Discussing hidden bias in observational studies. *Ann. Intern. Med.* 115:901–5
Rosenbaum PR. 2002. *Observational Studies*. New York: Springer
Rosenbaum PR. 2004. Design sensitivity in observational studies. *Biometrika* 91:153–64
Rosenbaum PR. 2005. Heterogeneity and causality: unit heterogeneity and design sensitivity in observational studies. *Am. Stat.* 59:147–52
Rosenbaum PR. 2007. Sensitivity analysis for M-estimates, tests, and confidence intervals in matched observational studies. *Biometrics* 63:456–64
Rosenbaum PR. 2010. *Design of Observational Studies*. New York: Springer
Rosenbaum PR. 2012. Optimal matching of an optimally chosen subset in observational studies. *J. Comput. Graph. Stat.* 21:57–71
Rosenbaum PR. 2013. Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics* 69:118–27
Rosenbaum PR. 2015a. How to see more in observational studies: some new quasi-experimental devices. *Annu. Rev. Stat. Appl.* 2:21–48
Rosenbaum PR. 2015b. Two R packages for sensitivity analysis in observational studies. *Obs. Stud.* 1:1–17
Rosenbaum PR. 2017a. Imposing minimax and quantile constraints on optimal matching in observational studies. *J. Comput. Graph. Stat.* 26:66–78
Rosenbaum PR. 2017b. *Observation and Experiment*. Cambridge, MA: Harvard Univ. Press

Rosenbaum PR, Ross RN, Silber JH. 2007. Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Am. Stat. Assoc.* 102:75–83
Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55
Rosenbaum PR, Rubin DB. 1984. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* 79:516–24
Rosenbaum PR, Rubin DB. 1985a. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am. Stat.* 39:33–38
Rosenbaum PR, Rubin DB. 1985b. The bias due to incomplete matching. *Biometrics* 41:103–16
Rosenbaum PR, Silber JH. 2009. Amplification of sensitivity analysis in matched observational studies. *J. Am. Stat. Assoc.* 104:1398–405
Rosenbaum PR, Silber JH. 2013. Using the exterior match to compare two entwined matched control groups. *Am. Stat.* 67:67–75
Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Ed. Psych.* 66:688–701
Rubin DB. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Am. Stat. Assoc.* 74:318–28
Rubin DB. 2007. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat. Med.* 26:20–36
Sanni AM, Groenwold RHH, Klungel OH. 2014. Propensity score methods and unobserved covariate balance. *Health Serv. Res.* 49:1074–82
Sekhon JS. 2009. Opiates for the matches: matching methods for causal inference. *Annu. Rev. Political Sci.* 12:487–508
Silber JH, Rosenbaum PR, McHugh MD, Ludwig JM, Smith HL, et al. 2016. Comparison of the value of nursing work environments in hospitals across different levels of patient risk. *JAMA Surg.* 151:527–36
Silber JH, Rosenbaum PR, Ross RN, Ludwig JM, Wang W, et al. 2014. Template matching for auditing hospital cost and quality. *Health Serv. Res.* 49:1446–74
Stuart EA. 2010. Matching methods for causal inference. *Stat. Sci.* 25:1–21
Stuart EA, Rubin DB. 2008. Matching with multiple control groups with adjustment for group differences. *J. Educ. Behav. Stat.* 33:279–306
Traskin M, Small DS. 2011. Defining the study population for an observational study to ensure sufficient overlap: a tree approach. *Stat. Biosci.* 3:94–118
Tukey JW. 1980. We need both exploratory and confirmatory. *Am. Stat.* 34:23–25
Vandenbroucke JP. 2004. When are observational studies as credible as randomised trials? *Lancet* 363:1728–31
Vazirani VV. 2010. *Approximation Algorithms*. New York: Springer
Walker AM. 2013. Matching on provider is risky. *J. Clin. Epidemiol.* 66:565–68
Williamson DP, Shmoys DB. 2011. *Design of Approximation Algorithms*. Cambridge, UK: Cambridge Univ. Press
Wolsey LA. 1998. *Integer Programming*. New York: Wiley
Wu CF, Hamada MS. 2011. *Experiments: Planning, Analysis, and Optimization*. New York: Wiley
Yang D, Small DS, Silber JH, Rosenbaum PR. 2012. Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics* 68:628–36
Yu R, Silber JH, Rosenbaum PR. 2019. Matching methods for observational studies derived from large administrative databases. *Stat. Sci.* In press
Zubizarreta JR. 2012. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Am. Stat. Assoc.* 107:1360–71
Zubizarreta JR, Cerdá M, Rosenbaum PR. 2013. Effect of the 2010 Chilean earthquake on posttraumatic stress: reducing sensitivity to unmeasured bias through study design. *Epidemiology* 24:79–87
Zubizarreta JR, Keele L. 2017. Optimal multilevel matching in clustered observational studies: a case study of the effectiveness of private schools under a large-scale voucher system. *J. Am. Stat. Assoc.* 112:547–60

Zubizarreta JR, Neuman M, Silber JH, Rosenbaum PR. 2012. Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia. *J. Am. Stat. Assoc.* 107:901–15

Zubizarreta JR, Paredes RD, Rosenbaum PR. 2014a. Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Ann. Appl. Stat.* 8:204–31

Zubizarreta JR, Reinke CE, Kelz RR, Silber JH, Rosenbaum PR. 2011. Matching for several nominal variables in a case-control study of readmission following surgery. *Am. Stat.* 65:229–38

Zubizarreta JR, Small DS, Rosenbaum PR. 2014b. Isolation in the construction of natural experiments. *Ann. Appl. Stat.* 8:2096–121

Annual Review of Statistics and Its Application

Volume 7, 2020

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics