2021

Contributions To Multivariate Matching In Observational Studies

Ruoqi Yu
University of Pennsylvania

Contributions To Multivariate Matching In Observational Studies

Abstract

Matching is a common approach to reduce bias in observed covariates to draw reliable causal inferences in observational studies. This thesis consists of three papers discussing new methods for conducting, evaluating, and improving matching designs in observational studies. The first paper presents new optimal matching techniques for large-scale observational data. This new method reduces the computational complexity and preserves appealing properties in terms of balancing covariates. After constructing a matched sample, it is essential to assess the covariate balance of the matched data since lack of balance in covariates can induce a bias of the estimated treatment effect. The second paper discusses a formal evaluation of covariate balance. This new assessment evaluates whether the match is adequate compared to randomized experiments and identifies the major problems, guiding how to improve the covariate balance. If diagnostics suggest that the current match is not satisfactory, how can we improve the quality of matched samples? The final paper utilizes the idea of directional penalties, which can improve covariate balance in a matched sample effectively, even for a large observational study.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Statistics

First Advisor

Paul R. Rosenbaum

Subject Categories

Statistics and Probability

CONTRIBUTIONS TO MULTIVARIATE MATCHING IN OBSERVATIONAL STUDIES

Ruoqi Yu

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

Paul R. Rosenbaum, Robert G. Putzel Professor; Professor of Statistics

Graduate Group Chairperson

Nancy R. Zhang, Ge Li and Ning Zhao Professor; Professor of Statistics

Dissertation Committee

Paul R. Rosenbaum, Robert G. Putzel Professor; Professor of Statistics

Dylan S. Small, Class of 1965 Wharton Professor of Statistics; Department Chair

Jeffrey H. Silber, Professor of Health Care Management, Pediatrics, Anesthesiology and Critical Care; Director, Center for Outcomes Research; Nancy Abramson Wolfson Endowed Chair in Health Services Research, Children's Hospital of Philadelphia

CONTRIBUTIONS TO MULTIVARIATE MATCHING
IN OBSERVATIONAL STUDIES

ACKNOWLEDGMENT

First and foremost, I would like to thank my advisor Paul Rosenbaum for his guidance and support in every stage of my growth to be an independent researcher. Paul is a role model – he shared his passion for statistics and his understanding of the field and showed me how to be a good scholar. Besides introducing me to a wide range of exciting research problems, Paul shared insightful feedback during our weekly meetings, which guided me to sharpen my thinking and brought my work to a higher level. Paul also gave incredible patience and uplifting encouragement during the challenging times and provided wise and thoughtful advice on any aspect of my academic life. Without him, this dissertation would not be possible.

I am also deeply thankful to the other members of my dissertation committee. Dylan Small was first an instructor of several foundational courses and then a mentor and co-author in more recent years. His creative insights on research problems and interesting collaborative examples have always been inspirational. Dylan's encouragement and support were crucial for me in several stages of my studies. Jeffrey Silber introduced me to rich datasets and exciting questions in modern medical research, which motivated my methodological research and allowed me to see how statisticians can make a real impact. Jeffrey provided me insightful advice and thrilling research opportunities at the Center for Outcomes Research. This unique experience of working with his group helped me grow as a good collaborator to engage with researchers in other fields.

I feel fortunate to have been part of the Statistics Department at Wharton as a graduate student. Gidget Murray and Maggie Saia of Wharton Doctoral Programs cheerfully guided me through the administrative process of obtaining my degree. I would like to thank the faculty of the Statistics Department, particularly Mark Low, Eric Tchetgen Tchetgen, Jian Ding, Weijie Su, Bhaswar Bhattacharya, Linda Zhao, Warren Ewens, Paul Shaman, and Bob Stine, and its staff, particularly Noelle Felipe, Gabby Frisone, Adam Greenberg, Nick

Ongpauco, Carol Reich, Sarin Sieng, and Tanya Winder, for their advice and support in different aspects of my life as a graduate student. I am also very grateful for the friendship of my fellow graduate students in the Statistics Department during our shared journey, especially my cohort-mates – Somabha Mukherjee, Saeed Sharifi-Malvajerdi, Matteo Sordello, Yichen Wang, and Mateo Wirth. I wish to thank several former students of the department, including Xinyao Ji, Samuel Pimentel, and Jose Zubizarreta.

Finally, I am deeply grateful for the unconditional love and endless support of my parents and Shulei. They have always been there for me, through the good times and more challenging ones. Thank you for everything.

ABSTRACT

CONTRIBUTIONS TO MULTIVARIATE MATCHING IN OBSERVATIONAL STUDIES

Ruoqi Yu

Paul R. Rosenbaum

Matching is a common approach to reduce bias in observed covariates to draw reliable causal inferences in observational studies. This thesis consists of three papers discussing new methods for conducting, evaluating, and improving matching designs in observational studies. The first paper presents new optimal matching techniques for large-scale observational data. This new method reduces the computational complexity and preserves appealing properties in terms of balancing covariates. After constructing a matched sample, it is essential to assess the covariate balance of the matched data since lack of balance in covariates can induce a bias of the estimated treatment effect. The second paper discusses a formal evaluation of covariate balance. This new assessment evaluates whether the match is adequate compared to randomized experiments and identifies the major problems, guiding how to improve the covariate balance. If diagnostics suggest that the current match is not satisfactory, how can we improve the quality of matched samples? The final paper utilizes the idea of directional penalties, which can improve covariate balance in a matched sample effectively, even for a large observational study.

TABLE OF CONTENTS

ACKNOWLEDGMENT iii

ABSTRACT v

LIST OF TABLES x

LIST OF ILLUSTRATIONS xiii

CHAPTER 1 : Introduction 1

CHAPTER 2 : Matching Methods for Observational Studies Derived from Large Administrative Databases 4

2.1 Introduction: the Problem; an Example; Outline 4

2.2 An Informal Discussion of Optimal Matching for Large Databases 8

2.3 Practical Aspects of Matching in Large Databases 19

2.4 Network Structure: A Sparse Bipartite Graph Expanded for Near-fine Balance 21

2.5 Constructing the Matched Sample in the Medicaid Example 30

2.6 Summary 39

CHAPTER 3 : Evaluating and Improving a Matched Comparison of Antidepressants and Bone Density 40

3.1 Introduction: Problem and Empirical Example 40

3.2 Comparing Covariate Balance with Complete Randomizations 44

3.3 From Diagnosed Imbalances to Better Matched Samples 49

3.4 A Simple but Useful Implementation 50

3.5 Simulation 52

3.6 Analysis of Bone Density and Antidepressants 56

3.7 Discussion 59

CHAPTER 4 : Directional Penalties for Optimal Matching in Observational Studies 62

4.1 Introduction: Example, Outline, the Simplest Case 62

4.2 Optimal Matching with Asymmetric Adjustments to Distances 70

4.3 Smoking and Homocysteine 78

4.4 Discussion and Extensions 80

APPENDIX 82

BIBLIOGRAPHY 95

LIST OF TABLES

TABLE 1 : In addition to matching for 463 principal surgical procedures and 973 principal diagnoses, the match controls the demographic covariate and comorbid conditions below. The table shows the covariate mean for children in children's hospitals (treated) and children in adult hospitals (control), for 38,841 matched controls and 159,527 controls before matching. The standardized difference is the absolute difference in means divided by an equally weighted pooled standard deviation before matching. Standardized differences above 0.2 standard deviations are in **bold**. 7

TABLE 2 : Balance in 463 Principal Procedures, 973 Principal Diagnoses, and their 463 × 973 interactions. The imbalance in the actual matched sample is compared to the minimum imbalance and the mean imbalance in 10,000 randomized experiments. For each covariate, by each measure, the matched sample is closer to balance than the most balanced of 10,000 randomized experiments formed from the same data. 36

TABLE 3 : Mortality within 30 days of surgery in 38,841 matched pairs of two children, one receiving surgery in a children's hospital, the other in an adult hospital. The table counts pairs, not children. 38

TABLE 4 : Balance table for marginal distributions of 19 covariates and propensity score (pscore), with estimated p-values based on 2,000 simulated randomized experiments: Before match, basic match (Base), and final match (Iter 3). Standardized mean differences (SMDs) greater than 0.1 in absolute value and p-values less than 0.1 are in **bold**. . 43

TABLE 5 : Comparison of bias, variance, and mean squared error of estimated average treatment effects on the treated based on 100 replications for matching on X_1 (Benchmark, with package `DiPs`), the proposed iterative algorithm with univariate GFKS (GFKS-1) and bivariate GFKS (GFKS-2), inverse probability weighting with the propensity score estimated by random forests (Random forest, with package `randomForest`), covariate balance propensity score (CBPS, with package `CBPS`) and minimal weights (MW, with package `sbw`) in the simulations. Here, $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_5)$ in the treated group and $\mathbf{X} \sim N((1, 0.2, ..., 0.2), \mathbf{I}_5)$ in the control group. Consider two simulation experiments by setting response as $Y_1 = X_1 + \epsilon_1$ and $Y_2 = X_1^2 + \epsilon_2$, where $\epsilon_1, \epsilon_2 \sim N(0, 1)$ are independent errors. 57

TABLE 6 : Covariate means for 548 daily smokers individuals (Treated), all 2211 never smokers (All), and four matched samples of 1096 never smokers, each matched 2-to-1 to the daily smokers (M1-M4). Also shown are: (i) the corresponding standardized differences in covariate means, treated-minus-control, where the demoninator is an unchanging pooled within group standard deviation before matching), and (ii) the conventional pooled, two sample t-statistics. Standardized differences ≥ 0.1 and t-statistics ≥ 2 in absolute value are in **bold**. 63

TABLE 7 : Robust Mahalanobis distances before and after matching. "All" refers to the 548×2211 distances before matching. For each match, there are 2×548 distances within matched sets in these 2-to-1 matches. Values are the mean, minimum, quartile 1, median, quartile 3 and maximum. 66

TABLE 8 : Bias after matching in one Normally distributed covariate when matching with a possibly asymmetric caliper $x_t - x_c \in [-\eta_1, \eta_2]$. The bias before matching is μ. The smallest absolute bias in each column is in **bold**. 70

TABLE 9 : Directional penalties λ_k used in the three matched samples, M1, M2 and M3. For three covariates, Female, Black and BMI, only the directional penalty, (4.4) or (4.5), was used. For the propensity score, a directional penalty, a hockey stick penalty (4.3) and a sometimes asymmetric caliper (4.6) was used. In the hockey stick (4.3) and caliper (4.6), the propensity score was scaled by the denominator of the standardized difference, namely the pooled, equally weighted standard deviation, s, before matching. 74

TABLE 10 : Estimated average treatment effects on the treated comparison for four outcomes: femur bone mineral density, femur bone mineral content, femoral neck bone mineral density and femoral neck bone mineral content. 95

LIST OF ILLUSTRATIONS

FIGURE 1 : Five bipartite graphs, where the vertical axis is the propensity score. There is a decision variable for each potential pairing of a treated subject and a potential control, that is, a decision variable for each edge. Graph (i) has all possible pairings. Graph (ii) has reduced the number of edges by cutting the graph into four parts at the quartiles, where these parts will be matched separately. Graph (iii) has a caliper that is just a little too small, so pair matching is not feasible. Graph (iv) has the smallest feasible caliper. Graph (v) has both the smallest caliper and the smallest upper bound on the number of edges for treated units. 9

FIGURE 2 : A bipartite graph matching exactly for gender, expanded for near fine balance of race, v, black or other. The optimal caliper is now 0.1925, and with this caliper the minimum number of neighbors is $\nu = 3$. The duplicate edges connect γ to γ', with capacity 1, so they insist that a control may be matched at most once. The solid grey edges retain feasibility through a penalized bypass, β, of the fine balance constraints. 16

FIGURE 3 : Creating the bipartite graph by exact matching for 463 Principal Procedures with an optimal caliper on the rank of the propensity score. There are 38,841 treated nodes and 159527 control nodes. (i) The propensity score before matching. (ii) Ranks of the propensity score before matching. (iii) Distribution of the number of edges for each treated unit with an optimal caliper on the propensity score, before and after determining the minimal number, $\nu = 105$, of near neighbors. (iv) The "after" boxplot from panel (iii) scaled so that detail is visible. 32

FIGURE 4 : Change in covariate imbalance from before matching to after matching for the 29 covariates in Table 1. The point of the arrow is after matching. A vertical line is at 0.1. After matching, all standardized differences are less than 0.1, and all large imbalances before matching have been greatly reduced. 36

FIGURE 5 : Comparison of individual two-sample t test (t-individual), individual Kolmogorov-Smirnov test (KS-individual), adjusted t test (t-adjusted), adjusted Kolmogorov-Smirnov test (KS-adjusted), univariate GFKS (GFKS-1) and bivariate GFKS (GFKS-2) in the simulations. Sim-1: $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_{10})$ in the treated group and $\mathbf{X} \sim N((\delta, 0, ..., 0), \mathbf{I}_{10})$ in the control group; Sim-2: $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_{10})$ in the treated group and $\mathbf{X} \sim N((\delta, \delta/2, ..., \delta/2), \mathbf{I}_{10})$ in the control group; Sim-3: $\mathbf{X} \sim t_3(\mathbf{0}, \mathbf{I}_{10})$ in the treated group and $\mathbf{X} \sim t_3((\delta, \delta/2, ..., \delta/2), \mathbf{I}_{10})$ in the control group; Sim-4: $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_{10})$ in the treated group and $\mathbf{X} \sim N(\mathbf{0}, \Sigma^*)$ in the control group, where Σ^* is a 10×10 matrix such that $\Sigma^*_{ii} = 1$, $\Sigma^*_{12} = \Sigma^*_{21} = \delta$, and $\Sigma^*_{ij} = 0$ otherwise. 54

FIGURE 6 : Boxplots of p-values of GFKS tests, standardized mean differences and two-sample t statistics for M0 (Basic match), M1 (Iteration 1), M2 (Iteration 2), M3 (Iteration 3). 60

FIGURE 7 : Four covariates in match M4. Treated = daily smokers, matched = matched controls in match M4, and unmatched = never smokers not used in match M4. One unmatched never smoker with a BMI of 130.2 is not displayed. For education, 1 is less than ninth grade, 3 is high school or equivalent, and 5 is a college graduate. Income is recorded as the ratio of family income to the poverty level, capped at 5×poverty. BMI, body mass index 67

FIGURE 8 : Estimated power and type I error of two-sample t test on individual covariates (t-individual), Kolmogorov-Smirnov test on individual covariates (KS-individual), adjusted t test (t-adjusted), adjusted Kolmogorov-Smirnov test (KS-adjusted), univariate GFKS (GFKS-1), bivariate GFKS (GFKS-2), and random forest based permuatation test (CPT (Random forest)). 93

CHAPTER 1 : Introduction

Many questions of interest in statistical applications (e.g., economics, social sciences, and biomedical research) are essentially about causality rather than association (Imbens and Rubin, 2015). Ideally, we study causal relationships with randomized experiments, which are not always practical or ethical. As such, causal effects are often studied with non-randomized observational studies. The quality and strength of evidence obtained with an observational study mainly depend on its design (Rosenbaum, 2010). Therefore, investigators should design observational studies carefully to approximate randomized experiments to obtain objective causal inference and more reliable answers to scientific questions (Rubin, 2007). Towards this goal, matching methods are developed as a common approach in the design stage to construct similar treated and control groups in terms of observed covariates. These tools allow transparent and interpretable inferences about the effects of interventions and provide opportunities to study the impact of potential unobserved confounding variables.

This thesis organizes three papers discussing new methods for conducting, evaluating, and improving matching samples in observational studies. Each chapter can be read independently without knowledge of the content of the other chapters.

As computing tools have rapidly developed in recent decades, data sets have grown in size while also becoming more accessible for analysis. In current practice, large matched samples are constructed by subdividing the population and solving a series of smaller problems, which may not be feasible and may sacrifice the good quality of matched samples. The first paper presents a new optimal matching technique for large observational studies. This new method utilizes a single match using everyone in the study and accelerates the computations in a different way. In particular, we use an iterative form of Glover's algorithm for a doubly convex bipartite graph to determine an optimal caliper for the propensity score, radically reducing the number of candidate matches; then, we optimally match in a large but

much sparser graph. This new method reduces the computational complexity and preserves appealing properties in terms of balancing covariates. For instance, matching techniques like fine-balance can be used on a much larger scale, improving their effectiveness. The method is applied to a large administrative data set from US Medicaid, matching children receiving surgery at a children's hospital to similar children receiving surgery at a hospital that mostly treats adults. In the example, we form 38,841 matched pairs from 159,527 potential controls within 30 minutes, controlling for 29 covariates plus 463 Principal Surgical Procedures, plus 973 Principal Diagnoses. This paper is a joint work with Jeffrey Silber and Paul Rosenbaum, and was published (with discussion and rejoinder) in 2020 in Volume 35, Issue 3 of *Statistical Science*.

After conducting a matched sample, it is essential to assess the covariate balance of the matched data since the lack of balance in covariates can induce a bias in the estimated treatment effects. When should we worry about a residual imbalance? This evaluation step is usually done informally, in ways that have several limitations. First, there are many diagnostics, even if covariates are assessed one at a time, which raises multiplicity issues. Also, joint distributions of covariates, even bivariate distributions, are often ignored. Besides, it is an open question whether diagnostics identify the major problems. The second paper discusses a formal assessment of covariate balance to address these issues. Unlike the common informal diagnostics, the proposed method compares marginal distributions and joint distributions of the matched sample with those of the benchmark, complete randomizations. In doing so, it can distinguish unsolvable matching problems, problems that can be solved by a better match, and those that are small compared to the imbalances that occur by chance in randomized experiments. This new method controls the probability of falsely identifying a covariate imbalance among many comparisons, yet it has a high probability of correctly detecting and identifying a major problem. This paper is forthcoming in *Biometrics* DOI: 10.1111/biom.13374.

If diagnostics suggest the current match is not satisfactory, how can we improve the qual-

ity of the matched samples? Multivariate matching in observational studies tends to view covariate differences symmetrically. For instance, a difference of 10 years in age is thought equally problematic whether the treated subject is older or younger than the matched control. However, correcting the bias may be easier if matching tries to avoid the typical case that creates the bias. The third paper describes several easily used, asymmetric, directional penalties and illustrates how they can improve covariate balance in a matched sample. The investigator starts with a matched sample built in a conventional way, then diagnoses residual covariate imbalances in need of reduction, and achieves the needed reduction by slightly altering the distance matrix with directional penalties, creating a new matched sample. This proposal is compatible with the methods developed in the first paper so that it can improve large matched samples efficiently. We also explore the connection between directional penalties and a widely used technique in integer programming, namely Lagrangian relaxation of problematic linear side constraints in a minimum cost flow problem. In effect, many directional penalties are Lagrange multipliers, pushing a matched sample in the direction of satisfying a linear constraint that would not be satisfied without penalization. This paper is a joint work with Paul Rosenbaum, and was published in 2019 in Volume 75, Issue 4 of *Biometrics*.

CHAPTER 2 : Matching Methods for Observational Studies Derived from Large Administrative Databases

2.1. Introduction: the Problem; an Example; Outline

2.1.1. Matching for observational studies derived from administrative data sets

As administrative records have moved from file cabinets to computers, administrative data sets have grown in size while also becoming more accessible for analysis. For instance, using US Medicare data, Silber et al. (2016) formed 25,076 matched pairs of two patients comparing surgical outcomes at hospitals with superior and inferior nursing, finding lower morality and reduced use of the intensive care unit at hospitals with superior nursing. Using data from the Pediatric Health Information System (PHIS), Silber et al. (2018) formed 23,582 matched pairs of two children, comparing the surgical outcomes of children on Medicaid to the outcomes of similar children with other forms of health insurance.

Matched observational studies are commonly constructed using propensity scores (Rosenbaum and Rubin, 1985b), externally estimated prognostic or risk scores (Hansen, 2008), covariate distances (Rubin, 1980), fine balance constraints (Rosenbaum et al., 2007; Yang et al., 2012; Zubizarreta, 2012; Pimentel et al., 2015a), and minimum cost flow algorithms that minimize the total distance within matched pairs or matched sets (Rosenbaum, 1989; Hansen and Klopfer, 2006; Lu et al., 2011). These techniques are straightforward and work well with a few thousand people, but they encounter computational difficulties with administrative data sets containing tens or hundreds of thousands of people. For reviews of matching methods, see Rosenbaum (2010) and Stuart (2010).

In current practice, 100,000 people are not matched in a single optimization; rather, people are subdivided into, say, fifty bins by matching exactly for a few discrete or rounded covariates; then, within each bin, thousands of people are matched optimally. This approach is neither unreasonable nor impractical, but it has aspects that are not attractive. Exact matching in bins gives overriding importance to the covariates that define the bins,

and there may be no scientific basis for this. Other covariates of equal importance may be inadequately matched because close matches between bins are forbidden. Categorizing continuous covariates, such as the propensity score, to make exact-match bins forbids close matches on the propensity score that cross category boundaries, while tolerating larger gaps inside categories. If you divide Medicare surgeries into bins by matching exactly for ICD-9 or ICD-10 principal surgical procedures, then you find that some surgeries, such as knee replacement surgery, are so common that its bin is still too large to match, whereas other surgeries are so rare that their bins need to be merged before the matching bin is large enough to match. Creating bins of practical size then has subjective aspects that may be left to a statistical programmer, with the consequence that some of the decisions that led to the match are not automatic, hence not reproducible by someone else.

More importantly, there are substantial statistical advantages to matching everyone at once. A matching technique called "fine balance" tries to balance covariates without pairing individuals who have the same values of these covariates (Rosenbaum et al., 2007). Fine balance makes groups comparable by counter-balancing — an imbalance in one pair is counterbalanced in another — as in a Latin square design, rather than seeking to pair identical individuals, as in a blocked design. There are many more opportunities for fine balance when more people are matched at the same time. Splitting 100,000 people into 50 bins unnecessarily limits what fine balance can do.

2.1.2. Surgery for children: Are outcomes better in children's hospitals?

A child may have surgery at a conventional hospital that mostly treats adults, or at a hospital dedicated to the treatment of children, such as Boston Children's Hospital or the Children's Hospital of Philadelphia. Does this choice matter? Do outcomes differ? We are interested in those surgical procedures that offer a genuine choice. A handful of specialized or especially risky surgical procedures for children are almost invariably performed at children's hospitals, and we will exclude these, focusing instead on the vast majority of procedures commonly performed on children at adult hospitals.

We look at data from Medicaid for 2009-2012. We have 203,163 children admitted for surgical procedures in which both the Principal (Surgical) Procedure and the Principal Diagnosis were not missing, and of these, 41,319 procedures were performed in a children's hospital, or about 20%. So 4 in 5 surgeries on children are performed at adult hospitals. There were 504 distinct surgical procedures, 3 of which were never performed at children's hospitals. We excluded 38 of the 504 surgical procedures where the majority of children were treated at children's hospitals, consistent with our goal of focusing on those procedures that are typically done at adult hospitals, leaving $504 - 3 - 38 = 463$ procedures.

After this exclusion, there were 198,368 surgical admissions, of which 38,841 were at children's hospitals, and there remained 463 distinct surgical procedures and 973 distinct principal diagnoses. Additionally, Table 1 lists other covariates, including demographic variables such as age, sex and race, comorbid conditions such as cancer and congenital anomalies, and the intensity of health care services in the past six months, such as operations, emergency department (ED) visits, and office visits. In Table 1, operations in the past six months distinguish two lists of operations, a narrow list of clearly relevant procedures, and a broad list including additional procedures. Notably, before matching, the children treated at children's hospitals rather than adult hospitals are younger, have more congenital anomalies (18.9% versus 7.9%) and other comorbid conditions, and have more visits to a hospital's emergency room in the past six months.

The standardized differences in Table 1 are the absolute treated-minus-control difference in means for a covariate divided by a pooled standard deviation before matching. The pooled standard deviation gives equal weight to the treated and control groups, and it always refers to the distribution before matching. In contrast, the numerator of the standardized difference is different before and after matching, so it is a standardized measure of improvement in balance in a covariate afforded by matching. This measure is traditional, and was used in Cochran and Rubin (1973) and Rosenbaum and Rubin (1985b).

We will form 38,841 matched pairs of two children, one in a children's hospital, one in

Table 1: In addition to matching for 463 principal surgical procedures and 973 principal diagnoses, the match controls the demographic covariate and comorbid conditions below. The table shows the covariate mean for children in children's hospitals (treated) and children in adult hospitals (control), for 38,841 matched controls and 159,527 controls before matching. The standardized difference is the absolute difference in means divided by an equally weighted pooled standard deviation before matching. Standardized differences above 0.2 standard deviations are in **bold**.

Covariate	Covariate Mean			Standardized Difference	
	Treated	Controls			
		Matched	All	Matched	All
Sample size	38841	38841	159527	38841	159527
Year admitted	2010.753	2010.725	2010.492	0.025	**0.238**
Age	8.338	8.489	10.310	0.027	**0.350**
Male	0.551	0.559	0.563	0.018	0.024
Black	0.159	0.157	0.184	0.004	0.067
Hispanic	0.301	0.291	0.290	0.021	0.024
Race, other	0.177	0.156	0.134	0.060	0.121
Autoimmune disorder	0.003	0.002	0.002	0.019	0.025
Blood disorder	0.046	0.034	0.046	0.055	0.003
Cancer	0.063	0.054	0.035	0.043	0.128
Cerebral palsy	0.072	0.057	0.028	0.070	**0.203**
Chromosomal anomaly	0.027	0.017	0.011	0.074	0.120
Congenital heart disease	0.091	0.078	0.039	0.053	**0.215**
Congenital anomaly	0.189	0.161	0.079	0.085	**0.328**
Diabetes	0.010	0.007	0.011	0.029	0.011
Enteritis/digestive disorder	0.019	0.016	0.010	0.030	0.077
Epilepsy/seizure	0.086	0.071	0.056	0.059	0.118
HIV	0.001	0.001	0.001	0.003	0.004
Immunocompromised	0.014	0.006	0.004	0.078	0.098
Major Organ Dysfunction	0.036	0.030	0.019	0.039	0.108
Mental retardation	0.131	0.107	0.076	0.079	0.182
Metabolic disorder	0.025	0.019	0.020	0.044	0.036
Muscular dystrophy	0.002	0.001	0.001	0.023	0.039
Neurodegenerative Disease	0.052	0.043	0.024	0.043	0.142
Other respiratory	0.011	0.007	0.004	0.046	0.081
Mean count of health services in the past 6 months					
Hospitalizations	0.184	0.148	0.128	0.058	0.089
Operations, broadly defined	0.084	0.067	0.057	0.051	0.080
Operations, narrowly defined	0.041	0.035	0.024	0.027	0.078
Emergency Department visits	2.681	2.366	2.075	0.068	0.131
Office visits	4.706	4.695	4.095	0.001	0.073

an adult hospital. The match will balance the 463 procedures, the 973 diagnoses, their $463 \times 973 = 450,499$ interactions, plus the covariates listed in Table 1. As the ratio of interaction categories to children in the study is 2.3, there is no realistic hope of modeling all of the interactions, but they can be balanced.

In current practice, a matching problem as large as this would be divided into 20 to 50 smaller problems. In sharp contrast, using new methods proposed in this paper, we will match the 198,368 children in a single optimization.

2.1.3. Outline of the paper: Concepts in pictures, formal results, application

Section 2.2 uses a toy example with 30 individuals and a few drawings to indicate the changes we suggest for matching in large administrative databases. A toy example is useful because a person can inspect a graph with 30 nodes and see what is happening, but real matching problems are vastly larger. This discussion is divided in half, with §2.2.1 removing edges from a graph, and with §2.2.2 bringing in the key concept of fine balance. The practical aspects of creating a match are briefly sketched in §2.3. Then, §2.4 develops the topic formally and in greater generality. A goal in §2.4 is to quantify the reduction in computational effort produced by the ideas informally introduced in §2.2. We illustrate the technique in §2.5 using the Medicaid data mentioned in §2.1.2. Proofs are given in the Appendix A.1.

2.2. An Informal Discussion of Optimal Matching for Large Databases

2.2.1. A motivating picture illustrating some issues and methods

Dense graphs offer too much choice, including obviously bad choices.

To fix ideas, Figure 1 is a picture of a toy version of the problem, omitting for the moment the important issue of fine balance. The example uses public data from the 2005-2006 National Heath and Nutrition Examination Survey (NHANES), with 7 daily smokers and 23 nonsmokers as potential controls. This example is a random sample of size 30 from the `nh0506` data in the `R` package `bigmatch`, obtained by "`set.seed(20)`" followed by "`nhs<-nh0506[sample(1:(dim(nh0506)[1]),30),]`". The reader may find it helpful to

try the methods we describe using the small data set `nh0506`; it describes 2475 people.

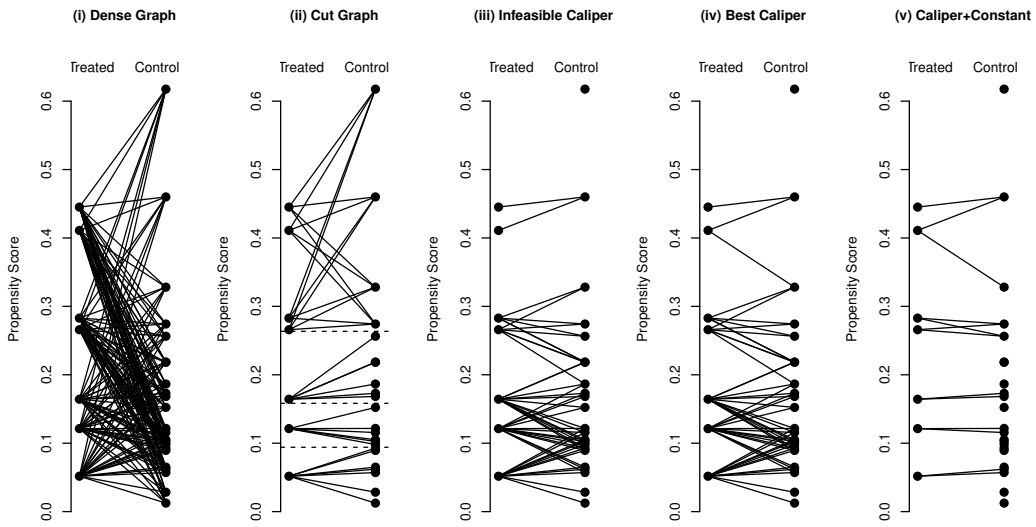


Figure 1: Five bipartite graphs, where the vertical axis is the propensity score. There is a decision variable for each potential pairing of a treated subject and a potential control, that is, a decision variable for each edge. Graph (i) has all possible pairings. Graph (ii) has reduced the number of edges by cutting the graph into four parts at the quartiles, where these parts will be matched separately. Graph (iii) has a caliper that is just a little too small, so pair matching is not feasible. Graph (iv) has the smallest feasible caliper. Graph (v) has both the smallest caliper and the smallest upper bound on the number of edges for treated units.

In Figure 1, 30 people are represented by dots or nodes, 7 treated and 23 controls, two of which are controls with virtually identical propensity scores 0.2186 and 0.2183 so their nodes are not visibly different in Figure 1. Each panel of Figure 1 is a so-called bipartite graph, meaning two parts, treated and control. In a bipartite graph, the edges connect nodes in different parts. Figure 1(i) has every possible edge, or $161 = 7 \times 23$ edges, so it is said to be a complete and dense bipartite graph. Candidate matches are represented by line segments or edges. In optimal bipartite matching, each edge is a binary decision variable: Should this treated node be matched to this connected control node, or to someone else? In the simplest case, a pair matching, we pair every treated node to a different control node, so we pick 7 edges that do not share a node in Figure 1(i). Each edge has a cost or distance

attached to it, where the distance measures how close a treated subject is to a control in terms of measured covariates. The cost or distance may involve the propensity score, a Mahalanobis distance of some kind, and other considerations. In §2.2.1, we seek a pair matching that minimizes the total cost over the 7 chosen edges, a standard combinatorial optimization problem; however, in §2.2.2 we impose additional balance constraints on the match. The problem is not trivial because two treated nodes may both want the same potential control, so you cannot pair each treated node to the closest control. Matching with multiple controls is discussed in §2.4.6. For pair matching, Figure 1(i) would entail optimizing a function of $161 = 7 \times 23$ binary decision variables subject to various constraints that require 7 nonoverlapping treatment/control pairs. If we matched 1000 treated people to 2000 potential controls, Figure 1(i) would have $2 \times 10^6 = 1000 \times 2000$ edges, a practical size for optimal matching. If we matched 30,000 treated people to 60,000 potential controls in a small administrative data base, Figure 1(i) would have $1.8 \times 10^9 = 30,000 \times 60,000$ edges, and optimal matching using Figure 1(i) would not be practical in 2018. The difficulty of optimal matching grows much faster than linearly with the number of edges, so a problem with 1.8×10^9 edges is much more than 1000 times harder than a problem with 2×10^6 edges; see §2.4.4 for specifics. Expressing the same thought in other words, it would be much easier to solve 1000 problems of size 2×10^6 than to solve one problem of size 1.8×10^9.

Cutting up a large graph into many smaller graphs works but has unattractive limitations.

It might take only a glance to realize that many of the $161 = 7 \times 23$ possible pairings are terrible, perhaps differing greatly on the propensity score; therefore, many candidate pairings really do not deserve serious consideration. We do not want to match the upper left treated node to the lower right control node because their propensity scores are far apart. So we might be willing to change the optimal matching problem in Figure 1(i) into a different problem that can be solved more quickly in large data sets. To emphasize, we will not solve the optimization problem in Figure 1(i), but replace that problem by another reasonable problem that can be solved.

Figure 1(ii) is a toy version of current practice in large data sets. Figure 1(ii) splits the treated and control population into 4 subpopulations or bins, here using the quartiles of the propensity score, which appear in Figure 1(ii) as horizontal dashed lines. Notice that Figure 1(ii) removes all edges from Figure 1(i) that would have crossed a dashed horizontal line. Figure 1(ii) has four connected components: a treated subject is only connected to — can only be matched to — controls in the same stratum defined by quartiles of the propensity score. Figure 1(ii) is a graph with four connected components, each of which is a complete bipartite graph, with a total of 35 edges or decision variables, rather than 161 in Figure 1(i). Figure 1(ii) has fewer edges or decision variables in total, but additionally there are four unrelated small problems that can be solved one-by-one, rather than one large problem. For instance, the problem in the lowest bin of Figure 1(ii) is trivial: pick the one control with the smallest covariate distance to the one treated unit. In the lowest bin, there is no competition among treated units that want the same control. The problem in the top quartile of Figure 1(ii) is a little harder: there are $4! = 24$ possible pairings of the 4 treated subjects and 4 controls, and one of these minimizes the total distance within the four pairs. If 30,000 treated people and 60,000 potential controls were divided into 30 bins of sizes 1000 and 2000, then a graph like that in Figure 1(ii) might consist of 30 separate subproblems each with $2 \times 10^6 = 1000 \times 2000$ edges or decision variables, and each subproblem would be of practical size for optimal matching. True, one would need to solve 30 problems, rather than one problem, but each problem could be solved in reasonable time, unlike a graph analogous to Figure 1(i) with $1.8 \times 10^9 = 30,000 \times 60,000$ edges.

Forming bins based on observed covariates is practical and not unreasonable, but it can restrict the possible matches in undesirable ways. This is already visible in the toy illustration in Figure 1(ii). The top bin in Figure 1(ii) has 4 treated units and 4 potential controls, thereby forcing all four controls to be used, leaving open only who is matched to whom. Matching does not reduce bias in the top bin of Figure 1(ii) because all four controls are used.

Moreover, the one control in Figure 1(ii) with the highest propensity score is not close to any treated unit, but as only four controls are available in the top bin, that one control must be included in the match anyway. This is despite the fact that the bottom treated unit in the top bin is very close to a control just barely on the opposite side of the bin boundary, and the second bin has an abundance of potential controls. It would be better to cross the bin boundary and not use the one control with the highest propensity score, but the quartile dividers do not permit this.

The situation can be even worse. If the one control with the highest propensity score had not been in Figure 1(ii), then the top bin would have 4 treated units and 3 controls, so matching all 7 treated units to 7 distinct controls would be impossible with the bin boundaries in Figure 1(ii). Pair matching of all treated units might be feasible in Figure 1(i), but cutting to produce Figure 1(ii) might make pair matching infeasible. Here, the word infeasible is being used in its technical sense: we are optimizing an objective function subject to constraints, but the set of matchings that satisfy the constraints is empty.

Calipers can help, but they must be defined carefully to avoid infeasibility.

If we required a matched control to have an age that differs by at most two years from the age x of its matched treated unit, then we would have imposed a caliper of $x \pm 2$. Cochran and Rubin (1973) discussed caliper matching. Rosenbaum and Rubin (1985b) advocated matching using the Mahalanobis distance within calipers defined by the propensity score. This strategy ensures a close match on the propensity score, but if several such matches are available, it seeks a close match also on other covariates in the Mahalanobis distance.

In general, a caliper is a function $\kappa : \mathbf{R} \longmapsto \mathbf{R}^2$ sending x to $\kappa_1(x) \leq \kappa_2(x)$ where a treated subject with covariate value x may be matched to any control with covariate values in the interval $[\kappa_1(x), \kappa_2(x)]$. The common choice is $x \longmapsto [x - w, x + w]$ for some fixed $w \geq 0$, such as $[x - 2, x + 2]$ for a two-year caliper on age. We could have other choices of $\kappa(\cdot)$, perhaps a very short caliper for very young children, and a longer caliper for people in middle age: perhaps we regard a 1-year old as very different from a 3 year old, but regard a

32 year old as close enough to a 34 year old. If treated subjects are, on average, older than potential controls, then we might prefer an asymmetric caliper, say $x \longmapsto [x-1,\, x+3]$, to offset a tendency of controls to be younger even inside a short caliper.

A caliper would eliminate some edges in Figure 1(i), but unlike Figure 1(ii), a caliper need not produce disconnected components. As the caliper becomes tighter — as we redefine $\kappa_2(x)$ to be closer to $\kappa_1(x)$ — more edges or decision variables are removed, but if we continue too far in this direction, then no pair matching may exist. Narrower calipers accelerate computation but risk infeasibility.

Figures 1(iii) and 1(iv) are obtained from Figure 1(i) by imposing calipers on the propensity score of, respectively, 0.08288 and 0.08293. Although these two calipers both round to 0.0829, Figures 1(iii) and 1(iv) differ in an important way. The caliper 0.08288 is too small: the two treated units with the largest propensity scores in Figure 1(iii) can only be matched to the same single control, so there is no pair matching of distinct individuals. In contrast, the caliper is only a tad larger in Figure 1(iv), but matching is feasible. Define the optimal caliper of the form $\pm w$ as the smallest caliper $w \geq 0$ such that pair matching is feasible. Then the optimal caliper w in Figure 1(i) is in the short interval [0.08288, 0.08293], and the caliper of 0.08293 is feasible.

One new technique in the current paper is a very fast algorithm that finds a short interval, like [0.08288, 0.08293], containing the optimal caliper in a large dense bipartite graph, thereby removing the maximum number of edges that can be removed by a caliper of the form $\pm w$ without generating infeasibility. With many fewer decision variables, this new, sparser graph is then optimized to minimize the total distance within matched pairs, constrained by the optimal caliper and by additional fine balance constraints. The new approach entails an iterative use of a variant of Glover (1967)'s algorithm for matching in a convex bipartite graph. In a doubly convex graph, it is possible to implement Glover's algorithm so that it runs in time proportional to the number of nodes, and this is much faster than the second step of minimum distance matching in either a dense or sparse graph.

Although Figure 1 depicts this technique in terms of a caliper on the propensity score of the form $\pm w$, the same technique has more general applications that we describe.

Figure 1(iv) is attractive compared to Figure 1(ii). The one control whose propensity score is far higher than everyone else is no longer a candidate for matching in Figure 1(iv), whereas its use was mandated in Figure 1(ii). There are no boundaries in Figure 1(iv) that prevent matching individuals who are close, as there were in Figure 1(ii).

Optimal restriction on the number of nearest neighbors inside a caliper.

A limitation of Figure 1(iv) is that some treated units still have many edges or decision variables. This limitation occurs where matching is easy, that is, where the treated and control distributions of the propensity score overlap extensively, and this limitation becomes more of a problem in larger graphs. Any subgraph of Figure 1(iv) maintains the optimal caliper of 0.08293, but not every subgraph would permit a feasible pair match; for instance, Figure 1(iii) is an infeasible subgraph of Figure 1(iv). How could we find a subgraph of Figure 1(iv) so that it discards edges, maintains feasibility, and retains nearest neighbors?

Suppose that we retain at most the ν nearest neighbors of each treated unit in Figure 1(iv). In a minimum caliper graph, like Figure 1(iv), how small can ν be while pair matching remains feasible? It is clear from Figure 1(iii) that $\nu = 1$ is too small, because the two treated units with the highest propensity scores have the same potential control as their $\nu = 1$ nearest neighbor. Figure 1(v) shows that $\nu = 2$ is feasible: a pair match is possible in Figure 1(v).

A second iterative application of Glover's algorithm can determine the minimum feasible ν. That is, the first application of Glover's algorithm determines the optimal caliper, and then the second application determines the smallest feasible ν among subgraphs of the optimal-caliper graph. As seen in Figure 1(v), the treated subject with the largest propensity score has only one neighbor, not $\nu = 2$ neighbors, because the caliper has sensibly eliminated distant controls as neighbors.

Knowing the minimum feasible ν does not require use of this minimal ν. Rather, it informs the investigator that matching in an optimal caliper graph will remain feasible if attention is restricted to at most ν nearest neighbors. For instance, between Figure 1(iv) and Figure 1(v) is a feasible graph satisfying the optimal caliper and with at most $\nu = 3$ nearest neighbors, and this intermediate graph would offer more choice among matched controls, perhaps resulting in a smaller Mahalanobis distance on covariates other than the propensity score, or perhaps with other desired properties such as covariate balance.

With 30,000 treated units and 60,000 potential controls, the dense graph would have $1.8 \times 10^9 = 30,000 \times 60,000$ edges or decision variables. With $\nu = 100$, there would be $3 \times 10^6 = 30,000 \times 100$ edges or decision variables, comparable to a complete bipartite graph that has one twentieth as many nodes or people. With $\nu = 100$, each treated subject would be offered 100 potential controls from which to choose one, so considerations besides the caliper on the propensity score would have substantial influence on the final match.

2.2.2. A second motivating figure incorporating other matching techniques

Best calipers and ν with exact matching for a nominal covariate.

Figure 2 adds two features to the bipartite graphs in Figure 1 that aid in matching. The first feature, exact matching for a nominal covariate, is discussed in the current section, while the second feature, near-fine balance, is discussed in the next section.

Figure 2 removes edges in Figure 1(i) that connect a treated man to a control woman, or a treated woman to a control man, forcing an exact match for gender. An exact match for gender is possible because there are two treated men and eight control men, and five treated women and fifteen control women.

With fewer edges in the initial graph, the optimal caliper on the propensity score is now larger, 0.1925 rather than 0.08293 in Figure 1(iv). Also, the smallest feasible ν within the optimal caliper has risen from $\nu = 2$ in Figure 1(v) to $\nu = 3$ in Figure 2. In Figure 2, no treated individual is connected to a control of the opposite gender, nor to a control differing on the propensity score by more than 0.1925, nor does any treated unit have more than

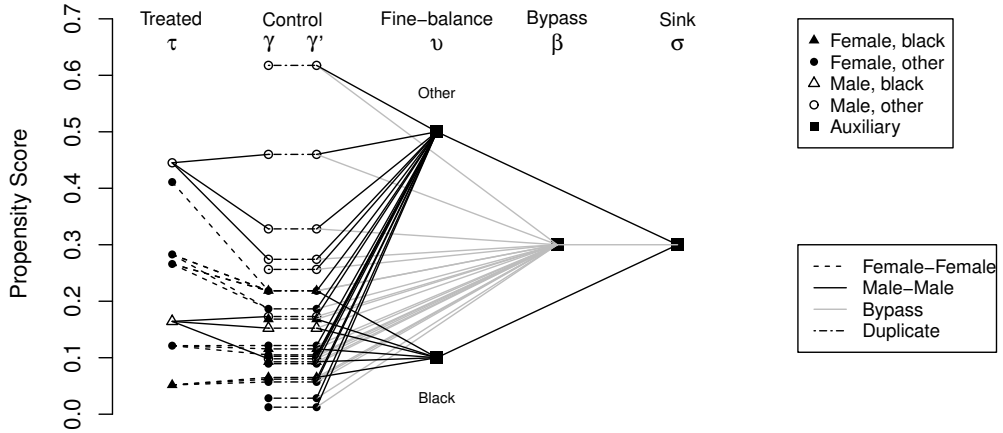


Figure 2: A bipartite graph matching exactly for gender, expanded for near fine balance of race, υ, black or other. The optimal caliper is now 0.1925, and with this caliper the minimum number of neighbors is $\nu = 3$. The duplicate edges connect γ to γ', with capacity 1, so they insist that a control may be matched at most once. The solid grey edges retain feasibility through a penalized bypass, β, of the fine balance constraints.

$\nu = 3$ controls as neighbors. Moreover, the caliper and ν are the best possible: a smaller caliper or ν would make pair matching infeasible.

Near-fine balance for a nominal covariate implemented as a soft constraint.

The structure to the right in Figure 2 will impose near-fine balance for race, black or other, as a soft constraint. The treated group in Figures 1 and 2 contains 2 blacks and 5 others. Fine balance means that the control group will be forced to contain 2 blacks and 5 others, ignoring whether blacks are matched to blacks or to others (Rosenbaum et al., 2007). In other words, fine balance is a constraint on the marginal distribution of race, not a constraint on who is matched to whom. Fine balance is not always feasible. If there were fewer than 2 blacks among the controls or fewer than 5 others among the controls, then fine balance would be infeasible. Figure 2 imposes additional constraints, the exact matching for gender, the caliper on the propensity score, and the limit $\nu = 3$ on the number

of neighbors. Even if the potential controls did include 2 blacks and 5 others, fine balance might be infeasible when conjoined to these other constraints.

Near-fine balance means coming as close as possible to fine balance (Yang et al. 2012). Near-fine balance becomes fine balance whenever fine balance is feasible; otherwise, the fine-balance constraint is minimally relaxed. A simple definition of near-fine balance requires that the total of the absolute differences in frequencies, $|\text{treated} - \text{control}|$, over the categories, black or other, is minimized. For instance, if fine balance is infeasible, the next best frequencies in matched controls would be either 1 black and 6 others or 3 blacks and 4 others, both with the minimal difference of $2 = |2-1| + |5-6| = |2-3| + |5-4|$.

Pimentel et al. (2015a) generalized the concept of near-fine balance, introducing a tree-structured hierarchy of near-fine balance constraints, allowing the user to express a preference for certain kinds of deviations from fine balance over other deviations. The refined balance method of Pimentel et al. (2015a) can be applied in conjunction with the new methods described informally in §2.2.1 and formally in §2.4.2; however, the detailed description of refined balance requires a considerable amount of otherwise unneeded notation, so we formulate the problem in terms of the simpler notion of near-fine balance. For refined balance, the auxiliary structure on the right in Figure 2 has multiple layers of near-fine balance nodes and some additional structure.

With an optimal caliper on the propensity score, the bipartite graph in, say, Figure 1(iv), is feasible but barely so, as seen by comparison with Figure 1(iii), so adding a fine balance constraint may make pair matching infeasible when fine balance would have been feasible in Figure 1(i) without the caliper. Let Υ be the number of possible values of the fine balance covariate; so $\Upsilon = 2$ in Figure 2. We can determine whether fine balance is feasible on its own in Figure 1(i) by constructing a $2 \times \Upsilon$ contingency table recording treatment or control by the fine balance covariate; however, no simple tabulation shows whether fine balance is feasible with a caliper, as in Figure 1(iv). As a consequence, we implement near-fine balance as a soft constraint in Figure 2. A soft constraint is implemented by altering the objective

function — here, the total covariate distance within matched pairs — so that it penalizes violations of fine balance. This is discussed in greater detail in §2.4, where the grey edges in Figure 2 will permit penalized violations of fine balance. The minimum cost flow algorithm tries to avoid penalized violations of fine balance, but tolerates the minimum number of violations needed to produce feasibility. The implementation of near-fine balance as a soft constraint departs from the hard constraint used by Yang et al. (2012) and is a variant of the soft constraint used by Pimentel et al. (2015a). A soft constraint is necessary here because the bipartite graph is thinned by calipers and near neighbors.

Implementation details with substantial consequences for performance.

When thinking about the computational effort required for optimal matching, attention usually focuses on the well-studied speed of the optimization itself. We did some timing exercises for large optimal matching problems, discovering that a substantial fraction of the time was spent setting up the optimization problem, rather than solving it. Specifically, much of the time was spent computing the robust Mahalanobis distances that label edges with the cost of pairing individuals. We reduced this time in two ways. First, by removing most edges as in Figure 1(v), we greatly reduced the number of Mahalanobis distances that need to be computed. Second, the Mahalanobis distance is a quadratic form, so the most straightforward form of computation involves $O\left(P^2\right)$ arithmetic operations for P covariates. This can be reduced to $O\left(P\right)$ computations per distance through a Cholesky decomposition. Although these are simple changes, they have a big effect on the speed of computations, an effect that falls outside of formal calculation of the time required to solve a minimum cost flow problem.

If we seek a single optimal caliper in the presence of $\Xi \geq 2$ exact match categories, such as the $\Xi = 2$ genders in Figure 2, then the caliper must be feasible within every category, so the optimal caliper is the maximum of Ξ optimal calipers for the categories one at a time. In Figure 2, we may find the optimal caliper separately for women and for men. Each caliper is found using a binary search, so it starts with an interval of feasible calipers

and cuts the interval in half repeatedly. Suppose that we find the caliper for women first. If the caliper we found for women is feasible for men, then we can stop, because the best caliper overall must be greater than or equal to the caliper for women alone. Because the ratio of potential controls to treated is $15/5 = 3$ for women in Figure 2, but it is $8/2 = 4$ for men, it makes sense to find the caliper for women first, guessing that the caliper will be larger for women, hoping therefore to avoid a search for the optimal caliper for men. The same considerations apply when optimizing the number ν of near neighbors, rather than the caliper. This shortcut matters more in the example in §2.5 where $\Xi = 463$ principal procedures are exactly matched.

2.3. Practical Aspects of Matching in Large Databases

Section 2.4 discusses a network structure and a few results that permit matching in large databases, and these ideas are implemented in the `R` package `bigmatch`. One can make effective use of the `bigmatch` package without reading §2.4, albeit with incomplete knowledge of precisely what the package is doing. The current section is intended to assist a reader who wants to get started immediately. In the `bigmatch` package, the examples for the `nfmatch` function go through all the needed steps in the small data set `nh0506` with 2475 people mentioned at the beginning of §2.2.

Essentially, the `bigmatch` package does two things. First, it creates a sparser but nonetheless feasible graph for matching. Returning to the tiny illustration in Figure 1, `bigmatch` starts by producing a graph like Figure 1(v) rather than like Figure 1(i); however, the actual graph is vastly larger in every sense than Figure 1(v). Although `bigmatch` "removes" most of the edges, it is careful to ensure that matching is still possible; it avoids removing too many edges. Call this the first step.

Second, in a graph like Figure 1(v), the `bigmatch` package offers a suite of standard techniques for optimal matching in observational studies, such as propensity score calipers, near-fine balance, minimizing a robust covariate distance, exact matching, near-exact (or almost-exact) matching. For discussion of these standard methods, see Rosenbaum (2010,

Part II). From the user's point of view, these standard methods work in their standard way. Inside `bigmatch`, there are various nonstandard implementations, essentially to avoid computing or storing information in Figure 1(i) that plays no role in Figure 1(v). Call this the second step. If aspects of the second step are unfamiliar, then try them out using the `nh0506` data in the `bigmatch` package.

The first step has two tasks: (i) pick a caliper on the propensity score (or some other score) yielding Figure 1(iv); then (ii) pick a limit ν on the number of near neighbors, moving from Figure 1(iv) to Figure 1(v). The `optcal` function in the `bigmatch` package does task (i): it uses Glover's algorithm iteratively to find the smallest feasible caliper on the propensity score while also respecting any requirements you have set for exact matching. It returns this caliper to you as a number. Also returned is an interval showing the precision with which the caliper was determined. You need not use this caliper – you may use a larger one – but if you use a smaller caliper, then no pair matching exists that will satisfy the smaller caliper. The `optconstant` function takes a caliper you specify – perhaps the optimal caliper just determined or perhaps a larger one – and determines the minimum feasible value of ν, the upper limit on the number of near neighbors. You need not use this minimum feasible ν – you may use a larger one – but if you use a smaller ν, then no pair matching exists that will satisfy it. You now know the lower limits on the caliper and ν, and you are ready for step two.

In step two, you give to `nfmatch` a caliper and ν that are at least as large as the minimums determined in step one, and you specify your other matching requirements, and it computes the optimal match subject to your specifications.

The minimum feasible caliper and ν yield a sparse graph, and perhaps the fastest computation in step two. However, speed is one consideration among others. Setting the caliper and ν to be higher than their minimum feasible values gives `nfmatch` more latitude in searching for a close, balanced match, perhaps producing a better match in terms of covariate balance. It is reasonable to construct and compare a few matched samples, picking the

most satisfactory one providing, of course, that you do not look at outcomes until after that decision is made and the study's design is finalized and fixed.

2.4. Network Structure: A Sparse Bipartite Graph Expanded for Near-fine Balance

2.4.1. The matching problem in a bipartite graph $\mathcal{B}$

There are T treated units, $\mathcal{T} = \{\tau_1, \ldots, \tau_T\}$, and C potential controls, $\mathcal{C} = \{\gamma_1, \ldots, \gamma_C\}$, where $\mathcal{T} \cap \mathcal{C} = \emptyset$, and $T \leq C$. Write $|\mathcal{S}|$ for the number of elements of a finite set $\mathcal{S}$, so that, for instance, $|\mathcal{T}| = T$. We would like to match every treated unit $\tau \in \mathcal{T}$ to a different control $\gamma \in \mathcal{C}$; that is, each potential control is used at most once. A match is a 1-to-1 function, $\mu : \mathcal{T} \to \mathcal{C}$, so that $\gamma = \mu(\tau) \in \mathcal{C}$ is matched to $\tau \in \mathcal{T}$, and $\mu(\tau) \neq \mu(\tau')$ whenever $\tau \neq \tau'$. Write $\mathcal{M}$ for the set of matched controls, $\mathcal{M} = \{\mu(\tau) : \tau \in \mathcal{T}\} \subseteq \mathcal{C}$ with $|\mathcal{M}| = T$. The small changes required for matching with multiple controls are discussed separately in §2.4.6. In Figures 1 and 2, $T = 7$ and $C = 23$, so $\mu(\cdot)$ will pair the seven treated units to seven different controls.

An edge $e = (\tau, \gamma)$ with $\tau \in \mathcal{T}$ and $\gamma \in \mathcal{C}$ is a possible pairing of treated subject τ with control γ, and we must decide whether we want this pairing in the matched sample, or a different one. The dense bipartite graph analogous to Figure 1(i) has nodes $\mathcal{T} \cup \mathcal{C}$ and every possible pairing: the edges $\mathcal{B}$ of this dense graph consist of all $T \times C$ ordered pairs (τ, γ) with $\tau \in \mathcal{T}$ and $\gamma \in \mathcal{C}$; that is, $\mathcal{B}$ is the direct product, $\mathcal{B} = \mathcal{T} \times \mathcal{C}$ so $|\mathcal{B}| = T \times C$. In Figure 1(i), there are $|\mathcal{B}| = T \times C = 7 \times 23 = 161$ potential pairs, from which we will select seven edges with different controls, so $|\mathcal{M}| = T = 7$. The sparse bipartite in Figure 1(v) has the same nodes, $\mathcal{T} \cup \mathcal{C}$, but the set of edges, $\mathcal{B} \subset \mathcal{T} \times \mathcal{C}$, is much smaller.

There is a real valued score, $\rho : \mathcal{T} \cup \mathcal{C} \to \mathbf{R}$, and we would like $|\rho(\tau) - \rho(\gamma)|$ to be small if $\tau \in \mathcal{T}$ is matched to $\gamma \in \mathcal{C}$. Commonly, $\rho(\cdot)$ is either the propensity score computed from observed covariates, as in Figure 1, or a transformation of the propensity score such as its logit or its rank. Additionally, there is a nonnegative distance $\delta : \mathcal{T} \times \mathcal{C} \to [0, \infty)$, and we would like $\delta(\tau, \gamma)$ to be small if $\tau \in \mathcal{T}$ is matched to $\gamma \in \mathcal{C}$. Commonly, $\delta(\tau, \gamma)$ is a robust Mahalanobis distance computed from observed covariates, perhaps with penalties to

enforce additional constraints (Rosenbaum, 2010, §8-9). Usually, we give some priority to $\rho(\cdot)$, because a close match on the true propensity score can, by itself, balance all observed covariates, but if many controls $\gamma \in \mathcal{C}$ are close to $\tau \in \mathcal{T}$ in terms of $\rho(\cdot)$, then it makes sense to seek a control who is also close on key covariates as measured by $\delta(\tau, \gamma)$; see Rosenbaum and Rubin (1985b).

There are two nominal covariates, ξ with $\Xi \geq 1$ nominal levels, $1, \ldots, \Xi$, and υ with $\Upsilon \geq 1$ nominal levels, $1, \ldots, \Upsilon$. Nominal covariate ξ will be matched exactly, while nominal covariate υ will be nearly finely balanced. In Figure 2, ξ was gender, female or male, and υ was race, black or other. To avoid silly cases, we assume $\Xi \leq C$ and $\Upsilon \leq C$, but typically Ξ and Υ are much smaller than C. In Figure 2 the values are $\Xi = \Upsilon = 2$, but in §2.5 they are $\Xi = 463$ and $\Upsilon = 973$. Each individual in $\mathcal{T} \cup \mathcal{C}$ has a value of $\xi(\cdot)$ and a value of $\upsilon(\cdot)$; that is, $\xi : \mathcal{T} \cup \mathcal{C} \to \{1, \ldots, \Xi\}$ and $\upsilon : \mathcal{T} \cup \mathcal{C} \to \{1, \ldots, \Upsilon\}$. In practice, either $\xi(\cdot)$ or $\upsilon(\cdot)$ might not be present, but it is notationally and algorithmically convenient to view this as a special case, not a new case. Specifically, if $\xi(\cdot)$ has only one level, $\Xi = 1$, then for all practical purposes there is no exact-match covariate. If $\upsilon(\cdot)$ has only one level, $\Upsilon = 1$, then for all practical purposes there is no near-fine-balance covariate.

How does the exact match variable enter the structure of the bipartite graph? We only consider pairs that are exactly matched for $\xi(\cdot)$ but we may not consider all such pairs. Saying the same thing precisely, every edge $e = (\tau, \gamma) \in \mathcal{B}$ in the graph will have $\xi(\tau) = \xi(\gamma)$, but $\xi(\tau) = \xi(\gamma)$ does not ensure that $e = (\tau, \gamma) \in \mathcal{B}$. If $\Xi = 1$, then the exact match covariate does not restrict the graph. In Figure 2, there is no edge connecting a man to a woman.

Definition 1 *Pair matching is feasible in $\mathcal{B}$ if there exists a 1-1 function $\mu : \mathcal{T} \to \mathcal{C}$ with $\{\tau, \mu(\tau)\} \in \mathcal{B}$ for $\tau \in \mathcal{T}$.*

What is fine balance? Fine balance means that $\upsilon(\tau) = k$ occurs in the treated group with

the same frequency that $\upsilon(\gamma) = k$ occurs in the matched control group $\mathcal{M}$,

$$|\{\tau \in \mathcal{T} : \upsilon(\tau) = k\}| = |\{\gamma \in \mathcal{M} : \upsilon(\gamma) = k\}|, \text{ for } k = 1, \ldots, \Upsilon. \tag{2.1}$$

Notice that (2.1) is a property of the match $\mathcal{M}$ as a whole, not a property of individual pairs; that is, a property of $\mathcal{M}$ but not $\mu(\cdot)$. Condition (2.1) could hold, yet many or all matched pairs $\{\tau, \mu(\tau)\} \in \mathcal{B}$ may have $\upsilon(\tau) \neq \upsilon\{\mu(\tau)\}$. In Figure 2, blacks need not be paired with blacks, but we would like the number of blacks to be the same in the treated and control groups. Sometimes, condition (2.1) is not feasible; it cannot be done. Write

$$d_k = |\{\tau \in \mathcal{T} : \upsilon(\tau) = k\}| - |\{\gamma \in \mathcal{M} : \upsilon(\gamma) = k\}|, \tag{2.2}$$

so $d_k = 0$ for $k = 1, \ldots, \Upsilon$ when (2.1) holds. Near-fine balance means that we minimize $\sum_{k=1}^{\Upsilon} |d_k|$ when fine balance (2.1) is infeasible.

Proposition 2 *If $\mu : \mathcal{T} \to \mathcal{C}$ is a feasible pair match in $\mathcal{B}$ with matched controls $\mathcal{M} = \{\mu(\tau) : \tau \in \mathcal{T}\} \subseteq \mathcal{C}$, then the deviations d_k from fine balance satisfy*

$$0 = \sum_{k=1}^{\Upsilon} d_k \quad \textit{and} \quad \sum_{k=1}^{\Upsilon} |d_k| = 2 \sum_{k=1}^{\Upsilon} \max(0, d_k).$$

The proof of Proposition 2 can be found in Appendix A.1.

2.4.2. Glover's algorithm used iteratively to determine an optimal caliper and near neighbors

For a bipartite graph $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$ with nodes $\mathcal{T} \cup \mathcal{C}$ and edges $\mathcal{B} \subseteq \mathcal{T} \times \mathcal{C}$, the neighborhood $\phi(\tau) \subseteq \mathcal{C}$ of $\tau \in \mathcal{T}$ is $\phi(\tau) = \{\gamma \in \mathcal{C} : (\tau, \gamma) \in \mathcal{B}\}$. The bipartite graph $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$ is said to be convex if it is possible to number or order the elements $\gamma_1, \ldots, \gamma_C$ of $\mathcal{C}$ so that $\gamma_i \in \phi(\tau)$ and $\gamma_j \in \phi(\tau)$ with $i < j$ implies $\gamma_{i+1} \in \phi(\tau)$, ..., $\gamma_{j-1} \in \phi(\tau)$. A convex bipartite graph $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$ is said to be doubly convex if it is also convex with the roles of $\mathcal{T}$ and $\mathcal{C}$ reversed.

In §2.4.1, sort the nodes of $\mathcal{T}$ first by the nominal variable $\xi(\tau)$ and, within levels of $\xi(\cdot)$, by the score $\rho(\tau)$. Use the parallel procedure to sort the nodes of $\mathcal{C}$. If we form $\mathcal{B}$ by

including (τ, γ) in $\mathcal{B}$ if and only if $\xi(\tau) = \xi(\gamma)$ and $|\rho(\tau) - \rho(\gamma)| \leq \varkappa$ for a fixed number $\varkappa > 0$, then the graph is doubly convex. We will determine the smallest $\varkappa$ such that pair matching is feasible in $\mathcal{B}$.

Given a convex or doubly convex bipartite graph $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$, Glover (1967) proposed an algorithm that determines whether pair matching is feasible in $\mathcal{B}$ in the sense of Definition 1. Actually, Glover's algorithm does this and more, but this is all we need. For our purposes, Glover's algorithm takes as input $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$ and returns a 1 if pair matching is feasible or a 0 if it is not.

Lipski and Preparata (1981) used a doubly-ended queue to obtain a fast implementation of Glover's algorithm in a doubly convex bipartite graph, with running time $O(T + C)$, so it is linear in the number of nodes. It takes longer to sort the nodes by $\xi(\cdot)$ and $\rho(\cdot)$ than it does to execute this version of Glover's algorithm. In the current problem, the sort needs to be done once, but Glover's algorithm will be used repeatedly. Of greater importance, both sorting and Glover's algorithm are much faster than solving the minimum cost flow problem to produce an optimal match with near-fine balance, so the time spent determining the optimal caliper is negligible by comparison.

We determine the optimal caliper $\varkappa$ by binary search. Set $\varkappa_{\min} = 0$ and $\varkappa_{\max} = \max_{\iota \in \mathcal{T} \cup \mathcal{C}} \rho(\iota) - \min_{\iota \in \mathcal{T} \cup \mathcal{C}} \rho(\iota)$, and pick an $\epsilon > 0$. Let $\text{glover}(\varkappa) = 1$ if pair matching is feasible in the bipartite graph $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$ with exact variable $\xi(\cdot)$ and caliper $\varkappa$ on the score $\rho(\cdot)$; otherwise, let $\text{glover}(\varkappa) = 0$. If $\text{glover}(0) = 1$, stop; pair matching is feasible with caliper $\varkappa = 0$. If $\text{glover}(\varkappa_{\max}) = 0$, stop; pair matching is infeasible for every value of $\varkappa$. Otherwise:

1. If $\varkappa_{\max} - \varkappa_{\min} < \epsilon$, stop and use caliper $\varkappa_{\max}$, which is feasible and within ϵ of the optimal caliper.

2. Otherwise, define $\overline{\varkappa} = (\varkappa_{\max} + \varkappa_{\min})/2$. If $\text{glover}(\overline{\varkappa}) = 1$, set $\varkappa_{\max} \leftarrow \overline{\varkappa}$ and go to step 1, but if $\text{glover}(\overline{\varkappa}) = 0$, set $\varkappa_{\min} \leftarrow \overline{\varkappa}$ and go to step 1.

In Figure 1, with $[\varkappa_{\min}, \varkappa_{\max}] = [0.08288, 0.08293]$, pair matching was infeasible with caliper 0.08288 in Figure 1(iii), but it was feasible with caliper 0.08293. If $\rho(\iota)$ is a probability, such as the propensity score, then $\varkappa_{\max} \leq 1$, and the interval $[\varkappa_{\min}, \varkappa_{\max}]$ has length at most 2^{-I} after I iterations of step 2. For instance, after $I = 7$ iterations, the the interval $[\varkappa_{\min}, \varkappa_{\max}]$ has length at most $2^{-7} = 0.0078125 < 0.01$.

Now, with $\varkappa$ in hand, consider restricting the number ν of neighbors, as in Figure 1(v). For each fixed $\tau \in \mathcal{T}$, sort $|\rho(\tau) - \rho(\gamma)|$ into increasing order, and define $o_v(\tau)$ to be the νth of the C sorted values of $|\rho(\tau) - \rho(\gamma)|$. Define the bipartite graph $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$ where (τ, γ) is in $\mathcal{B}$ if and only if $\xi(\tau) = \xi(\gamma)$, $|\rho(\tau) - \rho(\gamma)| \leq \min\{\varkappa, o_v(\tau)\}$. Having found and fixed the optimal caliper, $\varkappa$, as above, we may determine the minimum feasible value of ν by a second iterative application of Glover's algorithm.

In Figure 1, with optimal caliper $\varkappa = 0.08293$, the minimum feasible number of near neighbors is $\nu = 2$ in Figure 1(v). Exact matching for gender in Figure 2 increased this to $\varkappa = 0.1925$ and $\nu = 3$.

There are many minor but useful variations on this theme. A single outlier among the scores, $\rho(\tau)$, $\tau \in \mathcal{T}$, may result in a large optimal caliper, $\varkappa$; however, this possibility is avoided if the scores are replaced by their ranks. In Figure 2, we computed a single optimal caliper for use with both women, $\xi(\cdot) = 1$, and men, $\xi(\cdot) = 2$; however, one could determine a different optimal caliper for each exact group, thereby reducing either the caliper for men or the caliper for women. It is useful to know, and easy to determine, the minimum feasible number of near neighbors, ν; however, once this is known, one might decide to include in $\mathcal{B}$ a larger number of near neighbors, say 2ν, in the hope of obtaining a smaller deviation from fine balance, $\sum_{k=1}^{\Upsilon} |d_k|$, or a smaller total covariate distance, $\sum_{\tau \in \mathcal{T}} \delta\{\tau, \mu(\tau)\}$, when the final match is constructed in §2.4.4.

2.4.3. Added structure imposing near-fine balance as a soft constraint

How does the near-fine balance variable $\upsilon(\cdot)$ change the structure of the graph? We make a distinct copy γ' of each control γ, collecting these in a set $\mathcal{C}' = \left\{\gamma_1', \ldots, \gamma_C'\right\}$, adding the

γ' to the set of nodes and adding C edges $\left(\gamma_j, \gamma_j^{'}\right)$ from each control to its copy, placing these C duplicate edges in a set $\mathcal{O}$. Writing γ' signifies the one duplicate corresponding with a specific $\gamma \in \mathcal{C}$; it does not signify a generic member of $\mathcal{C}'$. Of course, the duplicate belongs to the same category of the fine balance variable, $\upsilon(\gamma) = \upsilon(\gamma')$. We add new nodes $1, \ldots, \Upsilon$, one for each category of $\upsilon(\cdot)$ and an edge connecting each copy $\gamma' \in \mathcal{C}'$ to the one category $\upsilon(\gamma')$ that contains it, that is, the edge $\{\gamma', \upsilon(\gamma')\}$. To implement near-fine balance as a soft constraint, we allow some controls $\gamma' \in \mathcal{C}'$ to bypass their category $\upsilon(\gamma')$ by introducing a new bypass node β and an edge (γ', β) from each control $\gamma' \in \mathcal{C}'$ to β. Finally, we introduce another node, σ, called a sink, an edge (β, σ) from the bypass node to the sink, and an edge (υ, σ) from each fine-balance category $\upsilon \in \{1, \ldots, \Upsilon\}$ to the sink σ. In Figure 2, there are $\Upsilon = 2$ near-fine balance categories, black and other, whereas the grey edges bypass these two categories and reach the sink by a different route.

In the end, there is a network similar to Figure 2 with nodes $\mathcal{N}$ and directed edges $\mathcal{E}$ given by:

$$\begin{aligned} \mathcal{N} &= \mathcal{T} \cup \mathcal{C} \cup \mathcal{C}' \cup \{1, \ldots, \Upsilon\} \cup \{\beta, \sigma\} \\ \mathcal{E} &= \mathcal{B} \cup \mathcal{O} \cup \left\{\left(\gamma', \upsilon\left(\gamma'\right)\right) : \gamma' \in \mathcal{C}'\right\} \cup \{(\upsilon, \sigma) : \upsilon \in \{1, \ldots, \Upsilon\}\} \\ &\quad \cup \left\{\left(\gamma', \beta\right) : \gamma' \in \mathcal{C}'\right\} \cup \{(\beta, \sigma)\}. \end{aligned} \tag{2.3}$$

We refer to $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$ as the bipartite graph, and to $(\mathcal{N}, \mathcal{E})$ as the matching graph. Both are directed graphs. In §2.4.4, capacities, costs and divergences are added to $(\mathcal{N}, \mathcal{E})$, and with these added structures we speak of the matching network. A key element is that we will construct $\mathcal{B}$ so that it is fairly sparse, yet pair-matching will be feasible in $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$. Then we impose near-fine balance with a soft constraint in $(\mathcal{N}, \mathcal{E})$ so that whenever pair matching is feasible in $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$, it remains feasible with near-fine balance in the network $(\mathcal{N}, \mathcal{E})$. In this way, as we make $\mathcal{B}$ sparse, we do not lose feasibility.

2.4.4. Optimal matching by minimum cost flow in a network

In the minimum cost flow problem, each edge $e \in \mathcal{E}$ has a finite nonnegative integer capacity, $\text{cap}(e) \geq 0$, and a nonnegative real valued cost, $\text{cost}(e) \geq 0$. Edge e can transport $\text{cap}(e)$ units at a cost per unit of $\text{cost}(e)$. Each node $n \in \mathcal{N}$ has an integer divergence, $\text{div}(n)$. A feasible flow $f(\cdot)$ is an nonnegative integer-valued function of the edges, $f : \mathcal{E} \to \{0, 1, 2, \ldots\}$, that respects the capacities for each $e \in \mathcal{E}$,

$$0 \leq f(e) \leq \text{cap}(e) \text{ with } f(e) \in \{0, 1, 2, \ldots\} \text{ for each } e \in \mathcal{E} \tag{2.4}$$

and the divergences for each node $n \in \mathcal{N}$,

$$\text{div}(n) = \sum_{n'':(n,n'')\in\mathcal{E}} f\{(n, n'')\} - \sum_{n':(n',n)\in\mathcal{E}} f\{(n', n)\} \text{ for each } n \in \mathcal{N}. \tag{2.5}$$

A positive divergence, $\text{div}(n) > 0$, means that node n supplies $\text{div}(n)$ units of flow, while a negative divergence, $\text{div}(n) < 0$, means that node n absorbs $-\text{div}(n)$ units of flow. If $\text{div}(n) = 0$, then node n passes along all the flow it receives from other units. A feasible flow may or may not exist. The cost of a feasible flow is the total cost of the flow over the edges, $\text{cost}(f) = \sum_{e\in\mathcal{E}} f(e)\,\text{cost}(e)$. The minimum cost flow problem is to find a feasible flow of minimum cost or determine that no feasible flow exists. Attractive textbook discussion of minimum cost flow problems is given by Bertsekas (1998) and Korte and Vygen (2012).

In the network (2.3) or in Figure 2, set:

$$\begin{aligned} &\text{div}(\tau) = 1 \text{ for } \tau \in \mathcal{T}, \\ &\text{div}(\sigma) = -T, \\ &\text{div}(n) = 0 \text{ for } n \notin \mathcal{T} \cup \{\sigma\} \end{aligned} \tag{2.6}$$

so one unit of flow emanates from each of the T treated units $\tau \in \mathcal{T}$, all T units of flow are absorbed by the sink, σ, and all other nodes pass along the all of the flow that they receive.

Also, set

$$\begin{aligned}
\text{cap}\left\{(\tau,\gamma)\right\} &= 1 \text{ for } (\tau,\gamma) \in \mathcal{B}, \\
\text{cap}\left\{\left(\gamma,\gamma'\right)\right\} &= 1 \text{ for } \left(\gamma,\gamma'\right) \in \mathcal{O}, \\
\text{cap}\left\{\left(\gamma',\upsilon\left(\gamma'\right)\right)\right\} &= 1 \text{ for } \gamma' \in \mathcal{C}', \\
\text{cap}\left\{\left(\gamma',\beta\right)\right\} &= 1 \text{ for } \gamma' \in \mathcal{C}', \\
\text{cap}\left\{(k,\sigma)\right\} &= \left|\left\{\tau \in \mathcal{T} : \upsilon\left(\tau\right) = k\right\}\right| \text{ for } k \in \left\{1,\ldots,\Upsilon\right\}, \\
\text{cap}\left\{(\beta,\sigma)\right\} &= T.
\end{aligned} \tag{2.7}$$

Combining (2.4), (2.5), (2.6), and (2.7) says the following about a feasible flow, $f\left(\cdot\right)$. No control $\gamma \in \mathcal{C}$ can receive more than one unit of flow, because it must transfer all its flow to γ', and $\text{cap}\left\{(\gamma,\gamma')\right\} = 1$. Because $f\left(\cdot\right)$ takes on nonnegative integer values, $f\left\{(\tau,\gamma)\right\} = 1$ for at most T nonoverlapping pairs, (τ,γ). The pair match will be defined by $\mu\left(\tau\right) = \gamma$ if and only if $f\left\{(\tau,\gamma)\right\} = 1$.

Select a large positive number, $\Psi > 0$, as a penalty, and define the costs as follows:

$$\begin{aligned}
\text{cost}\left(e\right) &= \delta\left(\tau,\gamma\right) \geq 0 \text{ for } e = (\tau,\gamma) \in \mathcal{B}, \\
\text{cost}\left(e\right) &= \Psi > 0 \text{ for } e = \left(\beta,\sigma\right), \\
\text{cost}\left(e\right) &= 0 \text{ for } e \notin \mathcal{B} \cup \left\{(\beta,\sigma)\right\}.
\end{aligned} \tag{2.8}$$

Definition 3 *The matching network refers to nodes $\mathcal{N}$ and directed edges $\mathcal{E}$ given by (2.3), divergences given by (2.6), capacities given by (2.7) and costs given by (2.8). A flow in the matching network is feasible if it satisfies (2.4), (2.5), (2.6), and (2.7). A minimum cost flow is a feasible flow that minimizes* $\text{cost}\left(f\right) = \sum_{e\in\mathcal{E}} f\left(e\right) \text{cost}\left(e\right)$ *among feasible flows.*

Proposition 4 *If pair matching is feasible in $\mathcal{B}$, then there exists at least one feasible flow in the matching network $(\mathcal{N}, \mathcal{E})$. Conversely, every feasible flow $f\left(\cdot\right)$ in the matching network $(\mathcal{N}, \mathcal{E})$ defines a feasible pair matching in $\mathcal{B}$ as follows: $\mu\left(\tau\right) = \gamma$ if and only if*

$f\{(\tau,\gamma)\}=1$.

Recall the definition of the deviation d_k from fine balance in (2.2). Proposition 5 says that we obtain from a minimum cost flow the closest match in terms of covariate distance $\sum_{\tau\in\mathcal{T}}\delta\{\tau,\mu(\tau)\}$ among all matches that minimally deviate from fine balance, as measured by $\sum_{k=1}^{K}|d_k|$. That is, the soft constraint imposed using Ψ in the costs (2.8) has prioritized the considerations represented by $\mathcal{B}$ and it has avoided infeasibility. Proposition 5 provides a needed extension of related existing results. Specifically, Yang et al. (2012) imposed near-fine balance with a hard constraint that can create infeasibility if the bipartite graph $(\mathcal{T}\cup\mathcal{C},\mathcal{B})$ is not complete, that is, not like Figure 1(i). As here, Pimentel et al. (2015a) used a soft constraint, but the structure of the network $(\mathcal{N},\mathcal{E})$ is somewhat different.

Proposition 5 *If $\Psi>\sum_{(\tau,\gamma)\in\mathcal{B}}\delta(\tau,\gamma)$, then every minimum cost feasible flow $f(\cdot)$ in the network $(\mathcal{N},\mathcal{E})$ yields a pair match $\mu(\tau)=\gamma$ in $\mathcal{B}$ that minimizes the deviation from fine balance; that is, it minimizes $\sum_{k=1}^{\Upsilon}|d_k|$. Moreover, among pair matches in $\mathcal{B}$ that minimize $\sum_{k=1}^{\Upsilon}|d_k|$, a match obtained from a minimum cost feasible flow minimizes $\sum_{\tau\in\mathcal{T}}\delta\{\tau,\mu(\tau)\}$.*

2.4.5. Computational effort

Consider a growing sequence of ever larger feasible matching networks $(\mathcal{N},\mathcal{E})$ each of the form given by Definition 3. Each of these networks uses some feasible caliper $\varkappa$ and some feasible number ν of near neighbors inside that specific caliper, $\varkappa$. It is not assumed that minimal feasible values of $\varkappa$ and ν are used. Our sequence of networks $(\mathcal{N},\mathcal{E})$ has a corresponding sequence of feasible $\varkappa$'s and ν's. The second sentence of Proposition 6 entertains the possibility that our growing sequence of networks has a single uniform bound $\overline{\nu}$ on ν, $\nu\leq\overline{\nu}$. Recall that $C\geq T$ is the number of potential controls.

Proposition 6 *The time required to find the minimum cost flow in Proposition 5 is bounded by $O\left\{\nu C^2+C^2\log(C)\right\}$. In particular, if ν is uniformly bounded, $\nu\leq\overline{\nu}$, then the time required is bounded by $O\left\{C^2\log(C)\right\}$.*

In contrast, if $(\mathcal{T} \cup \mathcal{C},\, \mathcal{B})$ were complete with $\mathcal{B} = \mathcal{T} \times \mathcal{C}$, as in Figure 1(i), Korte and Vygen (2012, Theorem 9.13) gives a bound of $O\left\{|\mathcal{N}| \cdot |\mathcal{E}| + |\mathcal{N}|^2 \cdot \log(|\mathcal{N}|)\right\} = O\left(C^3\right)$, a much larger bound than $O\left\{C^2 \log(C)\right\}$. Proposition 6 indicates that even with growing values of ν we have a time bound of $O\left\{C^2 \log(C)\right\}$ providing $\nu = O\{\log(C)\}$.

In `R`, minimum cost flow problems may be solved using the Fortran code Relax IV of Bertsekas and Tseng (1988), which implements the auction algorithm of Bertsekas (1981). The Fortran code is included in Hansen's `optmatch` package, and an `R` function, `callrelax`, for calling the Fortran code, is included in Pimentel (2016)'s `rcbalance` package. The `bigmatch` package associated with the current paper uses Relax IV. Strictly speaking, the time bound in Proposition 6 is not applicable with the auction algorithm, but the work of Bertsekas and Tseng suggests its performance is competitive. The time bound is attained using the algorithm in Korte and Vygen (2012), Theorem 9.13.

2.4.6. Extension to multiple controls

A conceptually simple way to match with two controls is to duplicate each treated subject, so $\mathcal{T} = \{\tau_1, \ldots, \tau_T\}$ is replaced by $\mathcal{T}^* = \left\{\tau_1, \tau_1^{'}, \ldots, \tau_T, \tau_T^{'}\right\}$, so $|\mathcal{T}^*| = 2T$, and then apply the method for pair matching described earlier. Because Glover's algorithm in §2.4.2 is so fast when applied to a doubly convex bipartite graph, there seems to be little harm in using it in this way to determine the optimal $\varkappa$ and ν. To match with $\omega \geq 2$ controls, $\mathcal{T}$ can be duplicated ω times. Duplication is unwise when solving the minimum cost flow problem because it entails storing the same edge several times, and instead one should set $\mathrm{div}(\tau_t) = \omega$ for $t = 1, \ldots, T$ in (2.6) to match with $\omega \geq 2$ controls. This is implemented in the `R` package `bigmatch`.

2.5. Constructing the Matched Sample in the Medicaid Example

2.5.1. Finding the minimal caliper $\varkappa$ and number of neighbors ν

In the Medicaid example in §2.1.2, the first step is to construct the bipartite graph $(\mathcal{T} \cup \mathcal{C},\, \mathcal{B})$ analogous to Figure 1(v), but with $|\mathcal{T}| = T = 38,841$ treated children and $|\mathcal{C}| = C = 159,527$ potential controls. Here, there are roughly $C/T = 4.1$ potential controls for each treated child. The complete bipartite graph analogous to Figure 1(i) is far too large: each

treated child has $C = 159,527$ potential controls, making $T \times C = 38,841 \times 159,527 = 6.20 \times 10^9$ edges in $\mathcal{B}$. However, in the graph analogous to Figure 1(v), each treated child has at most $\nu = 105$ potential controls, with $3.84 \times 10^6 \leq \nu T = 105 \times 38,841 = 4.078 \times 10^6$ edges in $\mathcal{B}$. It took 6.1 minutes to determine the optimal caliper on the rank of the propensity score, then an additional 1.4 minutes to determine the minimum feasible number of neighbors, $\nu = 105$. The best match in the network analogous to Figure 2 was found by solving a single minimum cost flow problem in an additional 32.5 minutes.

In the graph analogous to Figure 1(v), $\mathcal{B}$ contained fewer than $\nu T = 4.078 \times 10^6$ edges. How does that compare to a divided graph analogous to Figure 1(ii)? A complete bipartite graph with $T = 1000$ and $C = 4000$ would have 4×10^6 edges, so if the problem with $T = 38,841$ and $C = 159,527$ were split into 40 subproblems of size roughly $T = 1000$ and $C = 4000$, then each of the 40 subproblems would have about the same number of edges, about 4 million, as the one problem using the $\mathcal{B}$ that we construct. Most importantly, each of the 40 subproblems would separately use fine balance, so fine balance would be more constrained and would accomplish much less. For instance, fine balance can do nothing in the top stratum of Figure 1(ii), because all four controls must be used.

Figure 3 depicts aspects of the construction of $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$. Figure 3(i) shows the distribution of the estimated propensity score before matching, based on a logit regression of the binary indicator of childrens-versus-adult hospital on the covariates in Table 1 and the 463 categories of Principal Procedures. For the reason mentioned in §2.4.2, the caliper is defined in terms of the rank of the propensity score in Figure 3(ii); however, the propensity score itself could have been used.

We wanted to match exactly for the 463 surgical procedures, so we sorted the data first by procedure, then by the propensity score (or equivalently by its rank). We then used Glover's method to find a single optimal caliper on the rank of the propensity score of $\varkappa = 170,925.1$ for uniform use with all 463 procedures; i.e., this is the smallest feasible caliper in the sense that distinguished Figure 1(iii) and Figure 1(iv). Note that $\varkappa / (T + C) =$

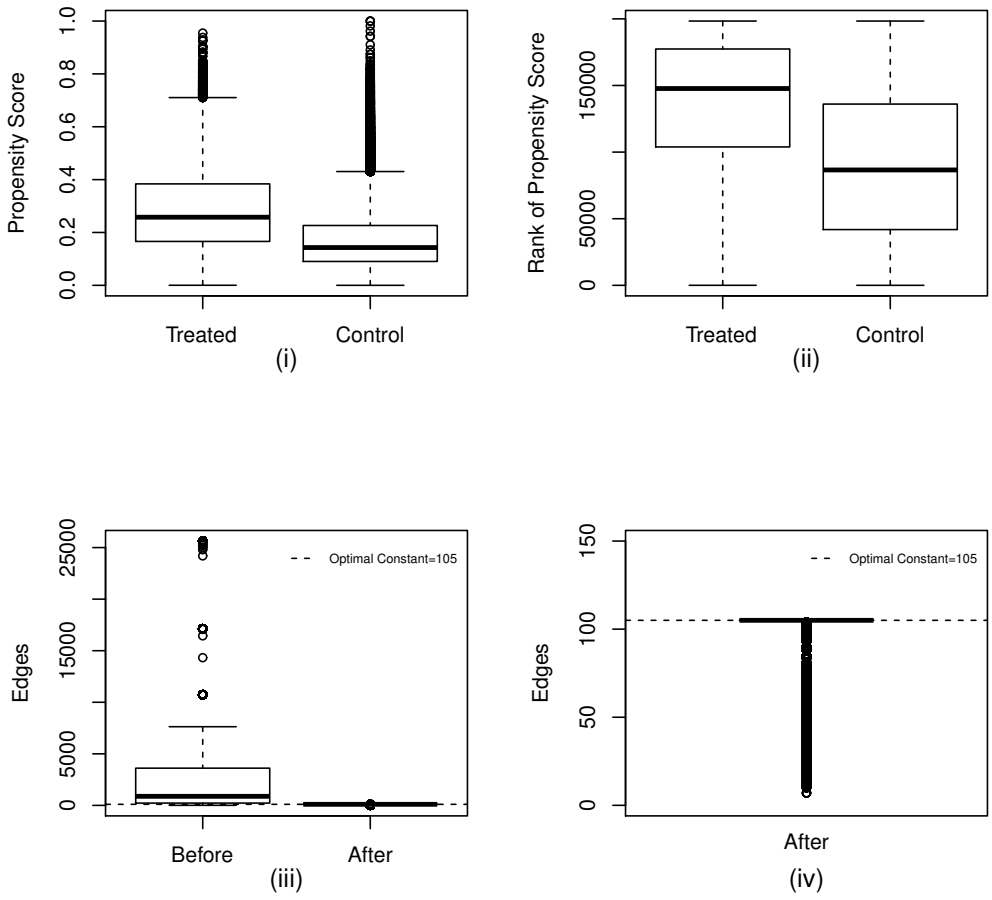


Figure 3: Creating the bipartite graph by exact matching for 463 Principal Procedures with an optimal caliper on the rank of the propensity score. There are 38,841 treated nodes and 159527 control nodes. (i) The propensity score before matching. (ii) Ranks of the propensity score before matching. (iii) Distribution of the number of edges for each treated unit with an optimal caliper on the propensity score, before and after determining the minimal number, $\nu = 105$, of near neighbors. (iv) The "after" boxplot from panel (iii) scaled so that detail is visible.

$170,925.1/198,368 = 0.86$, so this tightest feasible caliper is only eliminating the most extreme differences between ranks of propensity scores. The "Before" boxplot of Figure 3(iii) shows the distribution of the number of remaining edges for the $T = 38,841$ treated children, so many treated children have thousands of potential controls having the same surgical procedure inside the propensity caliper, but many other treated children have

nowhere near thousands of potential controls.

We then asked: If we restrict each treated child to have at most ν nearest neighbors, then how small can ν be while pair matching remains feasible? The answer turns out to be $\nu = 105$ nearest neighbors. For the actual problem, this corresponds with the step from Figure 1(iv) to Figure 1(v) in the toy problem, where $\nu = 2$. These $\nu = 105$ nearest neighbors have the same surgical procedure as the treated child, and the $\nu = 105$ closest ranks of the propensity score among control children with the same surgical procedure. The "After" boxplot of Figure 3(iii) retains at most $\nu = 105$ nearest neighbors, and Figure 3(iv) rescales this boxplot so its details are visible. In Figure 3(iv), almost all treated children now have exactly $\nu = 105$ nearest neighbors, but because of the caliper $\varkappa$ and exact matching for 463 procedures, a small number of treated children have fewer than $\nu = 105$ nearest neighbors. Here, if we reduced $\varkappa$ or ν, pair matching would be infeasible. We now turn to picking the best control child for each treated child, where each treated child has at most $\nu = 105$ potential controls to pick from.

2.5.2. Minimum distance matching with near-fine balance and near-exact pairing

Having determined the bipartite graph $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$ analogous to Figure 1(v), we define the network $(\mathcal{N}, \mathcal{E})$ analogous to Figure 2. By construction, pair matching is feasible in $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$, and every pair match in $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$ pairs children undergoing the same surgical procedure, where there are $\Xi = 463$ procedures.

For $e = (\tau, \gamma) \in \mathcal{B}$, define $\delta^*(\tau, \gamma) \geq 0$ to be a robust Mahalanobis distance based on the covariates in Table 1; see Rosenbaum (2010, §8.3) for a precise definition. Because $\mathcal{B}$ contains only about 4 million potential pairings, rather than about 6 billion potential pairings in the complete bipartite graph for $\mathcal{T} \cup \mathcal{C}$, we compute 4 million rather than 6 billion Mahalanobis distances. Had we split the original problem into 40 parts of about the same size and used a complete bipartite graph for each part, in parallel with Figure 1(ii), then each part would require the computation of about 4 million Mahalanobis distances, making a total of about 160 million distances. As noted previously, we also accelerate the

computation of Mahalanobis distances by orthogonalization.

The additional structure in $(\mathcal{N}, \mathcal{E})$ for near-fine balance attempts to balance 973 Principal Diagnoses. Where Figure 2 had two fine balance categories, black and white, in §2.1.2 the categories are $\upsilon = 1, \ldots, \Upsilon = 973$, with 973 additional nodes.

Near-fine balance ignores who is matched to whom. It is happy to counterbalance imbalances, to offset a mismatch in one pair by an opposite mismatch in the other. However, we prefer to pair two children with the same surgical procedure and also the same diagnosis, but this is not possible because there are far more interaction categories than there are children in the study, $\Xi \times \Upsilon = 463 \times 973 = 450,499 > 198,368 = T + C$. So, we apply an idea from Zubizarreta et al. (2011): we require both near-fine balance for diagnosis and also "near-exact" pairing for diagnosis. Near-exact pairing means that we maximize the number of exactly matched pairs, recognizing that we cannot match everyone exactly. Exact pairing for diagnosis would imply exact balance for diagnosis, but when exactness is absent, near-exact balance and near-exact pairing separate into two different goals.

Near-exact pairing is obtained by imposing a penalty $\Lambda > 0$ on pairs mismatched for the level of the fine-balance variable, $\upsilon(\cdot)$. That is, the robust Mahalanobis distances are penalized: for $e = (\tau, \gamma) \in \mathcal{B}$, define $\delta(\tau, \gamma) = \delta^*(\tau, \gamma) + \Lambda$ if $\upsilon(\tau) \neq \upsilon(\gamma)$ or $\delta(\tau, \gamma) = \delta^*(\tau, \gamma)$ if $\upsilon(\tau) = \upsilon(\gamma)$. Subject to other constraints, if Λ is large enough then a minimum cost flow will avoid as many mismatches for $\upsilon(\cdot)$ as it possibly can, then turn its attention to minimizing the total of Mahalanobis distances within pairs. In our formulation and application, the penalty, Ψ, for imbalance in Proposition 5 is much larger than the penalty, Λ, for inexactness, so balancing takes precedence in the hierarchy of constraints. The `bigmatch` package in `R` lets the user set both Ψ and Λ, for instance, reversing this precedence. A midsized penalty, $\Lambda = |\mathcal{B}|^{-1} \sum_{(\tau,\gamma)\in\mathcal{B}} \delta^*(\tau, \gamma)$, would not maximize the number of pairs matched for $\upsilon(\cdot)$, but instead would give about equal emphasis to $\upsilon(\cdot)$ and to the Mahalanobis distances.

The matching network is now complete. This initial match was not quite close enough

in terms of ED-visits in Table 1, so we gave this covariate a little more emphasis in the covariate distance, and the resulting match is the one we describe.

2.5.3. Quality of the match

Consider, now, the quality of the match in terms of the 29 covariates in Table 1, the $\Xi = 463$ surgical procedures, the $\Upsilon = 973$ principal diagnoses, and the $\Xi \times \Upsilon = 463 \times 973 = 450,499$ interaction categories.

Table 1 shows the covariate means and standardized differences in means for 29 covariates, before and after matching. Figure 4 depicts the changes in 29 standardized differences in means from Table 1. After matching, all 29 standardized differences were less than 0.1, and all large standardized differences before matching were greatly reduced. Rubin (1979)'s results suggest that covariate imbalances of less than 0.1 after matching can safely be removed by covariance adjustment of matched pair differences, whereas model-based adjustments alone cannot safely be relied upon to adjust for observed covariates that have large initial imbalances.

Table 2 examines imbalances in the $\Xi = 463$ procedures, the $\Upsilon = 973$ diagnoses, and their interactions. For a nominal variable $\theta(\cdot)$ with Θ levels, $\theta : \mathcal{T} \cup \mathcal{C} \rightarrow \{1, 2, \ldots, \Theta\}$, we may form a $2 \times \Theta$ contingency table from the matched sample, treated-versus-control by level of the nominal variable. In a completely randomized experiment, there is independence of row and column variables in this $2 \times \Theta$ contingency table, and this provides one benchmark against which the balance in the matched sample can be measured. In our matched sample this table has a total count of $2 \times 38,841$, with 38,841 treated children in the first row and 38,841 control children in the second row.

One measure of imbalance in this $2 \times \Theta$ table is the sum of the Θ absolute differences in the counts in the first and second row, essentially the so-called total variation distance. Indeed, for $\upsilon(\cdot)$ with Υ levels, the total variation distance is $\sum_{\upsilon=1}^{\Upsilon} |d_k|$ from (2.2) that has been the focus of attention all along. This measure can range from 0 if there is exact balance to $2 \times 38,841$ if the treated and control distributions have nonoverlapping support.

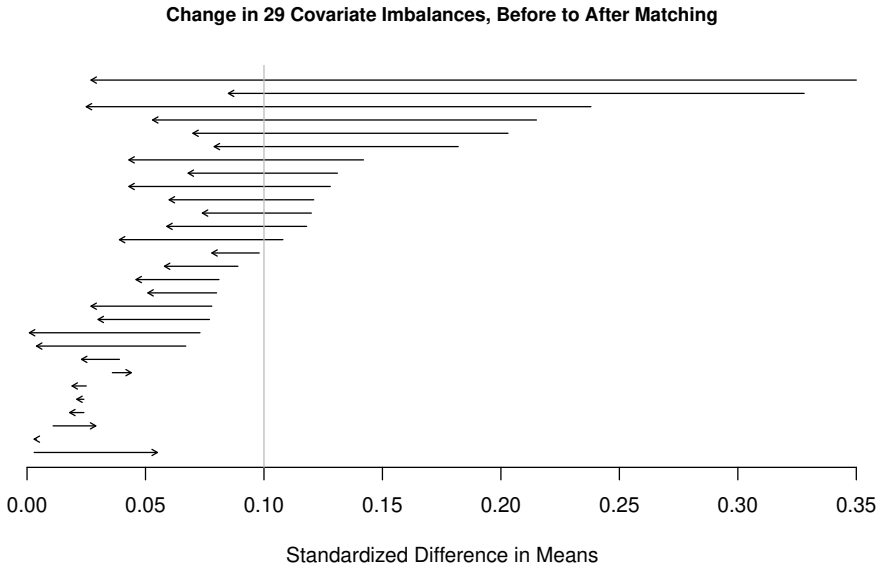


Figure 4: Change in covariate imbalance from before matching to after matching for the 29 covariates in Table 1. The point of the arrow is after matching. A vertical line is at 0.1. After matching, all standardized differences are less than 0.1, and all large imbalances before matching have been greatly reduced.

Table 2: Balance in 463 Principal Procedures, 973 Principal Diagnoses, and their 463×973 interactions. The imbalance in the actual matched sample is compared to the minimum imbalance and the mean imbalance in 10,000 randomized experiments. For each covariate, by each measure, the matched sample is closer to balance than the most balanced of 10,000 randomized experiments formed from the same data.

	Procedure	Diagnosis	Procedure $\times$ Diagnosis
Categories	463	973	463×973
Imbalance	0	846	11704
Minimum imbalance	2880	3354	13122
Mean imbalance	3479	3930	13802
Chi-squared statistic	0	307	8092
Miniumum chi-squared statistic	366	646	8446
Mean chi-squared statistic	462	776	8734

A second measure of imbalance is the usual chi-square statistic for testing independence of row and column variables, although for large Θ we cannot compare it to its usual asymptotic chi-square distribution with $\Theta - 1$ degrees of freedom.

As in Pimentel et al. (2015a), we compare our matched sample to 10,000 randomized ex-

periments each formed from the same data by randomly dividing the $2 \times 38,841$ individuals into two groups of size $38,841$. These 10,000 randomized experiments exhibit the degree of imbalance in observed covariates that a completely randomized experiment would produce. In these 10,000 randomized experiments, there is no systematic bias in the covariates, and all imbalances are due to chance. As seen in Table 1 and Figure 3(i), there were substantial biases in observed covariates before matching, but matching attempted to reduce this systematic imbalance. How does the imbalance in 10,000 randomized experiments compare to the imbalance in observed covariates in our matched sample? Obviously, randomization also tends to balance unobserved covariates, but matching for observed covariates cannot be expected to do this.

Table 2 makes this comparison for three nominal variables, surgical procedure with $\Xi = 463$ levels, principal diagnosis with $\Upsilon = 973$ levels, and their interaction with $\Xi \times \Upsilon = 463 \times 973 = 450,499$ levels. The surgical procedure has imbalance zero because it was exactly matched. For all three nominal variables, both in terms of total variation and in terms of chi-square, the imbalance in the matched sample is smaller than the smallest imbalance found in 10,000 randomized experiments.

2.5.4. Some remarks on alternative matched samples

We used the minimal feasible number of neighbors, $\nu = 105$. Feasibility requires $\nu \geq 105$, not $\nu = 105$, so we tried a match with the same caliper $\varkappa$ but with at most $\nu = 200$ near-neighbors. The balance in a table analogous to Table 2 was very slightly improved, but a table analogous Table 1 looked about the same. By definition, the match with $\nu = 105$ is closer in terms of the propensity score than the match with $\nu = 200$.

Our match used the idea from Zubizarreta et al. (2011) of requiring both near-fine balance for the 973 diagnoses and also near-exact pairing for these same diagnoses. In another variation of the match, if one required just near-fine balance for the 973 diagnoses, then the balance for 973 diagnoses alone was quite good, but the balance for the $\Xi \times \Upsilon = 463 \times 973 = 450,499$ interactions was worse than in the average of 10,000 randomized

experiments. Requiring both near-fine balance and near-exact pairing is helpful when trying to balance the interaction of an exactly matched covariate, like procedure, and a finely balanced covariate, like diagnosis.

2.5.5. Mortality within 30 days of surgery

Table 3 shows mortality within 30 days of surgery in 38,841 matched pairs, in the format associated with McNemar's test for paired binary data. The mortality rates are extremely low in both groups and differ negligibly. How large an effect is compatible with the data? How many deaths could be prevented or caused by having surgery in a children's hospital?

Goeman et al. (2010) proposed a general method of combining a test of the null hypothesis of no effect and an equivalence test. Their clever observation is that there is no multiple testing problem here, because the three underlying null hypotheses are mutually inconsistent, so at most one true null hypothesis is tested. As there is no sign of an effect in Table 3, the main question concerns equivalence: To what extent does Table 3 rule out large effects? We follow Pimentel et al. (2015a, §5), using the particular test in Rosenbaum (2002, §6) combined with the general method of Goeman et al. (2010).

Table 3: Mortality within 30 days of surgery in 38,841 matched pairs of two children, one receiving surgery in a children's hospital, the other in an adult hospital. The table counts pairs, not children.

		Child in an adult hospital		
		Dead	Alive	Total
Child in a	Dead	16	94	110
children's hospital	Alive	95	38636	38731
	Total	111	38730	38841

Were Table 3 from a paired randomized experiment, we would be 95% confident that surgery in an adult hospital caused a net increase of at most 25 deaths, or prevented at most 23 deaths, where $25/38,841 = 0.00064$ and $23/38,841 = 0.00059$. If we acknowledge that Table 3 is not from a randomized experiment, and allow for an unobserved covariate that doubles the odds of death and doubles the odds of treatment in a children's or adult hospital, then we are 95% confident that surgery in an adult hospital caused a net increase

of at most 43 deaths, or prevented at most 41 deaths, where $43/38,841 = 0.00111$ and $41/38,841 = 0.000106$. (This is $\Gamma = 1.25$ in Rosenbaum (2002) using the interpretation of Γ in Rosenbaum and Silber (2009).)

We also looked at "mortality-or-readmission within 30 days of surgery". Again, the rates differed negligibly in the two types of hospitals.

2.6. Summary

We proposed and illustrated a method for optimal matching in large administrative data sets describing hundreds of thousands of people. Within $\Xi = 463$ exact match categories, the method used Glover's algorithm to find the minimal feasible caliper on the propensity score, together with the minimal feasible number of nearest neighbors, $\nu = 105$; then, it minimizes a covariate distance while finely balancing $\Upsilon = 973$ additional categories. After matching, the $\Xi \times \Upsilon = 463 \times 973 = 450,499$ interaction categories were better balanced than the most balanced of 10,000 randomized experiments built from the same data.

CHAPTER 3 : Evaluating and Improving a Matched Comparison of Antidepressants and Bone Density

3.1. Introduction: Problem and Empirical Example

3.1.1. Evaluate and then improve covariate balance

A gold standard for estimating causal effects is to use randomized experiments, where there is no systematic difference in covariates between the treated and control groups, both observed and unobserved. To identify a causal relationship in observational studies, a common approach is to mimic randomized experiments in the design stage by creating similar treated and control groups for observed covariates, with either matching or weighting. Matching methods, including the propensity score and related techniques, exact and near-exact match, fine balance and near-fine balance, have been widely studied to improve covariate balance. For a thorough review, see Rosenbaum (2010) and Stuart (2010). Traditional weighting methods are based on the propensity score, known as inverse probability of treatment weighting (Rosenbaum, 1987). In recent years, weighting methods directly balancing covariates have been developed; see for example, Chan et al. (2016); Fan et al. (2016); Imai and Ratkovic (2014); Wang and Zubizarreta (2020); Zubizarreta (2015).

Matching is commonly understood to estimate the average treatment effect on the treated (ATT) in observational studies (Rosenbaum and Rubin, 1985a), whose identification assumptions are slightly weaker than the assumptions of unconfoundedness and positivity (Heckman et al., 1997). It is important to assess covariate balance in the matched samples, since lack of balance in the covariates can induce a bias of the matching estimator of a treatment effect (Imbens and Rubin, 2015, §18). Rubin (1979) suggests that model-based adjustments for matched samples which effectively balance covariates are more robust to model mis-specification than using model-based adjustments alone.

Balance checking in matching is usually done informally. The informal diagnostics, e.g., standardized mean differences and variance ratios, can provide useful information, but they

have several limitations. See Rosenbaum (2010, §8) for details of the common informal diagnostics. First, the informal approaches usually examine one variable at a time. This gives many comparisons, and hence produces multiple testing issues. It is also rare to check joint distributions for balance, even bivariate distributions, which can lead to a loss of important information. Furthermore, it can be difficult to know when to worry about an observed imbalance. Even a randomized experiment, the benchmark, would produce some imbalances by chance, particularly if there are many covariates. Finally, it is crucial that balance checks provide guidance on how to improve the match.

In the past two decades, several authors have proposed balance tests in observational studies (e.g., Rosenbaum, 2005; Hansen and Bowers, 2008; Austin, 2009; Chen and Small, 2020; Gagnon-Bartsch and Shem-Tov, 2019). The aim of this paper is to organize and formalize the assessment of covariate balance so that (i) balance of both marginal distributions and joint distributions of the matched sample is compared with that of complete randomizations; (ii) it is easy to evaluate multiple testing; and (iii) any detected imbalance suggests an improved match. Notably, methods developed in the current paper are not restricted to a particular matching procedure; they can be applied to evaluate the covariate balance of any matched sample. After deciding which covariates to match on, by whichever methods one prefers (Rubin, 2007; Stuart, 2010; Pearl, 2012; Pimentel et al., 2016), the proposed methods are applicable to evaluate and improve the balance of those covariates.

Exact matching on every coordinate of a high-dimensional covariate is neither possible nor needed (Rosenbaum and Rubin, 1983). If we knew that a variable was not a confounder, there would be no need to match for it (e.g., de Luna et al., 2011; Hahn, 2004). Matching is not always feasible, if the distributions in the treated and control groups are very different; for example, the two distributions have no common support. Even with just one covariate, there is an upper bound on bias reduction with pair matching (Rubin, 1973). We need to distinguish problems that cannot be solved by matching, those that can be solved by a better match, and those that are small compared to the imbalances that occur in randomized

experiments. Pimentel (2016) and Yu et al. (2020) checked the balance of several nominal variables by comparing a match to 10,000 simulated randomized experiments with the same marginal covariate distributions. The newly proposed method is a substantial generalization of this approach: the joint distribution of various imbalance measures of a match is compared to that of simulated completely randomized experiments built with the same covariates.

3.1.2. Example: Do antidepressants reduce bone density?

To illustrate matching methodology, effects of SSRIs on bone density of adults will be examined, with the data set described in §3.6. A 4-to-1 match is built with three conventional techniques (minimizing total covariate distance, with a propensity score caliper, and fine balancing on education). This basic match ("Base") greatly reduces the covariate imbalances, but the informal diagnosis results summarized in Table 4 suggest that there is some remaining bias in diabetes. The proposed iterative method is applied to improve this basic match in §3.6. At every iteration, diagnostics provide guidance on where adjustments should focus. In particular, for the basic match, our proposed diagnostics suggest that it is the interaction of diabetes and preference to weigh more that causes the imbalance in diabetes, instead of the marginal distribution of diabetes itself as suggested by the informal diagnosis. Adjusting this bivariate distribution leads to a better match, and similar procedures can be continued. After three adjustments, the final match ("Iter 3") exhibits better balance than the basic match ("Base") as seen in Table 4, and additionally exhibits balance for bivariate distributions not displayed in Table 4. With the proposed method, a match balancing both marginal distributions and bivariate distributions can be constructed in a more efficient way by targeting the most serious problem for each iteration of adjustments.

3.1.3. Outline

The rest of this paper is organized as follows. Sections 3.2 and 3.3 propose the general framework, where §3.2 focuses on how to measure covariate imbalance, and §3.3 discusses how to adjust a match if the measures diagnose a problem. Section 3.4 illustrates one specific implementation of the proposed method, with theoretical guarantees that it can identify the major problems in a reliable way. Simulations in §3.5 verify that the proposed

Table 4: Balance table for marginal distributions of 19 covariates and propensity score (pscore), with estimated p-values based on 2,000 simulated randomized experiments: Before match, basic match (Base), and final match (Iter 3). Standardized mean differences (SMDs) greater than 0.1 in absolute value and p-values less than 0.1 are in **bold**.

	Mean				SMD			SMD p-value		t-test p-value		Wilcoxon p-value		KS test p-value	
	All controls	Treated	Base	Iter 3	Before match	Base	Iter 3	Base	Iter 3	Base	Iter 3	Base	Iter 3	Base	Iter 3
age	47.514	53.845	53.607	53.949	**0.368**	0.014	-0.006	0.847	0.934	0.840	0.929	0.997	0.736	0.421	0.320
female	0.478	0.653	0.652	0.635	**0.359**	0.004	0.037	1.000	0.643	1.000	0.594	1.000	0.643	1.000	0.643
black	0.170	0.087	0.073	0.082	**-0.250**	0.041	0.014	0.533	0.902	0.533	0.902	0.533	0.902	0.533	0.902
Hispanic	0.301	0.144	0.144	0.147	**-0.383**	0.000	-0.007	1.000	0.921	1.000	0.921	1.000	0.921	1.000	0.921
povertyNA	0.084	0.047	0.037	0.037	**-0.149**	0.040	0.040	0.491	0.481	0.491	0.481	0.491	0.481	0.491	0.481
poverty	2.480	2.566	2.706	2.688	0.053	-0.086	-0.076	0.203	0.269	0.213	0.282	0.156	0.196	**0.058**	0.100
education	3.298	3.513	3.513	3.534	**0.175**	0.000	-0.018	1.000	0.792	1.000	0.780	1.000	0.777	1.000	0.962
height	167.910	166.847	166.756	166.934	**-0.109**	0.009	-0.009	0.882	0.900	0.882	0.901	0.974	0.823	0.557	0.631
weight	79.340	82.536	81.620	81.805	**0.171**	0.049	0.039	0.465	0.566	0.486	0.574	0.524	0.613	0.847	0.845
BMI	28.068	29.595	29.290	29.287	**0.256**	0.051	0.052	0.445	0.451	0.469	0.467	0.476	0.463	0.359	0.486
cotinine	54.756	64.333	52.358	53.176	0.078	0.098	0.091	0.164	0.165	0.167	0.166	**0.075**	0.128	0.104	0.221
vitaminD	63.461	73.513	72.719	72.125	**0.371**	0.029	0.051	0.654	0.434	0.670	0.463	0.808	0.630	0.604	0.516
diabetes	0.112	0.195	0.132	0.187	**0.233**	**0.176**	0.023	**0.011**	0.793	**0.015**	0.793	**0.011**	0.793	**0.011**	0.793
insurance	0.730	0.888	0.908	0.896	**0.410**	-0.052	-0.021	0.336	0.713	0.336	0.713	0.336	0.713	0.336	0.713
weighmore	0.081	0.051	0.026	0.047	**-0.122**	0.099	0.015	**0.050**	0.874	**0.082**	0.874	**0.050**	0.874	**0.050**	0.874
weighless	0.600	0.729	0.765	0.750	**0.275**	-0.077	-0.044	0.229	0.475	0.229	0.475	0.229	0.475	0.229	0.475
weightchange	-0.091	-0.660	-0.410	-0.362	-0.068	-0.030	-0.036	0.646	0.595	0.708	0.651	0.878	0.721	**0.013**	0.227
physicalact	0.285	0.191	0.177	0.187	**-0.221**	0.034	0.011	0.601	0.866	0.601	0.935	0.601	0.935	0.601	0.935
dietsup	0.464	0.625	0.629	0.623	**0.327**	-0.009	0.004	0.937	1.000	0.937	1.000	0.937	1.000	0.937	1.000
pscore	0.054	0.101	0.095	0.095	**0.786**	0.100	0.095	0.153	0.178	0.183	0.216	0.434	0.440	0.625	0.688

balance checks in §3.4 not only control the probability of falsely identifying a covariate imbalance, but also diagnose and identify the major problems with a high probability. The techniques are illustrated in §3.6 to study the effects of antidepressants on bone density.

3.2. Comparing Covariate Balance with Complete Randomizations

3.2.1. The joint distribution of M measures of imbalance

Suppose the goal is to balance d observed covariates $\mathbf{X} = (X_1, ..., X_d)$. Consider M imbalance measures, T_i, $i = 1, ..., M$. The joint behavior of $(T_1, ..., T_M)$ is of interest. The behavior these measures would exhibit in a completely randomized experiment is regarded as a benchmark, and the joint behavior of the M measures for a matched sample is compared with the benchmark. For more discussion on why we use completely randomized experiments as the benchmark, see Appendix A.2.1. The joint distribution of the M measures in a randomized experiment can be obtained empirically by repeatedly randomly assigning the matched individuals to new treatments. More concretely, consider a κ-to-1 match (Rosenbaum, 2010, §8). Suppose there are n treated subjects and m^* potential controls. A κ-to-1 match selects $m = \kappa n$ controls from m^* candidate controls. Each randomization is formed from the same matched data by randomly dividing the $m + n$ individuals into two groups of size n and m. These randomized experiments exhibit the degree of imbalance in observed covariates that a completely randomized experiment would produce. In these randomized experiments, there is no systematic bias in the covariates, and all imbalances are due to chance. How do these experiments compare to the observational match in terms of balance for observed covariates?

3.2.2. Combining M measures into a single summary measure

Evaluating the covariate balance of matched samples is equivalent to testing the global null hypothesis H_0 that the distributions of $\mathbf{X}$ are the same in the matched treated and control groups. The global null H_0 can be violated from different perspectives. In other words, H_0 can be decomposed into a series of M null hypothesis H_{0i}, $i = 1, ..., M$, such that $H_0 = \cap H_{0i}$. For example, H_{01} could be the null hypothesis that the distributions of X_1 are the same in matched treated and control groups, and H_{02} could be the null hypothesis that

the joint distributions of X_1 and X_2 are the same in the two groups, etc. Each hypothesis H_{0i} can be tested using T_i with p-value p_i. To avoid false rejections from testing many hypotheses, H_0 is tested using a single summary measure T^* summarizing the M p-values. Many papers in the literature have proposed methods that combine p-values.

A simple choice of T^* is the minimum p-value

$$p^* = \min p_i. \tag{3.1}$$

Of course, p^* is not a p-value but a test statistic, whose null distribution yields a true p-value. Berk and Jones (1978) showed p^*, as a statistic, has Bahadur efficiency at least that of any component T_i. They called this a relatively optimal combination. More precisely, with treated group size n and matched control group size $m = \tau n$ for some fixed $\tau \geq 1$, let L_n be the level attained by p^* (i.e., the corresponding p-value), and L_n^i be the level attained by T_i. Then, with probability one, $\liminf_n[-n^{-1}\log L_n]$ is at least as large as every $\liminf_n[-n^{-1}\log L_n^i]$.

There are several generalizations of the minimum p-value motivated by Fisher's product of p-values (Fisher, 1925). One generalization uses the product of the k smallest p-values,

$$p^k = \prod_{i=1}^{k} p_{(i)}, \tag{3.2}$$

where $p_{(i)}$'s are the ordered p-values. This is the rank truncated product (RTP) of p-values proposed by Dudbridge and Koeleman (2003). This may be more robust because several non-significant p-values may be more significant than one relatively small p-value, e.g., "Two 0.06 results are much stronger evidence against the null than one 0.05; and 10 p's of 0.10 are stronger evidence against the null than 5 p's of 0.05" (Rosenthal, 1990, p.133).

Another popular method of combining many p-values is to only use p-values below a specified threshold η, i.e., the truncated product method (TPM) for combining p-values proposed

by Zaykin et al. (2002). The combined statistic T^* then becomes

$$p_\eta^* = \prod_{i=1}^{M} p_i^{I(p_i \leq \eta)}. \tag{3.3}$$

Both RTP and TPM are special cases of Fisher's statistic, namely the product of all p-values (Fisher, 1925).

A test of the global null hypothesis H_0 can be conducted based on a chosen summary measure T^*. Under the independence of the p-values, the critical point can be derived for different choices of T^*; for example, see Dudbridge and Koeleman (2003); Zaykin et al. (2002). However, the situation here is much more complicated due to the strong dependence between the p-values. Sampling randomizations solves this problem, permits dependence, and also captures the underlying correlation structure of the statistics. Propositions 7 and 8 present level α tests of H_0, for continuous and discrete summary statistics T^*, respectively. When T^* is continuous, the probability integral transformation can be applied. Suppose B simulated randomizations are conducted. Without loss of generality, assume that $\alpha(B+1)$ is an integer.

Proposition 7 *Let F be the cumulative distribution function (CDF) of a continuous random variable. Suppose $X_1, ..., X_B \overset{i.i.d.}{\sim} F$, and Y is another independent copy drawn from F. Let $X_{(i)}$'s denote the order statistics of $X_1, ..., X_B$. Then, we have*

$$\mathbb{P}(Y < X_{(\alpha(B+1))}) = \mathbb{P}(Y \leq X_{(\alpha(B+1))}) = \alpha. \tag{3.4}$$

Proposition 7 can be derived using the fact that $P(X_{(j)} \leq Y < X_{(j+1)}) = 1/(B+1)$, for all $j = 0, ..., B$; for details see Fligner and Wolfe (1976). Let $T^*_{(i)}$ denote the i_{th} ordered T^* value from B randomizations. Consider

$$\phi_c(T^*) = \begin{cases} 1, \text{ if } T^* \leq T^*_{(\alpha(B+1))}; \\ 0, \text{ if } T^* > T^*_{(\alpha(B+1))}. \end{cases} \tag{3.5}$$

By Proposition 7, $\mathbb{E}[\phi_c(T^*)] = \alpha$, where $\mathbb{E}$ denotes the expectation under H_0. That is, $\phi_c(T^*)$ is a level α test of H_0 for a continuous summary statistic T^*.

On the other hand, when T^* is discrete, $\phi_c(T^*)$ may not control the type I error, i.e., $\mathbb{E}[\phi_c(T^*)] \geq \alpha$. Proposition 8 demonstrates why this is the case and suggests a level α test.

Proposition 8 *Let F be the CDF for a discrete random variable taking values $\{v_i : v_i \in \mathbb{R}\}$ with corresponding probabilities $\{q_i\}$. Pick an $\epsilon \geq \max q_i$. Let $X_1, ..., X_B \overset{i.i.d.}{\sim} F$, and Y be another independent copy of F. Then, we have*

$$\mathbb{P}(Y < X_{(\alpha(B+1))}) \leq \alpha \quad \text{and} \quad \alpha \leq \mathbb{P}(Y \leq X_{(\alpha(B+1))}) \leq \alpha + \epsilon. \tag{3.6}$$

The proof of Proposition 8 uses a continuous analogue of discrete random variables, as proposed by Fligner and Wolfe (1976). See Appendix A.2.2 for more details. Proposition 8 provides a different level α test of H_0 for a discrete summary statistic T^*:

$$\phi_d(T^*) = \begin{cases} 1, \text{ if } T^* < T^*_{(\alpha(B+1))}; \\ 1 \text{ with probability } \gamma, \text{ if } T^* = T^*_{(\alpha(B+1))}; \\ 0, \text{ if } T^* > T^*_{(\alpha(B+1))}. \end{cases} \tag{3.7}$$

Here, γ can be chosen to make sure $\mathbb{E}[\phi_d(T^*)] = \alpha$. For details of how to choose γ, see the proof of the Neyman-Pearson lemma in Lehmann and Romano (2006, §3). In practice, γ can be estimated with the simulated randomized experiments formed from the matched data.

3.2.3. Choice of individual measures of imbalance

This section presents several choices of T_i. When testing the marginal distribution of each X_i, natural choices are the familiar two-sample statistics, e.g., standardized mean differences, two-sample t statistics, Wilcoxon rank sum statistics, Kolmogorov-Smirnov statistics.

Next, consider the balance of the joint distribution of k covariates $\widetilde{\mathbf{X}} = [X_1, ..., X_k]$. One

approach uses a real-valued function of $\widetilde{\mathbf{X}}$, $f(\widetilde{\mathbf{X}})$; then this reduces to the marginal distribution case above. For instance, a univariate test may be applied to the principal components of $\widetilde{\mathbf{X}}$. Another option of $f(\widetilde{\mathbf{X}})$ is the ranking of multivariate data $\widetilde{\mathbf{X}}$, e.g., with a minimum spanning tree (Friedman and Rafsky, 1979). Alternatively, a learning algorithm may pick out some patterns of $\widetilde{\mathbf{X}}$ that predict the treatment. Then $f(\widetilde{\mathbf{X}})$ can be the predicted probability of receiving the treatment or the classification group. Several widely used approaches include logistic regression, decision trees, assembling methods (random forests, boosting, bagging), and support vector machines. For more discussion of classication methods, see Friedman et al. (2001).

An alternative way is to apply the multivariate Kolmogorov-Smirnov test to k-dimensional joint distributions. Let $\mathbb{P}_n$ and $\mathbb{P}_m$ denote the empirical probability functions of the two groups with sample size n and m, respectively. Then, for each $\mathbf{v} = (v_1, ..., v_k) \in \mathbb{R}^k$, $F_n(\mathbf{v}) = \mathbb{P}_n(X_1 \leq v_1, ..., X_k \leq v_k)$ and $G_m(\mathbf{v}) = \mathbb{P}_m(X_1 \leq v_1, ..., X_k \leq v_k)$ are the multivariate empirical distribution functions of the two groups. Bickel (1969) uses the test statistic

$$\sqrt{\frac{nm}{n+m}} \sup_{\mathbf{v} \in \mathbb{R}^k} |F_n(\mathbf{v}) - G_m(\mathbf{v})|. \tag{3.8}$$

Each $\mathbf{v} \in \mathbb{R}^k$ can split the space of $\mathbb{R}^k$ into 2^k orthants based on k orthogonal halfspaces $X_1 = v_1, ..., X_k = v_k$ (Roman et al., 2005). For example, an orthant in $\mathbb{R}^k$ is a ray in $\mathbb{R}$ and a quadrant in $\mathbb{R}^2$. Bickel's primary focus is on the lower orthants $\{(X_1, ..., X_k) : X_1 \leq v_1, ..., X_k \leq v_k\}$. Probabilities of other orthants of interest, e.g., $\mathbb{P}(X_1 > v_1, ..., X_k > v_k)$, can be derived from at most $2^k - 1$ lower orthants. That is, balance of other orthants can be obtained by adjusting several lower orthants. However, a single adjustment based on a simple cut at one orthant is preferred in matching. Therefore, it is helpful to consider the probability of 2^k orthants for each $\mathbf{v} \in \mathbb{R}^k$ simultaneously. Let $\mathcal{Q} := \{Q_{\mathbf{v}j} : \mathbf{v} \in \mathbb{R}^k, j = 1, ..., 2^k\}$ denote the collection of all such orthants. Then, T_i can be defined in the following form:

$$T_i = \sqrt{\frac{nm}{n+m}} \sup_{Q \in \mathcal{Q}} |\mathbb{P}_n(Q) - \mathbb{P}_m(Q)|, \tag{3.9}$$

where $\mathbb{P}_n(Q)$ and $\mathbb{P}_m(Q)$ are the empirical probabilities of the orthant Q in the two groups.

3.3. From Diagnosed Imbalances to Better Matched Samples

This section presents a general way to improve the match once an imbalance is identified. In the ideal situation of a randomized experiment, the treatment assignment is independent of any attribute of subjects in the study. That is, the treated and control groups cannot be classified by any variables. Improving a match is the mirror image of classifying treated and control subjects based on observed covariates. Once a feature distinguishing the matched treated and control groups is identified, a new match balancing that feature is constructed.

Formally, at every iteration l, the proposed test in §3.2.2 is performed with a summary statistic T^* to evaluate the global null H_0, giving a p-value p_l. If H_0 is rejected (i.e., $p_l \leq \alpha$), find the most problematic T_i's (i.e., T_i's giving smallest p-values), and adjust the match to balance the corresponding features – a new match balancing the identified imbalanced features from all iterations as closely as possible is conducted. The iterative algorithm can be stopped when H_0 is retained. If one would like to impose other covariate balance measures, the algorithm can be continued until a satisfactory match is obtained. See the application in §3.6.1 for an example. The risks involved in balance testing are relatively small; at worst we work to improve a match that is already adequate. Nonetheless, by the principle of testing in order, if the balance test for each iteration is done at level α, the chance of falsely rejecting at least one balanced match is at most α (Rosenbaum, 2008).

Iterative reviews of matched samples search for improvements of an existing match. It may discover an improvement or find that the two groups are very different and not matchable. If an unsolvable matching problem is identified, one can define a new target population to improve covariate overlap; see Fogarty et al. (2016, §3.3) for a review of possible solutions. In the adjustment process, one can choose to adjust for a single most imbalanced variable or fix multiple imbalances at once. The advantage of adjusting one at a time is that fixing one problem may balance other imbalanced features without further adjustments. On the other hand, this might require a few iterations to accomplish what might be accomplished in one.

There is a trade-off between the number of iterations and improving matches effectively.

Common strategies for balancing an identified feature include near-fine balance (Yang et al., 2012), near-exact match (Rosenbaum, 2010, §9.2), non-directional and directional penalties (Yu and Rosenbaum, 2019), etc. In this paper, simulations in §3.5.3 and the application in §3.6.1 use one specific choice – near-fine balance of the interaction of the identified binary variables from each iteration, as in Yang et al. (2012). Alternatively, one could use integer programming (Zubizarreta, 2012) to control the marginal balance at some threshold. This is a promising alternative to balance more variables when the sample size is not very large.

3.4. A Simple but Useful Implementation

This section illustrates a specific implementation of the general method proposed in Sections 3.2 and 3.3. Suppose the matched treated and control groups have sample size n and m, respectively. Denote the matching ratio as $\tau := m/n$, and assume τ is fixed. That is, the matching is one treated to $\tau \geq 1$ controls on average, i.e., $m = \tau n$. Let $\mathbb{P}_T$ and $\mathbb{P}_C$ be the probability distribution function of the d-dimensional covariates $\mathbf{X} = [X_1, ..., X_d]$ for the matched treated and control groups, respectively, and $\mathbb{P}_n$ and $\mathbb{P}_m$ be the empirical versions of $\mathbb{P}_T$ and $\mathbb{P}_C$, respectively. Let $[d] := \{1, ..., d\}$ denote the set of indices for all d covariates. To test the null hypothesis $H_0 : \mathbb{P}_T = \mathbb{P}_C$ versus $H_1 : \mathbb{P}_T \neq \mathbb{P}_C$, consider a max-type statistic evaluated on a finite set of values, motivated by the traditional Kolmogorov-Smirnov (KS) statistics. We call it a generalized KS statistic on a finite set, GFKS for short. A GFKS statistic T_I focusing on the maximum deviation between $\mathbb{P}_n$ and $\mathbb{P}_m$ on a subset of covariates $\mathbf{X}_I$, $I \subset [d]$. If $|I| = k$, T_I is a k-dimensional GFKS statistic. Specifically, let $\mathcal{V}_I$ be a finite set of values in $\mathbb{R}^{|I|}$. For each $\mathbf{v} \in \mathcal{V}_I$, there are $2^{|I|}$ corresponding orthants $Q_{\mathbf{v}j}$, $j = 1, ..., 2^{|I|}$, as illustrated in §3.2.3. Let $\mathcal{Q}_I := \{Q_{\mathbf{v}j} : \mathbf{v} \in \mathcal{V}_I, j = 1, ..., 2^{|I|}\}$ be the collection of all possible orthants defined by $\mathcal{V}_I$. For example, when $|I| = 1$, $\mathbf{V}_I$ can be chosen as a set of marginal quantiles of a continuous variable, and $\mathcal{Q}_I$ can be defined as $\mathcal{Q}_I := \{(-\infty, \mathbf{v}], (\mathbf{v}, \infty) : \mathbf{v} \in \mathcal{V}_I\}$. Then, the GFKS statistic on $\mathbf{X}_I$, T_I, is defined as

$$T_I := \sqrt{\frac{nm}{n+m}} \max_{Q \in \mathcal{Q}_I} |\mathbb{P}_n(Q) - \mathbb{P}_m(Q)|. \tag{3.10}$$

Denote the corresponding p-value as p_I.

The KS test, which focuses on the maximum deviation of two empirical distribution functions, is a popular approach for two-sample testing. However, the intense computation for testing multi-dimensional distributions limits use of the multi-dimensional KS test. In addition, as discussed in §3.2.3, the traditional k-dimensional KS test focuses on lower orthants, but it is better to look at all 2^k orthants when trying to improve a match. To overcome these limitations, a k-dimensional GFKS statistic is evaluated at $\mathcal{Q}_I$ – all $2^{|I|}$ orthants for a limited set of marginal quantiles $\mathcal{V}_I$, instead of lower orthants for every single observation. As a max-type statistic, GFKS, like the KS statistic, can not only indicate a problem, but also suggest an imbalanced binary feature that the next match must balance. For example, a bivariate GFKS statistic identifies a worst cut of the plane into one quadrant and its complement.

To combine all GFKS statistics, one-dimensional, two-dimensional, ..., up to s-dimensional, use the minimum p-value, T^*, as a summary statistic. Specifically, let $\mathcal{I} := \{I \subset [d] : |I| \leq s\}$ and $M := |\mathcal{I}|$. Then, the summary statistic T^* can be defined as

$$T^* := \min_{I \in \mathcal{I}} p_I. \tag{3.11}$$

The real worry in evaluating matched samples is distinguishing moderately large covariate imbalances from issues of multiple testing where something turns out to be significant just because we conduct many balance checks. The ability to distinguish these two cases can be measured by Bahadur slope (Bahadur, 1967; Wieand, 1976). For a similar discussion in sensitivity analysis, see Rosenbaum (2015). Therefore, to evaluate the performance of T^*, we consider the Bahadur slope of T^*

$$\lim_{n \to \infty} -\frac{2 \log L_n(T^*)}{n+m}, \tag{3.12}$$

where $L_n(T^*)$ denotes the attained level of T^* (i.e., p-value) based on n matched sets.

Proposition 9 *Suppose T^* is the minimum p-value of GFKS statistics T_I, $I \in \mathcal{I}$. Let $d_I = \max_{Q \in \mathcal{Q}_I} |\mathbb{P}_T(Q) - \mathbb{P}_C(Q)|$. Then, T^* has Bahadur slope*

$$\frac{4\tau}{(\tau+1)^2} \max_{I \in \mathcal{I}} d_I^2. \tag{3.13}$$

The Bahadur slope of the summary statistic T^* is suggested by Proposition 9. For the proof of Proposition 9, see Appendix A.2.2.

The proposed algorithm treats the smallest p-value as the primary focus, and adjusts for that at every iteration. The following proposition says that in the limit, we are nearly certain to select for the next match one of the cuts that produce the maximum Bahadur slope.

Proposition 10 *Let $\mathcal{J}^* = \{I \in \mathcal{I} : d_I = \max d_I\}$ denote the collection of I's for which T_I has the largest Bahadur slopes and $\mathcal{J} = \{I \in \mathcal{I} : d_I < \max d_I\}$ denote the rest of I's. Then, we have*

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{I^* \in \mathcal{J}^*} p_{I^*} < \min_{I \in \mathcal{J}} p_I \right) = 1. \tag{3.14}$$

Proposition 10 suggests that in sufficiently large samples, the largest Bahadur slopes lead to the smallest p-values with a large probability. See Appendix A.2.2 for the proof of Proposition 10. This provides a theoretical guarantee that using the minimum p-value of GFKS statistics can identify the most imbalanced cut of the worst-balanced covariate reliably; moreover, it suggests a binary variable that the next iteration match should adjust. For a detailed description of the iterative algorithm, see Appendix A.2.3. Theoretical properties of a more general way of combining p-values are also discussed in Appendix A.2.3.

3.5. Simulation

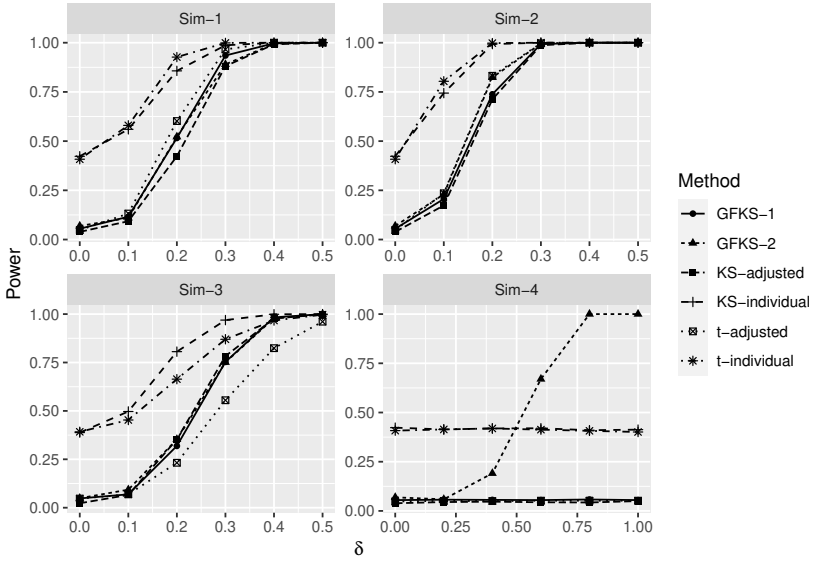
3.5.1. Power (type II error) and type I error

Simulations are conducted to study the power and type I error of the proposed balance checks in §3.4, i.e., minimum p-value of GFKS tests, both univariate ($|I| = 1$) and bivariate

($|I| \leq 2$). GFKS tests are also compared with two common practices - two-sample t-test and KS test on individual variables, as well as their multiple testing adjusted versions with the proposed general framework in §3.2. To evaluate the finite sample performance, the empirical probability that the null hypothesis is rejected, $\hat{p}$, is calculated based on 1,000 replications in various settings. Consider matched samples with 500 treated individuals and 500 controls; each individual has 10 observed covariates $\mathbf{X}$. We evaluate the GFKS statistics based on the finite orthant collection $\mathcal{Q}_I$ at (20%, 40%, 60%, 80%) quantiles of standard normal $N(0,1)$. P-values of GFKS tests are approximated based on $B = 1000$ simulated randomizations. Four experiments are considered – in the first three experiments, all imbalances are on the marginal distributions; in the last experiment, all marginal distributions are the same in the two groups and only the joint distribution of the first two covariates is out of balance. Specifically, let δ denotes the signal strength in the experiments. In the first simulation (Sim-1), $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_{10})$ in the treated group and $\mathbf{X} \sim N((\delta, 0, ..., 0), \mathbf{I}_{10})$ in the control group, where $\mathbf{I}_k$ denote the $k \times k$ identity matrix. In the second case (Sim-2), $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_{10})$ in the treated group and $\mathbf{X} \sim N((\delta, \delta/2, ..., \delta/2), \mathbf{I}_{10})$ in the control group. In the third scenario (Sim-3), $\mathbf{X} \sim t_3(\mathbf{0}, \mathbf{I}_{10})$ in the treated group and $\mathbf{X} \sim t_3((\delta, \delta/2, ..., \delta/2), \mathbf{I}_{10})$ in the control group, where $t_d(\mu, \Sigma)$ denote the multivariate t distribution with degree of freedom d, non-central parameter μ and covariance matrix Σ. In the last simulation (Sim-4), $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_{10})$ in the treated group and $\mathbf{X} \sim N(\mathbf{0}, \Sigma^*)$ in the control group, where Σ^* is a 10×10 matrix such that $\Sigma^*_{ii} = 1$, $\Sigma^*_{12} = \Sigma^*_{21} = \delta$, and $\Sigma^*_{ij} = 0$ otherwise. We plot the estimated probability $\hat{p}$ against the signal strength δ in these settings in Figure 5(a).

Results in Figure 5(a) show that the common practice of using individual t-test or individual KS test inflates the type I error in all four experiments, while the other four methods control the type I error at level $\alpha = 0.05$. In addition, the two GFKS tests and the adjusted KS test can detect the imbalances on the marginal distributions in a more reliable way. Specifically, the four methods have similar performance if the true distribution is normal (Sim-1 and Sim-2); nonetheless, GFKS and the adjusted KS test have a larger power than the adjusted t-test when the true distribution is not normal (Sim-3). In the last simulation (Sim-4),

(a) Estimated Power and Type I Error in 1000 Replications



(b) Empirical Probability of Identifying the Major Problems in 1,000 Replications

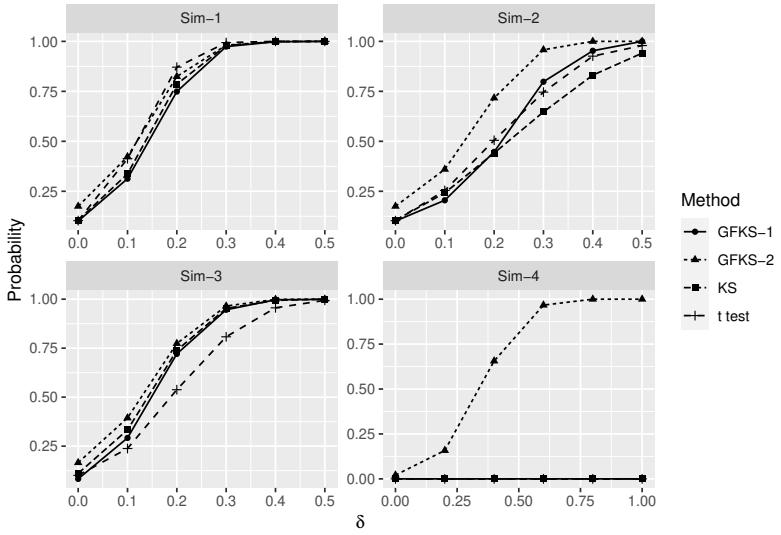


Figure 5: Comparison of individual two-sample t test (t-individual), individual Kolmogorov-Smirnov test (KS-individual), adjusted t test (t-adjusted), adjusted Kolmogorov-Smirnov test (KS-adjusted), univariate GFKS (GFKS-1) and bivariate GFKS (GFKS-2) in the simulations. Sim-1: $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_{10})$ in the treated group and $\mathbf{X} \sim N((\delta, 0, ..., 0), \mathbf{I}_{10})$ in the control group; Sim-2: $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_{10})$ in the treated group and $\mathbf{X} \sim N((\delta, \delta/2, ..., \delta/2), \mathbf{I}_{10})$ in the control group; Sim-3: $\mathbf{X} \sim t_3(\mathbf{0}, \mathbf{I}_{10})$ in the treated group and $\mathbf{X} \sim t_3((\delta, \delta/2, ..., \delta/2), \mathbf{I}_{10})$ in the control group; Sim-4: $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_{10})$ in the treated group and $\mathbf{X} \sim N(\mathbf{0}, \Sigma^*)$ in the control group, where Σ^* is a 10×10 matrix such that $\Sigma^*_{ii} = 1$, $\Sigma^*_{12} = \Sigma^*_{21} = \delta$, and $\Sigma^*_{ij} = 0$ otherwise.

bivariate GFKS outperforms all other methods, since the other tests only focus on marginal distributions. Therefore, GFKS tests, which can be easily generalized to test interactions, is a more powerful approach to evaluate the quality of matched samples.

Another simulation is conducted to compare the listed methods with an omnibus test, the machine learning based classification permutation test (Gagnon-Bartsch and Shem-Tov, 2019). Results are included in Appendix A.2.4.

3.5.2. Do balance checks identify the major problems?

In this section, whether the proposed balance checks in §3.4 can identify the major problems with a high probability is investigated. We consider the same four simulation settings as in §3.5.1. To measure the performance, the empirical probability that the smallest p-value identifies the major problems $\hat{q}$, which is calculated based on 1000 replications, serves as a criterion. Since the machine learning based classification permutation test cannot provide information on where the imbalance is and how to improve the match, the performance of GFKS tests is only compared with two-sample t-test and KS test on individual covariates. Results are summarized in Figure 5(b).

When only marginal distributions are imbalanced, the two GFKS tests and KS test can identify the most imbalanced covariate or a related interaction more robustly. The four tests have similar performance when the null distribution is normal (Sim-1 and Sim-2), while GFKS and KS outperform t-tests when the null distribution is not normal (Sim-3). In Sim-4 where there is no imbalance in the marginal distributions, bivariate GFKS have a larger probability to identify the problematic interaction than all methods focusing on the marginal distributions. In summary, GFKS is a more robust method to identify the major imbalances.

3.5.3. Effects on average treatment effect on the treated

Simulations in this section are conducted to study the effects of the proposed iterative adjustment algorithm with GFKS tests on estimating ATT. We also compare the performance of univariate and bivariate GFKS with that of three weighting methods using a propensity

score estimated by random forests, covariate balance propensity score (CBPS) proposed by Imai and Ratkovic (2014), and minimal weights (MW) proposed by Zubizarreta (2015) with the tuning algorithm in Wang and Zubizarreta (2020). The iterative procedure with GFKS tests is applied to adjust a optimal pair match with a propensity score caliper.

Suppose there are 500 treated individuals and 1,000 controls, and each individual has five observed covariates $\mathbf{X} = (X_1, ..., X_5)$, and two outcomes Y_1 and Y_2. Specifically, let $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_5)$ in the treated group and $\mathbf{X} \sim N((1, 0.2, ..., 0.2), \mathbf{I}_5)$ in the control group. Consider two simulation experiments by setting response as $Y_1 = X_1 + \epsilon_1$ and $Y_2 = X_1^2 + \epsilon_2$, where $\epsilon_1, \epsilon_2 \sim N(0, 1)$ are independent errors. Since both outcomes are related to only X_1, a benchmark is to estimate ATT by matching on X_1. We replicate each simulation 100 times. With the proposed method, a satisfactory match can be obtained within three iterations of adjustments in most replications. The bias, variance and mean squared error (MSE) of the estimated ATT in these 100 replications are summarized in the following table. Results in Table 5 show that when the response surface is linear (Y_1), the proposed algorithm (with univariate and bivariate GFKS) has similar performance as CBPS and SBW, which is better than weighting with a propensity score estimated by random forests. When the response surface is non-linear (Y_2), the proposed algorithm outperforms all three weighting methods. It can be concluded that the proposed algorithm is a more robust approach, since it can achieve good estimates of ATT without any model assumption on the response surface.

3.6. Analysis of Bone Density and Antidepressants

Antidepressants, one of the most commonly prescribed drug classes in the United States, are a popular treatment choice to reduce depression symptoms. Selective serotonin reuptake inhibitors (SSRIs), a first-line therapy for depression, is the most widely used class of antidepressants (Pirraglia et al., 2003) Researchers have found that SSRI use decreases bone density of older people (Diem et al., 2007; Haney et al., 2007) and adolescents (Feuer et al., 2015). To illustrate matching methodology, effects of SSRIs on bone density of adults are examined, with 277 people taking only SSRIs as "treated" and 4,613 people taking neither SSRIs nor tricyclic antidepressants (TCAs) as potential "controls". A 4-to-1

Table 5: Comparison of bias, variance, and mean squared error of estimated average treatment effects on the treated based on 100 replications for matching on X_1 (Benchmark, with package `DiPs`), the proposed iterative algorithm with univariate GFKS (GFKS-1) and bivariate GFKS (GFKS-2), inverse probability weighting with the propensity score estimated by random forests (Random forest, with package `randomForest`), covariate balance propensity score (CBPS, with package `CBPS`) and minimal weights (MW, with package `sbw`) in the simulations. Here, $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_5)$ in the treated group and $\mathbf{X} \sim N((1, 0.2, ..., 0.2), \mathbf{I}_5)$ in the control group. Consider two simulation experiments by setting response as $Y_1 = X_1 + \epsilon_1$ and $Y_2 = X_1^2 + \epsilon_2$, where $\epsilon_1, \epsilon_2 \sim N(0, 1)$ are independent errors.

Y_1	Benchmark	GFKS-1	GFKS-2	Random forest	CBPS	MW
Bias	-0.009	-0.060	-0.060	-0.192	-0.004	-0.000
Variance	0.005	0.006	0.006	0.005	0.005	0.005
MSE	0.005	0.009	0.010	0.042	0.005	0.005
Y_2	Benchmark	GFKS-1	GFKS-2	Random forest	CBPS	SBW
Bias	-0.055	-0.128	-0.126	-0.331	-0.969	-0.407
Variance	0.006	0.008	0.009	0.006	0.031	0.017
MSE	0.009	0.025	0.024	0.116	0.970	0.183

match is built, and bone mineral density and content are analyzed. The data are from National Health and Nutrition Examination Survey (NHANES) 2009-2010. We consider the 19 observed covariates (see Appendix A.2.5 for definitions), essentially those used in the literature (Diem et al., 2007; Feuer et al., 2015; Haney et al., 2007).

The goal is to balance marginal distributions of the 19 covariates and propensity score estimated using a linear logit model., as well as all $\binom{20}{2}$ bivariate distributions. The $M = 20 + \binom{20}{2} = 210$ GFKS statistics consider the orthants based on the values of discrete variables and the $(11\%, 35\%, 65\%, 89\%)$ quantiles of the continuous variables in the treated group, as suggested by Cochran (1968). P-values of GFKS statistics are approximated based on $B = 2,000$ simulated randomized experiments formed from the matched samples. The minimum of the p-values, T^*, is considered as the summary statistic; its p-value is approximated based on the simulated randomized experiments.

3.6.1. Our matches

First consider a basic 4-to-1 match M0 with three standard techniques: minimizing total rank-based Mahalanobis distance, with a 0.2 standard deviation propensity score caliper, and fine balancing on education. From Table 4, the initial match ("Base") is not bad,

greatly reducing the covariate mean imbalances in the original data set. However, the standardized mean differences (SMDs) suggest that there is some remaining imbalance in diabetes. Looking through p-values of the 210 GFKS tests provides more insight into why diabetes is not balanced – the smallest p-value 0.001 occurs at the interaction of diabetes and preference to weigh more, not at the marginal distribution of diabetes. The global test in §3.2 gives a p-value 0.06, which is less than $\alpha = 0.1$.

To get a better match, we find that the most imbalanced orthant is the group of people who don't have diabetes and don't prefer to weigh more. This feature occurs more often for people not taking any antidepressants. A new binary variable for this orthant is constructed and a new match M1 is built nearly finely balancing the interaction education and this binary variable. M1 improves M0 significantly. The p-value of T^* is 0.945, which is greater than $\alpha = 0.1$. The proposed global test implies M1 is adequate. More precisely, M1 is more balanced than 94.5% of the 2,000 simulated randomizations.

However, the SMD of propensity score in M1 increases slightly to 0.105. What causes the imbalance in propensity score? Following the same procedure, the interaction of weight change and propensity score is diagnosed with the smallest p-value $T^* = 0.086$. Specifically, there are more people in the control group than the treated group whose weight change is lower than or equal to 65% of the people taking SSRIs and propensity score is smaller than or equal to 89% of the people taking SSRIs. To adjust for that, we form a new binary variable based on this feature.

A new match M2 is built additionally near-fine balancing this binary variable. The marginal distribution of weight change is detected in M2, and the most imbalanced orthant is the group of people whose weight change is lower than or equal to 35% of the people taking SSRIs. This feature occurs more frequently in the treated group. A new binary variable for this orthant is constructed and nearly finely balanced in a new match M3. In Table 4, all SMDs for M3 ("Iter 3") are below 0.1. Additionally, all p-values are above 0.1, and the p-value for $T^* = 0.129$ is 0.982, which means that M3 is better than 98.2% of the 2,000

randomizations.

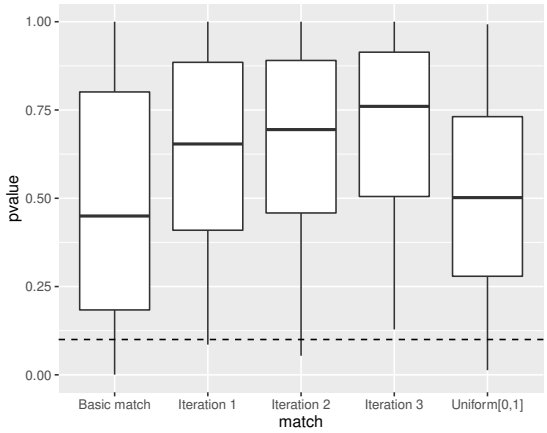
To compare the quality of the four constructed matched samples, Figure 6(a) shows the boxplots of 210 p-values of the four matched samples, along with a uniform distribution on $[0, 1]$. There is an upward trend, and all p-values are above 0.1 for M3. Figure 6(b) and (c) compare the 20 SMDs and the 20 two-sample t statistics in the four matches. M3 has SMDs falling into the ± 0.1 band and two-sample t statistics falling into the ± 2 band. Therefore, M3 is satisfactory if we use the common practices to evaluate its balance. Some other classical two-sample tests can also applied to each individual covariate for M3; the results are summarized in Table 4 in the main paper. It can be concluded that if the common practice of using SMD and two-sample t test, KS test, Wilcoxon rank test on each individual covariate is applied to evaluate the match, the quality of M3 is adequate. Notably, M3 cannot be classified by machine learning methods like random forests. The machine learning based classification permutation test with random forests (Gagnon-Bartsch and Shem-Tov, 2019) gives a p-value 0.998, which suggests the adequacy of the final match.

3.6.2. Outcome analysis

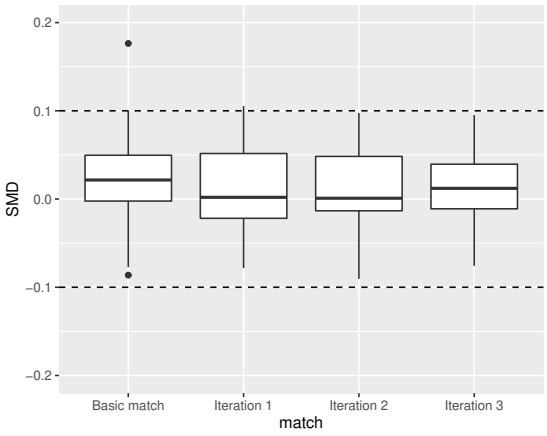
The outcomes are analyzed with match M3. We first consider whether using SSRIs decreases femur bone mineral density and analyze the outcomes in log scale. The inference uses the permutation distribution of Huber's M-statistic, as discussed by Maritz (1979); Rosenbaum (2007). Using the `senm()` and `senmCI()` functions in the package `sensitivitymult`, the one-sided p-value is 0.387 and two-sided 95% confidence interval is [-0.022, 0.016]. Similar analyses are conducted for other outcome variables: femur bone mineral content, femoral neck bone mineral density and femoral neck bone mineral content. If M3 were a randomized experiment, then none of the outcomes has a significant causal effect. Another analysis is conducted combining the four outcomes using Scheffé projections (Rosenbaum, 2016) with the `comparison()` function in the package `sensitivitymult`. With equal weights to these four outcomes, a p-value of 0.999 is obtained, suggesting no significant causal effect.

3.7. Discussion

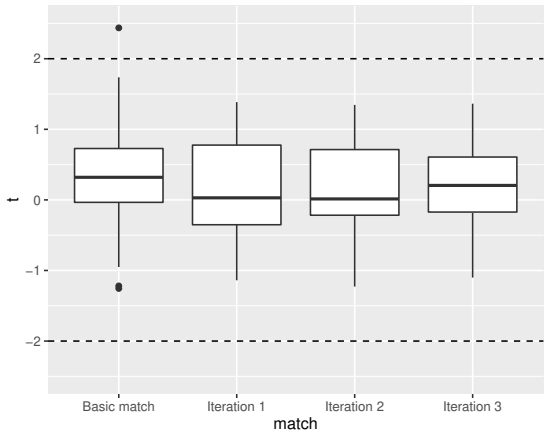
After conducting a matched sample, it is important to assess covariate balance in the re-



(a) 210 p-values of GFKS tests



(b) 20 standardized mean differences



(c) 20 two-sample t statistics

Figure 6: Boxplots of p-values of GFKS tests, standardized mean differences and two-sample t statistics for M0 (Basic match), M1 (Iteration 1), M2 (Iteration 2), M3 (Iteration 3).

sulting matched data. Commonly used informal diagnostics are useful but have limitations. To overcome these limitations, the current paper develops a formal framework for rigorous assessment of covariate balance. Unlike the informal diagnostics, the proposed method regards randomized experiments as a benchmark and uses a summary statistic to compare both marginal distributions and joint distributions of the matched sample to that of simulated randomized experiments built from the same data. In doing so, it can distinguish unsolvable matching problems, problems that can be solved by a better match and those that are small compared to the imbalances that occur by chance in randomized experiments. Moreover, the assessment not only evaluates whether the match is adequate, but also identifies the major problems, which provides guidance on how to improve the covariate balance. Any detected imbalance suggests an improved match that balances the feature distinguishing the two groups. To illustrate this general idea, we consider a simple but useful implementation using the minimum p-value of GFKS tests for the marginal and joint distributions as the summary statistic, and provide theoretical understandings of this implementation. There are many other variations of this general framework, but the implementation with GFKS has some appealing properties and works well as shown in the simulations.

CHAPTER 4 : Directional Penalties for Optimal Matching in Observational Studies

4.1. Introduction: Example, Outline, the Simplest Case

4.1.1. Motivating example: smoking and homocysteine

Bazzano et al. (2003) asked whether cigarette smoking causes elevated levels of homocysteine, a possible risk factor for cardiovascular disease, thrombosis, and Alzheimer's disease; see Hankey and Eikelboom (1999); Seshadri et al. (2002); Welch and Loscalzo (1998). Bazzano et al. (2003) used data from the National Health and Nutrition Examination Survey III (NHANES) conducted in 1988-1994. To illustrate directional penalties, consider an analogous matched comparison using NHANES 2005-2006. Among individuals aged 20 or more, there were 548 daily smokers, called treated, and 2211 never smokers, called controls.

The first two numeric rows of Table 6 show the means for seven covariates before matching, namely female, age, black, education, the ratio of family income to the poverty level capped at five times poverty, BMI or body mass index, and the propensity score estimated from these six covariates using a linear logit model. Education is a 5-point scale, with 1 for less than 9th grade, 3 for high school or equivalent, and 5 for college graduate. For instance, 43% of the 548 daily smokers were female, but 61% of the 2211 never smokers were female. In Table 6, education and the propensity score are abbreviated as Educ and P-score. Table 6 compares the covariate balance before matching to covariate balance in four 2-to-1 matched samples, M1-M4, that are described below.

Table 6 also displays differences in covariate means as a fraction of the standard deviation before matching. A standardized difference in means is the treatment-minus-control difference in means divided by the square root of the equally weighted average of the within-group variances before matching (see Cochran and Rubin, 1973, p.420). In Table 6, the numerator of standardized difference changes from one row to the next, but for each covariate its denominator is the same within each column, namely a pooled standard deviation of the covariate before matching. The smoking and control groups were quite different before

Table 6: Covariate means for 548 daily smokers individuals (Treated), all 2211 never smokers (All), and four matched samples of 1096 never smokers, each matched 2-to-1 to the daily smokers (M1-M4). Also shown are: (i) the corresponding standardized differences in covariate means, treated-minus-control, where the demoninator is an unchanging pooled within group standard deviation before matching), and (ii) the conventional pooled, two sample t-statistics. Standardized differences ≥ 0.1 and t-statistics ≥ 2 in absolute value are in **bold**.

	Sample	Covariates						
	Size	Female	Age	Black	Educ	Income	BMI	P-score
		Covariate Means						
Treated	548	0.43	45.29	0.21	3.04	2.34	27.62	0.24
All Controls	2211	0.61	45.68	0.25	3.40	2.77	29.19	0.19
Controls M1	1096	0.46	45.42	0.22	3.04	2.33	28.56	0.23
Controls M2	1096	0.44	45.45	0.22	3.04	2.36	27.71	0.23
Controls M3	1096	0.42	45.32	0.21	3.04	2.34	27.65	0.24
Controls M4	1096	0.43	45.35	0.21	3.04	2.33	27.63	0.23
		Standardized difference in means, treated-minus-control						
Before matching		**-0.38**	-0.02	**-0.10**	**-0.30**	**-0.27**	**-0.23**	**0.56**
Match M1		-0.06	-0.01	-0.03	0.00	0.01	**-0.13**	**0.13**
Match M2		-0.02	-0.01	-0.03	0.00	-0.01	-0.01	0.07
Match M3		0.01	-0.00	-0.02	0.00	0.00	-0.00	0.05
Match M4		-0.00	-0.00	-0.02	0.00	0.01	-0.00	0.05
		Pooled two-sample t-statistics						
Before matching		**-7.95**	-0.45	**-2.02**	**-5.91**	**-5.56**	**-4.66**	**12.22**
Match M1		-1.12	-0.14	-0.64	0.00	0.22	**-2.52**	**2.43**
Match M2		-0.35	-0.17	-0.64	0.00	-0.20	-0.27	1.36
Match M3		0.14	-0.03	-0.34	0.00	0.07	-0.08	0.94
Match M4		-0.04	-0.07	-0.30	0.00	0.19	-0.01	0.98

matching. Smokers were less often female, less often black, with lower education, lower income, lower BMI, and of course higher propensity scores. The difference on the propensity score was 0.56 standard deviations, a large difference, and the differences in female, education, poverty and BMI were all more than 0.2 standard deviations.

Finally, Table 6 displays two-sample t-statistics using the conventional pooled variance estimate. Unlike the standardized differences, t-statistics are affected by the reduced sample size in the matched comparison, and also the estimated variance changes for a covariate from one match to another. For a binary covariate, the square of the pooled t-statistic is close to the Pearson chi-square statistic testing independence in the 2×2 table

recording treatment × covariate. In a completely randomized experiment, we expect to see a statistically significant imbalance in a covariate at the 0.05 level for one covariate in twenty. In that sense, the t-statistics compare the covariate balance attained by matching to the covariate balance expected in a completely randomized experiment. Judged in this way, the t-statistics for BMI and the propensity score in match M1 are somewhat disappointing.

The rows M1, M2, M3 and M4 in Table 6 describe four 2-to-1 matched samples, so the sample size for a control mean is $1096 = 2 \times 548$ in each of these rows. Match M1 is fairly conventional, using three standard techniques. As suggested in Rosenbaum and Rubin (1985b) M1 includes a caliper on the propensity score, trying to find matches that differ on the propensity score by at most 0.2 times the standard deviation of the propensity score, and within this caliper it matches using a version of the Mahalanobis distance, specifically the robust, rank-based version in Rosenbaum (2010, §8.3). Although we speak of the Mahalanobis "distance," we always mean the quadratic form, not its square root, so this quadratic form is not a norm. Additionally, the five categories of Education are "finely balanced", meaning that the marginal distribution of education is exactly the same in smoker and matched control groups, although individuals may not be paired for education; see Rosenbaum et al. (2007); Zubizarreta (2012); Pimentel et al. (2015b). Match M1 minimizes the total of the rank-based Mahalanobis distances between smokers and their matched controls, with a penalty for violation of the caliper on the propensity score, subject to the constraint of fine balance for education. The match was constructed by finding a minimum cost flow in a network, a combinatorial optimization problem that can be solved quickly, in time proportional to the cube of the sample size. For details, see Rosenbaum (1989, §3.2), Rosenbaum et al. (2007), and Yang et al. (2012).

Match M1 has greatly reduced the imbalance in covariate means. For the covariate Female, the standardized difference in means, with the same denominator, dropped from $-.38$ before matching to -0.06 after matching. For the propensity score, the standardized difference dropped from 0.56 to 0.13. Of course, the standardized difference for Education is

now zero, because the marginal distributions were forced to be identical. Although some investigators might accept match M1, others might regard the residual imbalance in BMI and the propensity score as a bit too large, and might switch to pair or 1-to-1 matching to obtain better balance on the covariates, at the cost of halving the size of the control group. There are several reasons to prefer a 2-to-1 match to a pair match (Rosenbaum, 2013, 2017b, p.222). Can match M1 be improved, tidied up, without switching to pair matching?

Matches M2, M3 and M4 were obtained in almost the same way, with very little additional effort, after slightly adjusting the rank-based Mahalanobis distances using both directional and nondirectional penalties that we describe in the current paper. With no loss in sample size, the balance in matches M2-M4 are generally better than in M1. The matches were constructed one at a time; for instance, dissatisfaction with the standardized differences for BMI and the propensity score in M1 led to M2; then M3 improved M2, and so on. This is typical of the use of directional penalties: they are chosen to fix a problem identified using balance diagnostics applied to a previous match. A Normal distribution contains 95% of its probability on an interval that is ± 2 standard deviations. A standardized difference of 0.05 in match M4 is $0.0125 = 0.05/4 \doteq 1\%$ of that interval.

Table 7 shows the distribution of the $548 \times 2211 = 1,211,628$ treated-control Mahalanobis distances before matching, and the $2 \times 548 = 1096$ such distances within the 1-to-2 matched sets. For comparison, if two independent observations were drawn at random from the same 6-dimensional multivariate Normal distribution (for 6 covariates, excluding the redundant propensity score), then the expected (non-robust) Mahalanobis distance between them is $2 \times 6 = 12$. The distances in Table 7 are much smaller within matched sets, but the one largest distance in match M3 is 13.43, close to the upper quartile before matching. Can this be reduced? Match M4 slightly altered the distances in match M3, discouraging any pairing with a distance of more than 7. Match M4 looks similar to match M3 in terms of balance in Table 6, but the maximum of 1096 distances in match M4 is now 6.98, rather

than 13.43.

Table 7: Robust Mahalanobis distances before and after matching. "All" refers to the 548×2211 distances before matching. For each match, there are 2×548 distances within matched sets in these 2-to-1 matches. Values are the mean, minimum, quartile 1, median, quartile 3 and maximum.

	Mean	Min	Q1	Median	Q3	Max
All Controls	10.52	0.00	6.83	9.99	13.63	48.70
Match M1	1.39	0.00	0.27	0.84	1.58	19.49
Match M2	1.69	0.00	0.48	1.11	2.03	19.49
Match M3	1.68	0.00	0.54	1.15	2.14	13.43
Match M4	1.57	0.00	0.53	1.15	2.13	6.98

Figure 7 depicts four covariates for match M4. Notably, the treated and matched control groups look similar, but the unmatched controls are quite different with more education and income, somewhat higher BMIs and lower propensity scores. For instance, 8.4% of daily smokers and 8.4% of matched controls had at least a BA degree, but 40% of unmatched controls had a BA degree, nearly a 5-fold difference.

Matching has traditionally viewed covariate differences symmetrically, with a 1 unit difference in BMI regarded as the same whether positive or negative. In contrast, a directional penalty views covariate differences asymmetrically, preferring a 1 unit difference that works against the bias to a 1 unit difference that works in support of the bias. In Table 6, the smokers have somewhat lower BMIs, so a directional penalty prefers a control who is 1 unit too low to a control who is 1 unit too high.

Although directional penalties have intuitive appeal, they are also linked to a key mathematical idea in integer programming. Essentially, many directional penalties are Lagrangians when trying to impose a balance constraint on an optimal matching problem. The imbalances in match M1 suggested changes to the Lagrangian, producing match M2, which suggested the changes that produced match M3. Tidying up match M1 to obtain match M4 took a few simple steps and a few minutes at the computer.

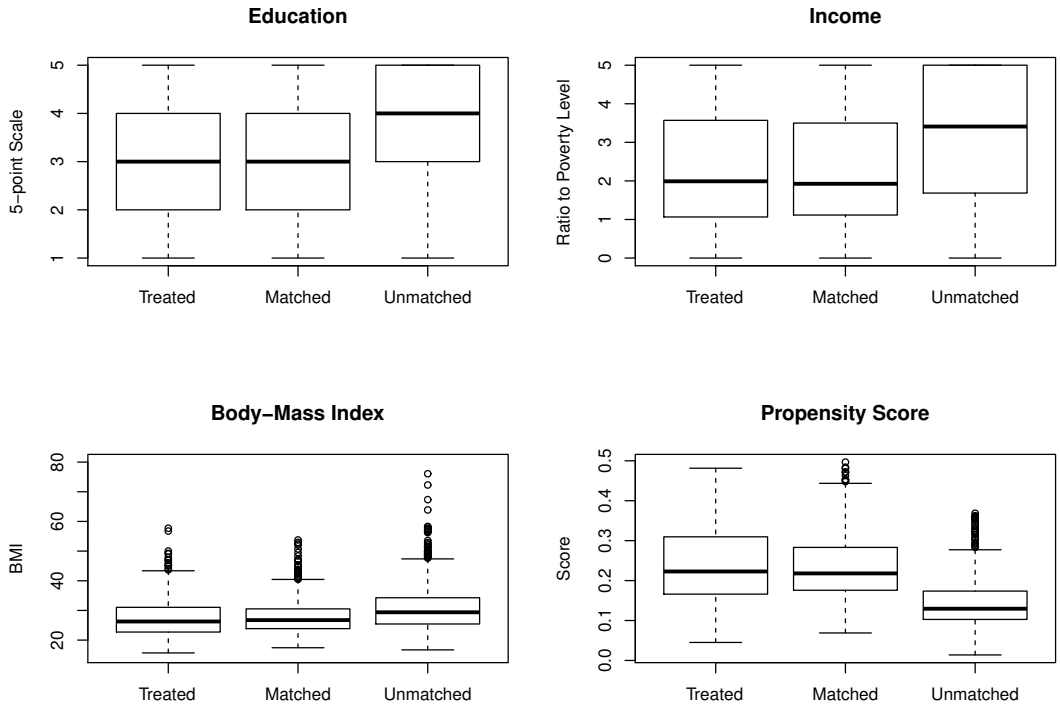


Figure 7: Four covariates in match M4. Treated = daily smokers, matched = matched controls in match M4, and unmatched = never smokers not used in match M4. One unmatched never smoker with a BMI of 130.2 is not displayed. For education, 1 is less than ninth grade, 3 is high school or equivalent, and 5 is a college graduate. Income is recorded as the ratio of family income to the poverty level, capped at 5×poverty. BMI, body mass index

4.1.2. Checking and improving an optimal solution

A theme in much of statistical analysis is that we find an optimal solution of one kind or another, then check whether the optimal solution is satisfactory. This sounds contradictory — how can the optimal solution be inadequate? — but it is not contradictory. For instance, we fit a linear model by least squares or M-estimation, that is, by optimizing a criterion; then, we check whether the fitted model adequately describes the data. The check of adequate fit uses formal regression diagnostics in an informal way, often aided by graphic display. The use of an optimal solution for a particular model is important because we want to move away from that model if the model is inadequate, not if our particular fit

of that model is inadequate — that is one reason we use least squares, not sort-of-small squares. We discard a framework if the best solution within that framework is inadequate.

Similar issues arise in matching. We find an optimal match within certain formal requirements; then, we check whether the match is adequate, and if it is inadequate we change the requirements. The assessment of adequacy uses various formal devices in an informal way. We want the distribution of covariates to be similar in treated and control groups, the individual matched pairs to be individually close on covariates, the pairs to be exactly matched for a few covariates to facilitate a few planned subgroup analyses, and so on. For discussion of the trade-off of close pairs and balanced covariate distributions (see Rosenbaum, 2017b, p.200-204). Matching, unlike modelling, is part of the design of an observational study, so it is done at the beginning of the study without access to outcomes, as in §4.1.1 (see Rubin, 2007).

Here, we are concerned with the actions we might take to improve a matched sample when diagnostic checks have suggested the need for an improved match. This task arises in every matching effort, no matter what methods are used to construct the initial match; yet, there is little written about the task. The question is: How should we change the matching requirements when diagnostics have suggested the initial match is not adequate?

4.1.3. Outline: working asymmetrically against the direction of remaining biases

As further motivation, §4.1.4 considers the simplest way to handle covariate imbalances asymmetrically, namely an asymmetric caliper for a single Normally distributed covariate. Realistic matching problems are discussed in §4.2, where minimum distance matching uses asymmetrically adjusted distances to favor pairs that work against the direction of the bias. Minimum distance matching is briefly reviewed in §4.2.1. Asymmetric linear summaries of covariate imbalance are introduced in §4.2.2, and multiples of these diagnostics are used to adjust the distances in the subsequent match. Sections 4.2.4-4.2.6 observe that directional penalties closely resemble a standard idea in integer programming, namely the optimization of a Lagrangian that attempts to enforce constraints by altering the objective function; see

Fisher (1981). The slightly technical §4.2.6 is largely motivational and heuristic, so it is not essential reading. We return to the example in §4.3, constructing the improved matches M2-M4 and drawing inferences about the effects of smoking. Our methods are implemented in an `R` package `DiPs` available from `CRAN`. The package includes the NHANES data, and examples in the package documentation reproduce analyses described here.

4.1.4. The simplest case: an asymmetric caliper for a Normal covariate

Traditionally, caliper matching for age meant that a treated subject could be matched a control who was either up to $\eta > 0$ years older or $\eta > 0$ years younger, so treated-minus-control difference in age, say $x_t - x_c$, was required to be in $[-\eta,\ \eta]$; see Cochran and Rubin (1973, §2.3) for discussion of this intuitive and venerable technique. For instance, with $\eta = 5$, a 22-year old could be matched to any control with an age in 22 ± 5. Such a caliper is symmetric. If treated subjects are typically older than controls, then an asymmetric caliper of the same length might be preferable, because the residual imbalances within the ± 5 caliper are likely to remain tilted in the original direction. Perhaps we should require $x_t - x_c \in [-\eta_1,\ \eta_2]$ where $\eta_1 \geq 0$ and $\eta_2 \geq 0$ with $\eta_1 \neq \eta_2$. Such a caliper is asymmetric.

In the trivial case of matching two individuals on one Normally distributed covariate, it is easy to demonstrate the advantage of an asymmetric caliper. Suppose $x_t \sim N(\mu,\ 1)$ for a treated individual and $x_c \sim N(0,\ 1)$ for a control, so the bias before matching is $\mathrm{E}(x_t - x_c) = \mu$. Suppose that we sample x_t, then independently sample x_c conditionally given that $x_t - x_c \in [-\eta_1,\ \eta_2]$, or equivalently given that $x_c \in [x_t - \eta_2,\ x_t + \eta_1]$. Exact matching takes $\eta_1 = \eta_2 = 0$, but the probability of an exact match is zero for every control reservoir of finite size, so we must set $\max(\eta_1, \eta_2) > 0$. If $\eta_1 = \eta_2 > 0$, then this is a traditional, symmetric caliper. By numerical integration, we find the bias in caliper matching by finding the conditional expectation of $x_t - x_c$ given the value of x_t and the fact that $x_t - x_c \in [-\eta_1,\ \eta_2]$, then averaging that conditional expectation over the distribution of x_t.

Table 8 shows the bias after matching for various initial biases, μ, and various calipers,

$[-\eta_1, \eta_2]$. When the initial bias is $\mu = 1$, a symmetric caliper $[-1, 1]$ leaves behind a bias of 0.25, but the equally long, asymmetric caliper $[-1.3, 0.7]$ leaves behind a bias of 0.02. Notably, a symmetric caliper, $[-1, 1]$, is best when the initial bias is zero, $\mu = 0$.

Table 8: Bias after matching in one Normally distributed covariate when matching with a possibly asymmetric caliper $x_t - x_c \in [-\eta_1, \eta_2]$. The bias before matching is μ. The smallest absolute bias in each column is in **bold**.

Caliper		Initial bias μ			
$-\eta_1$	η_1	1	0.5	0.25	0
-1.0	1.0	0.25	0.13	0.06	**-0.00**
-1.1	0.9	0.17	0.05	**-0.01**	-0.07
-1.2	0.8	0.09	**-0.02**	-0.08	-0.15
-1.3	0.7	**0.02**	-0.10	-0.16	-0.22
-1.4	0.6	-0.06	-0.17	-0.23	-0.30
-1.5	0.5	-0.14	-0.25	-0.31	-0.37

Table 8 yields several observations. The best asymmetric caliper of length 2 often removes much more bias than the symmetric caliper $[-1, 1]$ of length 2. The best caliper $[-\eta_1, \eta_2]$ of length 2 depends upon the initial bias, μ: the larger the initial bias, the larger the asymmetry needed to offset it. Pick the wrong caliper for a given μ and the bias may increase, rather than decrease. This suggests that asymmetric calipers can remove more bias than symmetric calipers, providing the asymmetric caliper is picked to be appropriate to the initial bias μ that is actually present. These observations apply, not just to calipers, but to asymmetric distances in general.

4.2. Optimal Matching with Asymmetric Adjustments to Distances

4.2.1. Minimum distance matching with $\kappa \geq 1$ controls

There are T treated subjects in a set $\mathcal{T} = \{\tau_1, \ldots, \tau_T\}$, each one is to be matched to $\kappa \geq 1$ controls, and there are $C \geq \kappa T$ potential controls $\mathcal{C} = \{\gamma_1, \ldots, \gamma_C\}$, with $\mathcal{T} \cap \mathcal{C} = \emptyset$ and no control is used in the match more than once. There is a "distance", δ_{tc}, between τ_t and γ_c, with $-\infty < \delta_{tc} \leq \infty$. For instance, δ_{tc} might be the Mahalanobis distance between the observed covariates of τ_t and γ_c; see Rubin (1980). More commonly, δ_{tc} might is set to ∞ if τ_t and γ_c differ by more than a specified caliper on the propensity score, and otherwise δ_{tc} is some form of Mahalanobis distance, as suggested by Rosenbaum and Rubin (1985b).

Write a_{tc} for the binary indicator of who is matched to whom, where $a_{tc} = 1$ if γ_c is one of the κ controls matched to τ_t, and $a_{tc} = 0$ otherwise. Then a match must satisfy: $\sum_{c=1}^{C} a_{tc} = \kappa$ for each t, and $\sum_{t=1}^{T} a_{tc} \leq 1$ for each c, and any such match is called a feasible match. The total distance between treated subjects and their matched controls is $\sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc}\,\delta_{tc}$, and we want this total distance to be small. To speak concisely of all of the a_{tc}, write $\mathbf{a}$ for the $T \times C$ matrix containing the a_{tc}. An optimal or minimum distance match is one that minimizes $\sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc}\,\delta_{tc}$ among all feasible matches $\mathbf{a}$. In `R`, such an optimal match $\mathbf{a}$ may be found using the `pairmatch` function of `optmatch` package (Hansen, 2007), which in turn calls the Fortran subroutine of Bertsekas and Tseng (1988). The `R` package `DiPs` associated with the current paper uses the same software, but includes also directional penalties and near-fine balance constraints.

An optimal match may be found using a minimum cost flow algorithm in computer time that is $O\left(C^3\right)$, that is, relatively quickly. See Bertsekas (1998); Korte and Vygen (2012) for general discussion of minimum cost flow algorithms, and see Rosenbaum (1989, 2010, §8) and Hansen and Klopfer (2006) for their application to matching in observational studies. For other types of optimal matching, see Lu et al. (2011); Kilcioglu and Zubizarreta (2016); Zubizarreta (2012).

Some algorithms require $\delta_{tc} \geq 0$; however, negative δ_{tc} present no real problem. If $-\infty < \min_{tc}\delta_{tc} = \delta_{\min} < 0$, then simply redefine all δ_{tc} to equal $\delta_{tc} - \delta_{\min} \geq 0$, and solve the problem with these redefined distance. The total distance for every feasible solution has been increased by $-\kappa T \delta_{\min} > 0$, so an optimal solution $\mathbf{a}$ with redefined distances is also an optimal solution with the original distances.

4.2.2. Asymmetric linear summaries of covariate imbalance

Each τ_t and γ_c has a value of K functions, x_{k,τ_t} and x_{k,γ_c}, $k = 1, \ldots, K$, of observed covariates, and we would like them to be balanced in treated and control groups. It is convenient to permit repetition; for instance, x_{k,τ_t} and x_{k,γ_c} for $k = 1, 2, 3$ might all refer to age.

Write d_{ktc} for a comparison of x_{k,τ_t} and x_{k,γ_c}, and consider the average d_{ktc} for matched individuals, those with $a_{tc} = 1$, namely

$$\overline{d}_k = \frac{1}{\kappa T} \sum_{t=1}^{T} \sum_{c=1}^{C} a_{tc}\, d_{ktc}. \tag{4.1}$$

The simplest, but arguably not the best, comparison of x_{k,τ_t} and x_{k,γ_c} is $d_{ktc} = x_{k,\tau_t} - x_{k,\gamma_c}$, so that $\overline{d}_k$ in (4.1) becomes the difference between the mean of x_{k,τ_t} for the treated individuals and the mean of x_{k,γ_c} for those controls used in the match, namely

$$\frac{1}{T} \sum_{t=1}^{T} x_{k,\tau_t} - \frac{1}{\kappa T} \sum_{t=1}^{T} \sum_{c=1}^{C} a_{tc}\, x_{k,\gamma_c}. \tag{4.2}$$

Although (4.2) is a natural measure of covariate imbalance, it permits a large positive difference, $x_{k,\tau_t} - x_{k,\gamma_c} \gg 0$, inside one matched set to cancel a large negative difference inside another, and perhaps we prefer to avoid both large differences of either sign, rather than allowing them to cancel. For hockey-stick measures of the form

$$d'_{ktc} = \max(0,\, x_{k,\tau_t} - x_{k,\gamma_c}) \geq 0, \qquad d''_{ktc} = \max(0,\, x_{k,\gamma_c} - x_{k,\tau_t}) \geq 0, \tag{4.3}$$

we have $x_{k,\tau_t} - x_{k,\gamma_c} = d'_{ktc} - d''_{ktc}$ and $|x_{k,\tau_t} - x_{k,\gamma_c}| = d'_{ktc} + d''_{ktc}$. Consider two averages, $\overline{d}'_k$ and $\overline{d}''_k$ in (4.1) with d'_{ktc} and d''_{ktc} in place of $x_{k,\tau_t} - x_{k,\gamma_c}$. Then $\overline{d}'_k + \overline{d}''_k$ is the average absolute difference, or (4.1) with $d_{ktc} = |x_{k,\tau_t} - x_{k,\gamma_c}|$, and $\lambda_1 \overline{d}'_k + \lambda_2 \overline{d}''_k$ can tilt against the direction of bias by adjusting $\lambda_1 \geq 0$, $\lambda_2 \geq 0$.

If $d_{ktc} = \text{sign}\,(x_{k,\tau_t} - x_{k,\gamma_c})$, where $\text{sign}\,(y)$ is 1, 0 or -1 as $y > 0$, $y = 0$, or $y < 0$, then $\overline{d}_k = 0$ in (4.1) if the median of $x_{k,\tau_t} - x_{k,\gamma_c}$ within matched sets is zero. Again, if $\text{sign}\,(x_{k,\tau_t} - x_{k,\gamma_c})$ is split into two parts,

$$d'_{ktc} = 1 \text{ if } x_{k,\tau_t} - x_{k,\gamma_c} > 0, \;\; d'_{ktc} = 0 \text{ otherwise}, \tag{4.4}$$

$$d''_{ktc} = 1 \text{ if } x_{k,\gamma_c} - x_{k,\tau_t} > 0, \;\; d''_{ktc} = 0 \text{ otherwise}, \tag{4.5}$$

then reducing both $\overline{d}_k^{'}$ and $\overline{d}_k^{''}$ in (4.1) to or below 1/2 controls the median difference while expressing a preference for zero difference in place of counterbalancing nonzero differences. Again, $\overline{d}_k^{'}$ and $\overline{d}_k^{''}$ may be given different weights to tilt against the direction of bias.

For two constants, $-\infty \leq -\eta_1 \leq 0 \leq \eta_2 \leq \infty$, define

$$d_{ktc} = \left[\begin{array}{l} 1 \text{ if } x_{k,\tau_t} - x_{k,\gamma_c} < -\eta_1 \\ 0 \text{ if } -\eta_1 \leq x_{k,\tau_t} - x_{k,\gamma_c} \leq \eta_2 \\ 1 \text{ if } x_{k,\tau_t} - x_{k,\gamma_c} > \eta_2 \end{array}\right. . \tag{4.6}$$

Then (4.6) implements a directional caliper; see §4.1.4. If $\overline{d}_k = 0$ with d_{ktc} defined by (4.6), then $x_{k,\tau_t} - x_{k,\gamma_c}$ is between $-\eta_1$ and η_2 within every matched set.

Balancing the means of functions of covariates can balance variances, covariances and quantiles. If the means of both x and x^2 are the same in matched groups, then the sample variance of x is also the same. See Zubizarreta (2012) for several examples.

Minimizing the total matched pair difference, $\sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc}\,\delta_{tc}$, may produce a match with a few large distances, δ_{tc}. If $\rho > 0$ is a moderate value for the distance, then setting $d_{ktc} = 1$ if $\delta_{tc} > \rho$ and $d_{ktc} = 0$ otherwise can be used to reduce or eliminate distances that exceed ρ. Rosenbaum (2017a) proposed a threshold algorithm to minimize $\sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc}\,\delta_{tc}$ subject to $\delta_{tc} \leq \rho$ with ρ picked to be as small as possible; however, it is simpler, and perhaps adequate, to pick a moderate but somewhat arbitrary value for ρ. In match M4 in §4.1.1, $\rho = 7$ was selected as close to the lower quartile of distances, 6.83, among all controls in Table 7.

4.2.3. The proposed method

The proposed method may now be stated. For K comparisons d_{ktc}, $k = 1, \ldots, K$, suppose that we would prefer $\overline{d}_k \leq \epsilon_k$, $k = 1, \ldots, K$ for fixed $\epsilon_k \geq 0$. Pick $\lambda_1 \geq 0$, $\ldots$, $\lambda_K \geq 0$. Define new distances $\delta_{tc}^{*} = \delta_{tc} + \sum_{k=1}^{K} \lambda_k\ (d_{ktc} - \epsilon_k)$ and minimize $\sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc}\,\delta_{tc}^{*}$ subject to $a_{tc} \in \{0, 1\}$ with $\sum_{c=1}^{C} a_{tc} = \kappa$ for each t, and $\sum_{t=1}^{T} a_{tc} \leq 1$ for each c; that is, solve the conventional matching problem in §4.2.1 with revised distances δ_{tc}^{*}. This can be done

quickly, in $O\left(C^3\right)$ steps. Do this a few times, adjusting the λ_k at each step, in an effort to satisfy or approximate $\overline{d}_k \leq \epsilon_k$, $k = 1, \ldots, K$, commonly increasing λ_k if $\overline{d}_k > \epsilon_k$. The four matches in Table 6 resulted from four adjustments to λ_k, as indicated in Table 9.

Table 9: Directional penalties λ_k used in the three matched samples, M1, M2 and M3. For three covariates, Female, Black and BMI, only the directional penalty, (4.4) or (4.5), was used. For the propensity score, a directional penalty, a hockey stick penalty (4.3) and a sometimes asymmetric caliper (4.6) was used. In the hockey stick (4.3) and caliper (4.6), the propensity score was scaled by the denominator of the standardized difference, namely the pooled, equally weighted standard deviation, s, before matching.

	Imbalance measure	λ_k in 4 matches			
Covariate	d_{ktc}	M1	M2	M3	M4
Directional penalty (4.4)					
Female	1 if $x_{kt} - x_{kc} = -1$, 0 otherwise	0	3	4	4
Black	1 if $x_{kt} - x_{kc} = -1$, 0 otherwise	0	1	2	2
BMI	1 if $x_{kt} < x_{kc}$, 0 otherwise	0	2	2	2
Propensity/s	1 if $x_{kt} > x_{kc}$, 0 otherwise	0	1	1	1
Hockey stick (4.3)					
Propensity	$\max(0, x_{kt} - x_{kc})$	0	4	20	20
Directional caliper (4.6)					
Propensity/s	$\eta_1 = -.2$, $\eta_2 = .2$	1000	0	0	0
Propensity/s	$\eta_1 = -.3$, $\eta_2 = .2$	0	1000	0	0
Propensity/s	$\eta_1 = -.5$, $\eta_2 = .15$	0	0	1000	1000
Maximum Mahalanobis distance penalty					
Max δ_{tc}	1 if $\delta_{tc} > 7$, 0 otherwise	0	0	0	1000

Sections 4.2.4-4.2.5 provide some motivation for this method, which is essentially a heuristic use of a few steps of Lagrangian relaxation. Alas, for integer programs, there is no assurance that there are $\lambda_1 \geq 0, \ldots, \lambda_K \geq 0$ such that Lagrangian relaxation will find a match $\mathbf{a}$ that minimizes $\sum_{t=1}^{T} \sum_{c=1}^{C} a_{tc}\, \delta_{tc}$ subject to $\overline{d}_k \leq \epsilon_k$, $k = 1, \ldots, K$, even if such a solution exists; moreover, there is no way to solve general integer programs in polynomial time. Nonetheless, the literature on integer programming makes extensive formal or heuristic use of Lagrangian relaxation, often to good effect. Our experience is that a few reasonable adjustments to the λ_k often tidy up a match, as was true even with M2 in Table 6. A reader uninterested in further motivation might skip §4.2.4-4.2.5, returning to the example in §4.3.

Proprietary, commercial solvers, such as Cplex and gurobi, make an effort, often successful, to solve integer programs. Because they are proprietary, the user does not know exactly what they do, but they likely use Lagrangian relaxation, branch-and-bound algorithms, cutting planes and other techniques. There is no guarantee that these solvers will obtain an optimal solution in practical time, but they often work well. Zubizarreta (2012) uses such solvers with great effect in optimal matching, and this is often an attractive approach. In contrast, our proposal starts with a conventional match that is not quite satisfactory, and tidies it up using a couple of quick heuristic steps of Lagrangian relaxation.

4.2.4. A computationally difficult problem: minimum distance matching with constraints on linear measures of imbalance

Consider finding a κ-to-1 match $\mathbf{a}$ to

$$\text{minimize} \quad \sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc}\,\delta_{tc} \tag{4.7}$$

subject to

$$a_{tc} \in \{0,1\}\,, \quad \sum_{c=1}^{C} a_{tc} = \kappa,\ t = 1,\ldots,T,\ \sum_{t=1}^{T} a_{tc} \leq 1,\ c = 1,\ldots,C, \tag{4.8}$$

$$\overline{d}_k \leq \epsilon_k,\ \text{for } k = 1,\ldots,K, \tag{4.9}$$

where $\epsilon_k \geq 0$ are specified numbers. We may rewrite (4.9) equivalently as

$$\sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc}\,v_{ktc} \leq 0,\ k = 1,\ldots,K \text{ where } v_{ktc} = d_{ktc} - \epsilon_k. \tag{4.10}$$

As noted in §4.2.1, the problem (4.7)-(4.8) without (4.9) can be solved by an algorithm that runs in polynomial time, that is, relatively quickly. One can solve (4.7)-(4.8) in $O\left(C^3\right)$ computational steps; moreover, if the number of finite δ_{tc} is of order $O\left\{C\log\left(C\right)\right\}$ then the solution can be found faster in $O\left\{C^2\log\left(C\right)\right\}$ steps; see Korte and Vygen (2012, Theorem 11.2).

In contrast, the problem (4.7)-(4.9) is a general integer program and can be much more difficult to solve. In general, integer programs cannot be solved in polynomial time; more precisely, the general integer programming problem is NP-complete (Schrijver, 1998, Theorem 18.1). Various techniques are used to approximate the solution to (4.7)-(4.9). Lagrangian relaxation is often a component of approximate solutions, and it entails solving several easy problems (4.7)-(4.8) with changes to the objective function; see Bertsekas (1998, §10.3), Fisher (1981), Korte and Vygen (2012, §5.6) or Wolsey (1998, §10). It turns out that the informal technique of directional penalties is an instance of Lagrangian relaxation, as seen in §4.2.5.

Write α_A for the minimum value of $\sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc}\,\delta_{tc}$ in the assignment problem (4.7)-(4.8), and write α_B for the minimum of $\sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc}\,\delta_{tc}$ in the problem with balance constraints (4.7)-(4.9). Because of the added constraints (4.9), $\alpha_A \leq \alpha_B$.

The problem (4.7)-(4.9) can be embellished with certain additional features commonly found in matching problems, without changing the basic structure. We do not add these features here, because they make the discussion more complex, not more enlightening. If fine balance constraints or near-fine balance constraints are added to (4.8) as hard constraints, as in Rosenbaum (1989, §3.2) and Yang et al. (2012), then the revised (4.7)-(4.8) can still be solved in $O\left(C^3\right)$ steps, and the revised (4.7)-(4.9) is still much harder, and all of the same issues and techniques apply. Matches M1-M4 in §4.1.1 imposed a fine balance constraint on education. The `R` package `DiPs` includes directional penalties and near-fine balance constraints.

4.2.5. Directional penalties as Lagrangians in a relaxation of the linear imbalance constraints

If the kth x is imbalanced, if $\overline{d}_k > \epsilon_k$, or equivalently if $\sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc}\,v_{ktc} > 0$ in (4.9), then a directional penalty adjusts some of the distances, δ_{tc}, in an effort to reduce $\overline{d}_k$; then, it solves (4.7)-(4.8) with these revised distances, and checks to see where things now stand. That is, if our current matched sample has failed to balance the proportion of people over

age 60 so $\overline{d}_k > \epsilon_k$, we increase the distance δ_{tc} whenever τ_t is over age 60 and γ_c is under age 60, and we match again with these new penalized distances. The current section shows that this process is a version of Lagrangian relaxation.

Fix $\lambda_k \geq 0$, $k = 1, \ldots, K$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)^T$, and define the Lagrangian $L(\mathbf{a}, \boldsymbol{\lambda})$ to be

$$L(\mathbf{a}, \boldsymbol{\lambda}) = \sum_{t=1}^{T} \sum_{c=1}^{C} a_{tc}\ \delta_{tc} + \sum_{k=1}^{K} \lambda_k \sum_{t=1}^{T} \sum_{c=1}^{C} a_{tc}\, v_{ktc},$$

and consider the problem of finding $\mathbf{a}$ to:

$$\text{minimize } L(\mathbf{a}, \boldsymbol{\lambda}) \text{ subject to the constraints (4.8).} \tag{4.11}$$

This problem (4.11) is commonly said to be a Lagrangian relaxation of the problem (4.7)-(4.9). Importantly, if we replace δ_{tc} by $\delta^*_{tc} = \delta_{tc} + \sum_{k=1}^{K} \lambda_k\, v_{ktc}$, then $L(\mathbf{a}, \boldsymbol{\lambda}) = \sum_{t=1}^{T} \sum_{c=1}^{C} a_{tc}\ \delta^*_{tc}$, so problem (4.11) becomes an instance of the matching problem (4.7)-(4.8) with revised distances δ^*_{tc}, so (4.11) can also be solved quickly, in $O\left(C^3\right)$ time.

4.2.6. A standard result: connecting the Lagrangian and the linear imbalance constraints

Proposition 11 is a standard result in integer programming; see, for instance, Bertsekas (1998, §10.3) or Wolsey (1998, §10.1). Write $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)^T \geq (0, \ldots, 0)^T = \mathbf{0}$ if $\lambda_k \geq 0$ for $k = 1, \ldots, K$. Let $\beta_{\boldsymbol{\lambda}}$ be the minimum value of the Lagrangian, $L(\mathbf{a}, \boldsymbol{\lambda})$, in (4.11). Proposition 11 says several things. First, it says that the two problems we can solve are lower bounds for the problem we cannot solve. Second, it indicates that if the solution to (4.11) is an optimal solution to (4.7)-(4.9), then we can easily recognize it as such, simply checking a few conditions. Finally, Proposition 11 suggests that, if we solve (4.11) several times with different $\boldsymbol{\lambda} \geq \mathbf{0}$, then we should reduce λ_k whenever the previous solution $\mathbf{a}$ satisfies $\sum_{t=1}^{T} \sum_{c=1}^{C} a_{tc}\, v_{ktc} < 0$.

Proposition 11 *For all* $\boldsymbol{\lambda} \geq \mathbf{0}$*,*

$$\beta_{\boldsymbol{\lambda}} \leq \alpha_B \text{ and } \alpha_A \leq \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \beta_{\boldsymbol{\lambda}} \leq \alpha_B. \tag{4.12}$$

If $\overline{\mathbf{a}}$ *is an optimal solution of (4.11) for a specific* $\boldsymbol{\lambda} \geq \mathbf{0}$*, if the balance constraints (4.9) are satisfied by this* $\overline{\mathbf{a}}$*, and if* $\overline{d}_k = \epsilon_k$ *whenever* $\lambda_k > 0$*, then* $\overline{\mathbf{a}}$ *is an optimal solution of (4.7)-(4.9) and* $\alpha_B = \beta_{\boldsymbol{\lambda}}$.

4.2.7. Adjusting λ_k *in light of Proposition 11*

Proposition 11 has both formal and heuristic uses. One formal use applies the bound in Proposition 11 to a branch-and-bound search, but this is likely to be impractical in large statistical matching problems. The heuristic uses simply suggest the kinds of adjustments to penalties, λ_k, that might move a match in the direction of satisfying linear constraints (4.9). In the current context, it is the heuristic value of Proposition 11 that matters.

We hope to achieve the targeted balance of $\overline{d}_k \leq \epsilon_k$ for each k. Informally, Proposition 11 suggests that we increase λ_k if $\overline{d}_k > \epsilon_k$ and decrease λ_k if $\overline{d}_k < \epsilon_k$. Proposition 11 suggests that: (i) we want no directional penalty, $\lambda_k = 0$, if the balance we want, $\overline{d}_k \leq \epsilon_k$, occurs naturally without penalization, or (ii) if there is imbalance, $\overline{d}_k > \epsilon_k$, then we want the λ_k that just barely gets the job done, $\overline{d}_k = \epsilon_k$. Proposition 11 says that if an optimal solution of (4.7)-(4.9) can be produced by penalization, then that solution will have the form just described.

4.3. Smoking and Homocysteine

4.3.1. Improving a matched sample in a few simple steps

Section §4.2.5 concluded that we can quickly optimize the Lagrangian by altering the covariate distances from δ_{tc} to $\delta^*_{tc} = \delta_{tc} + \sum_{k=1}^{K} \lambda_k \, v_{ktc}$. Proposition 11 leads us to adjust the λ_k to move towards satisfying the linear imbalance constraints. Table 9 shows how this was done in matches M1, M2, M3 and M4 of §4.1.1.

As noted in §4.1.1, match M1 was built in a conventional way, minimizing the total distance within pairs, subject to a symmetric caliper on the propensity score, requiring fine balance of education. Match M1 was not terrible, but Table 6 revealed some remaining imbalance in BMI and the propensity score, and much smaller imbalances in Female and Black.

Match M2 sought improvements in the following way. An asymmetric caliper on the

propensity score replaced the symmetric caliper with a large penalty, λ_k, to enforce it. This means that we tolerate a smaller difference in propensity scores when the treated subject has the higher propensity score, and a larger difference when the control has the higher propensity score; that is, we tolerate larger differences that work against the direction of the bias. Also, there is a hockey-stick penalty on the propensity score with a moderate λ_k. Finally, small sign penalties are imposed for Female, Black, BMI and the propensity score. In Table 6, the resulting match M2 is a big improvement, perhaps satisfactory as it is. Match M3 sought and obtained slight further improvements, increasing the asymmetry of the propensity score caliper, and slightly increasing the pressure on Female, Black and the propensity score. In terms of balance in Table 6, match M3 looks quite good.

Table 7 revealed that, within most matched sets, the rank-based Mahalanobis distance between treated and control individuals was small. Remember that, in the absence of bias, for multivariate Normal covariates using usual Mahalanobis distance, the expected distance between two randomly chosen individuals is twice the number of covariates, here $2 \times 6 = 12$. The mean covariate distance is a constant multiple of its total, and the optimization algorithm is attempting to minimize the total subject to constraints. Matches M1 to M4 look similar in terms of the mean and quartiles of covariate distance, but the maximum distance in matches M1 to M3 is not small. As discussed in §4.2.2, match M4 severely penalized covariate distances above 7, so none occur in the match, but aside from this improvement match M4 looks similar to match M3. Although Figure 7 is included only for match M4, it was examined for each match, as were tables of counts for discrete covariates.

4.3.2. Analysis of homocysteine and smoking

In match M4, plotting the homocysteine levels of smokers and matched controls suggested estimating effects on the log scale, that is, as a multiplicative effect. Even on the log scale, the distribution of log-homocysteine levels is long-tailed. Inferences used Huber's M-statistics, with his recommended ψ-function, whose influence function resembles a trimmed mean. Randomization inferences and sensitivity analyses were done for point estimates,

confidence intervals and P-values, as described in Rosenbaum (2013), and using the `senmCI` and `amplify` functions in the `sensitivitymult` package in `R`, with the default settings.

Assuming match M4 removed all biases, observed and unobserved, randomization inference gives an estimated multiplicative increase in homocysteine levels from smoking of $\exp(0.114) = 1.121$ or 12%, with 95% confidence interval of 8.8% to 15.5%, and one-sided P-value testing no effect of 9.7×10^{-14}. There is no reason to believe that matching removed unobserved bias in treatment assignment — after all, this is not a randomized experiment. An unobserved covariate that doubles the odds of smoking and doubles the odds of a higher homocysteine in a matched pair would yield a longer 95% confidence interval of [1.053, 1.194] or between 5.3% and 19.4%, with maximum one-sided P-value testing no effect of 9.3×10^{-8}. Allowing for a small bias in treatment assignment leaves a substantial treatment effect, more than a 5% increase. For the two-sided confidence interval to include 1 or no effect would take a bias in treatment assignment that triples the odds of smoking and more than triples the odds of a higher level of homocysteine in a matched pair. In words, only a moderately large bias in treatment assignment could explain away the entire effect. (These statements refer $\Gamma = 1.25$ and $\Gamma = 1.75$ by way of the `amplify` function.)

4.4. Discussion and Extensions

Large penalties have often been used in matching in observational studies in an effort to satisfy some condition as often as possible. For instance, in a conventional way, match M1 used a large penalty to enforce a symmetric caliper on the propensity score. In parallel, near-exact or almost exact matching uses a large penalty to match exactly as often as possible for some nominal covariate, such as gender; see Rosenbaum (2010, §9.2).

In contrast, because directional penalties work against the direction of bias, they can go too far, for instance replacing a large positive bias by a large negative bias. For this reason, directional penalties must be adjusted to achieve desired effects. The adjustment may combine a large penalty with adjustments to the degree of asymmetry of an asymmetric caliper. Alternatively, the adjustment may entail adjustments to the magnitudes of small

penalties, λ_k, in a manner consistent with the heuristic advice provided by Proposition 11. In either case, a few adjustments to directional penalties may quickly improve a match, as we saw in matches M1-M4. We judge a match to be improved by considering a collection of informal diagnostics, including balance summaries, comparisons of within-match covariate distances, boxplots of covariate distributions, and many other related summaries.

Recently, methods have been proposed for matching large administrative data bases consisting of hundreds of thousands of people; see Yu et al. (2020). The directional penalties proposed here may be used with those methods, as they simply adjust covariate distances.

We have discussed matching each treated individual to the same number $\kappa \geq 1$ of controls. When matching with a variable number of controls, say half 1-1 pairs and half 1-2 triples, Pimentel et al. (2015b) argue that the pairs should be balanced for covariates and separately the triples should be balanced for covariates, so that every weighted combination of the pairs and the triples is balanced for covariates. They decide whether a treated subject will be in a pair or a triple using Yoon's entire number derived from the estimated propensity score. In parallel, the directional penalties we describe may be applied twice, once to balance the pairs, separately to balance the triples.

APPENDICES

A.1. Appendix for Chapter 2: Proofs of Main Results

Proof of Proposition 2

Recall that $\upsilon : \mathcal{T} \cup \mathcal{C} \to \{1, \ldots, \Upsilon\}$ and $T = |\mathcal{T}| = |\mathcal{M}|$. So $T = \sum_{k=1}^{\Upsilon} |\{\tau \in \mathcal{T} : \upsilon(\tau) = k\}|$ and $T = \sum_{k=1}^{\Upsilon} |\{\gamma \in \mathcal{M} : \upsilon(\gamma) = k\}|$; hence $0 = \sum_{k=1}^{\Upsilon} d_k$. Trivially, $d_k = \max(0, d_k) + \min(0, d_k)$; so, $0 = \sum_{k=1}^{\Upsilon} d_k$ implies $\sum_{k=1}^{\Upsilon} \max(0, d_k) = -\sum_{k=1}^{\Upsilon} \min(0, d_k)$. Trivially, $|d_k| = \max(0, d_k) - \min(0, d_k)$, so that $\sum_{k=1}^{\Upsilon} |d_k| = \sum_{k=1}^{\Upsilon} \max(0, d_k) - \sum_{k=1}^{\Upsilon} \min(0, d_k) = 2\sum_{k=1}^{\Upsilon} \max(0, d_k)$.

Proof of Proposition 4

Let $\mu : \mathcal{T} \to \mathcal{C}$ be a match in $\mathcal{B}$, so μ is a 1-1 function, and let $\mathcal{M} \subset \mathcal{C}$ be the image of μ, so $\mathcal{M}$ is the subset of $T = |\mathcal{T}| = |\mathcal{M}|$ controls who are matched. We construct a feasible flow $f(\cdot)$ from $\mu(\cdot)$. For $(\tau, \gamma) \in \mathcal{B}$, set $f\{(\tau, \gamma)\} = 1$ if $\mu(\tau) = \gamma$, and set $f\{(\tau, \gamma)\} = 0$ otherwise. Set $f\{(\gamma, \gamma')\} = 1$ if $\gamma \in \mathcal{M}$ and set $f\{(\gamma, \gamma')\} = 0$ if $\gamma \in \mathcal{C} - \mathcal{M}$. Set $f\{(\gamma', \beta)\} = 1$ if $\gamma \in \mathcal{M}$ and set $f\{(\gamma', \beta)\} = 0$ if $\gamma \in \mathcal{C} - \mathcal{M}$. Set $f\{(\beta, \sigma)\} = T$. Set $f\{(\gamma', \upsilon)\} = 0$ for $\upsilon = 1, \ldots, \Upsilon$. This flow satisfies (2.4), (2.5), (2.6), and (2.7), so it is a feasible flow in $(\mathcal{N}, \mathcal{E})$. Conversely, let $f(\cdot)$ be a feasible flow in the matching network $(\mathcal{N}, \mathcal{E})$. Because $f(\cdot)$ is feasible with $\text{cap}\{(\tau, \gamma)\} = 1$ for each $(\tau, \gamma) \in \mathcal{B}$ and $\text{div}(\tau) = 1$ for each $\tau \in \mathcal{T}$, it follows that for each $\tau \in \mathcal{T}$ there exists a $\gamma \in \mathcal{C}$ such that $f\{(\tau, \gamma)\} = 1$. Define $\mu(\tau) = \gamma$ if $f\{(\tau, \gamma)\} = 1$; so, we have just shown that $\mu : \mathcal{T} \to \mathcal{C}$ is a function. To complete the proof, we need to show that $\mu(\cdot)$ is a 1-1 function. Fix $\gamma \in \mathcal{C}$; then, because $\text{cap}\{(\gamma, \gamma')\} = 1$, it follows that $f\{(\gamma, \gamma')\} \leq 1$, so that $\sum_{\tau:(\tau,\gamma)\in\mathcal{B}} f\{(\tau, \gamma)\} \leq 1$; so, there is at most one $\tau \in \mathcal{T}$ such $f\{(\tau, \gamma)\} = 1$.

Proof of Proposition 5

Let $f(\cdot)$ be a minimum cost feasible flow in $(\mathcal{N}, \mathcal{E})$, and let $g(\cdot)$ be any feasible flow in $(\mathcal{N}, \mathcal{E})$, so $\text{cost}(f) \leq \text{cost}(g)$. First, we show that the bypass flow is smaller for $f(\cdot)$, or more precisely, we show $f\{(\beta, \sigma)\} \leq g\{(\beta, \sigma)\}$. Let $h(\cdot)$ be any feasible flow. Using (2.8) and $\sum_{(\tau,\gamma)\in\mathcal{B}} \delta(\tau, \gamma) < \Psi$, it is seen that $\text{cost}(h) = \sum_{e\in\mathcal{E}} \text{cost}(e)\, h(e)$ is bounded above

and below by

$$h\{(\beta,\sigma)\}\cdot\Psi \leq \text{cost}(h) \leq h\{(\beta,\sigma)\}\cdot\Psi + \sum_{(\tau,\gamma)\in\mathcal{B}} \delta(\tau,\gamma) < [h\{(\beta,\sigma)\}+1]\cdot\Psi, \qquad \text{(A.1)}$$

and in particular this is true of feasible flows $f(\cdot)$ and $g(\cdot)$. Because they are feasible flows, $f(\cdot)$ and $g(\cdot)$ take nonnegative integer values (2.4). If $g\{(\beta,\sigma)\} < f\{(\beta,\sigma)\}$ then $f\{(\beta,\sigma)\} \geq g\{(\beta,\sigma)\}+1$, and using (A.1)

$$\text{cost}(g) < [g\{(\beta,\sigma)\}+1]\cdot\Psi \leq f\{(\beta,\sigma)\}\cdot\Psi \leq \text{cost}(f),$$

which is impossible because $f(\cdot)$ is a minimum cost feasible flow; so, we conclude $f\{(\beta,\sigma)\} \leq g\{(\beta,\sigma)\}$. In brief, a minimum cost feasible flow minimizes the bypass flow, $f\{(\beta,\sigma)\}$. Because $T = h\{(\beta,\sigma)\} + \sum_{k=1}^{\Upsilon} h\{(k,\sigma)\}$ for every feasible flow $h(\cdot)$, a minimum cost feasible flow $f(\cdot)$ has maximized $\sum_{k=1}^{\Upsilon} f\{(k,\sigma)\}$ and minimized $\sum_{k=1}^{\Upsilon} [\text{cap}\{(k,\sigma)\} - f\{(k,\sigma)\}]$. Recall from (2.2) and (2.7), that $d_k = |\{\tau\in\mathcal{T} : \upsilon(\tau)=k\}| - |\{\gamma\in\mathcal{M} : \upsilon(\gamma)=k\}|$ may be written

$$d_k = \text{cap}\{(k,\sigma)\} - |\{\gamma\in\mathcal{M} : \upsilon(\gamma)=k\}|. \qquad \text{(A.2)}$$

If $\text{cap}\{(k,\sigma)\} = f\{(k,\sigma)\}$ for $k = 1,\ldots\Upsilon$, then $0 = \sum_{k=1}^{\Upsilon} |d_k|$, so $\sum_{k=1}^{\Upsilon} |d_k|$ is minimized, as required. Otherwise, consider a fine balance category k with $f\{(k,\sigma)\} < \text{cap}\{(k,\sigma)\}$. If there were at least one $\gamma\in\mathcal{C}$ such that $\upsilon(\gamma)=k$ and $f\{(\gamma',\beta)\}=1$, then we could reduce the cost of $f(\cdot)$ by $\Psi>0$ by redefining $f\{(\gamma',\beta)\}=0$, $f\{(\gamma',k)\}=1$ and increasing $f\{(k,\sigma)\}$ by 1, thereby contradicting the fact that $f(\cdot)$ is a minimum cost flow; so, there is no $\gamma\in\mathcal{C}$ such that $\upsilon(\gamma)=k$ and $f\{(\gamma',\beta)\}=1$, and therefore $|\{\gamma\in\mathcal{M} : \upsilon(\gamma)=k\}| = f\{(k,\sigma)\}$. Hence, using (A.2), if $f\{(k,\sigma)\} < \text{cap}\{(k,\sigma)\}$, then $d_k > 0$. If $f\{(k,\sigma)\} = \text{cap}\{(k,\sigma)\}$, then $d_k \leq 0$. Since we have minimized $\sum_{k=1}^{\Upsilon} [\text{cap}\{(k,\sigma)\} - f\{(k,\sigma)\}] = \sum_{k=1}^{\Upsilon} \max(0,d_k)$, we have minimized $\sum_{k=1}^{K} |d_k|$ by Proposition 2. In brief, we have shown

that minimizing $f\{(\beta,\sigma)\}$ is equivalent to minimizing $\sum_{k=1}^{K}|d_k|$, and every minimum cost feasible flow $f(\cdot)$ minimizes $f\{(\beta,\sigma)\}$. The cost of any feasible flow $h(\cdot)$ is $\sum_{(\tau,\gamma)\in\mathcal{B}}\delta(\tau,\gamma)\ h\{(\tau,\gamma)\}+h\{(\beta,\sigma)\}\cdot\Psi$; so, if $f\{(\beta,\sigma)\}=g\{(\beta,\sigma)\}$ with $\text{cost}(f)\leq\text{cost}(g)$, then it follows that $\sum_{(\tau,\gamma)\in\mathcal{B}}\delta(\tau,\gamma)\ f\{(\tau,\gamma)\}\leq\sum_{(\tau,\gamma)\in\mathcal{B}}\delta(\tau,\gamma)\ g\{(\tau,\gamma)\}$, and the match $\mu(\cdot)$ obtained from $f(\cdot)$ has minimized $\sum_{\tau\in\mathcal{T}}\delta\{\tau,\mu(\tau)\}$, as required among matches that minimize $\sum_{k=1}^{\Upsilon}|d_k|$.

Proof of Proposition 6

A minimum cost flow problem of the type in Proposition 5 has a worst case time bound of $O\left\{|\mathcal{N}|\cdot|\mathcal{E}|+|\mathcal{N}|^2\cdot\log(|\mathcal{N}|)\right\}$; see Korte and Vygen (2012, Theorem 9.13) with the simplification that their B equals T in Proposition 5. The bipartite graph $(\mathcal{T}\cup\mathcal{C},\mathcal{B})$ contains $T+C=O(C)$ nodes because $T\leq C$, and at most νT edges in $\mathcal{B}$. The part of the network $(\mathcal{N},\mathcal{E})$ excluding $(\mathcal{T}\cup\mathcal{C},\mathcal{B})$ contains C duplicate nodes γ', and $\Upsilon+2$ auxiliary nodes, namely $1,\ldots,\Upsilon$, β, σ, or $O(C)$ nodes in total. It also contains C edges (γ,γ'), C edges (γ',β), C edges $\{\gamma',\upsilon(\gamma')\}$, $\Upsilon\leq C$ edges (k,σ), and one edge (β,σ), or $O(C)$ edges in total. So, putting the two parts together, $|\mathcal{N}|=O(C)$ and $|\mathcal{E}|=O(\nu T+C)=O(\nu C)$ again using $T\leq C$, so the result follows.

A.2. Appendix for Chapter 3

A.2.1. Using Completely Randomized Experiments as the Benchmark

In the current paper, we regard complete randomization as the benchmark, rather than paired randomized experiments. This is because the balance produced by complete randomization is a fixed benchmark or ruler that does not change as the match changes. On the other hand, if one uses the balance in paired randomization as a ruler, the ruler would change its scale for different matches. It would be the best that the ruler could remain the same size as measuring different things. In particular, one would like to conclude that balance is improved if the means of the matched treated and control groups become much closer. However, when the means of the two groups become closer after matching, the paired t-statistic (as in a paired randomized experiment) may increase, since matching also reduces the within-pair variance – the ruler is changing length as the object being measured

changes. To illustrate this point, consider a simulated age variable which follows $N(30, 20)$ in the treated group (sample size 500) and $N(40, 20)$ in the control group (sample size 1,000). Optimal pair matching with a propensity score caliper reduces the mean difference in age from 10 years to less than 0.5 years in this simulated example; yet it also reduces the within-pair standard deviation in age, so a difference of less than 0.5 years appears to be a large difference by that standard (a significantly large paired t-statistic of -4.238 is obtained). Worse than that, if we take two close pairs with ages (20.1, 20) and (50.1, 50) differing by 0.1 years, swap their controls to form (20.1, 50) and (50.1, 20), then the imbalance in mean is the same as before, but the within-pair standard deviation increases, so the paired t-statistic decreases. That is, making the match worse seems to make it better if the paired t-statistic is the standard. Therefore, a fixed ruler with completely randomized experiments is preferred to evaluate covariate balance of matched samples.

On the other hand, exact balance is a fixed ruler to measure covariate balance of a matched sample. We prefer to use the balance of completely randomized experiments instead of exact balance as the benchmark, since even a randomized experiment, where there is no systematic difference in both observed and unobserved covariates between the treated and control groups, would produce some imbalances by chance, particularly if there are many covariates. Comparison with randomization distinguishes imbalances which are large from imbalances that would be regarded as inconsequential in a randomized experiment.

A.2.2. Proofs of Main Results

Proof of Proposition 8

Proposition 8 can be proved using a continuous analogue of the discrete random variables, as proposed by Fligner and Wolfe (1976).

Let F^* be a continuous cumulative distribution function which agrees with F at $\{v_i\}$. If X_j^*, $j = 1, ..., B$, and Y^* are independent random samples from F^*, then $X_j = \min_{v_i \geq X_j^*} v_i$, $j = 1, ..., B$, and $Y = \min_{v_i \geq Y^*} v_i$ are independent random samples from F.

With this definition, $X^*_{(j)} \leq Y^*$ implies $X_{(j)} \leq Y$. Hence $\mathbb{P}(X^*_{(j)} \leq Y^*) \leq \mathbb{P}(X_{(j)} \leq Y)$, and this gives

$$\mathbb{P}(Y < X_{(\alpha(B+1))}) \leq \mathbb{P}(Y^* < X^*_{(\alpha(B+1))}) = \alpha.$$

Similarly, since $Y^* \leq X^*_{(j)}$ implies $Y \leq X_{(j)}$, we have

$$\mathbb{P}(Y \leq X_{(\alpha(B+1))}) \geq \mathbb{P}(Y^* \leq X^*_{(\alpha(B+1))}) = \alpha.$$

On the other hand, since $\epsilon \geq \max q_i$,

$$\mathbb{P}(Y \leq X_{(\alpha(B+1))}) \leq \alpha + \epsilon.$$

Lemma 1 and Proof

Lemma 1 For any finite collection of sets $\mathcal{Q}$, define

$$T = \sqrt{\frac{nm}{n+m}} \max_{Q \in \mathcal{Q}} |\mathbb{P}_n(Q) - \mathbb{P}_m(Q)|.$$

Then, the Bahadur slope of T is

$$\frac{4\tau}{(\tau+1)^2} \max_{Q \in \mathcal{Q}} (\mathbb{P}_T(Q) - \mathbb{P}_C(Q))^2.$$

Proof of Lemma 1 Since $\mathcal{Q}$ is finite and components in $\mathcal{Q}$ can be any shape in $\mathbb{R}^k$, the traditional technique of using continuous Gaussian processes (Anderson and Darling, 1952; Abrahamson, 1967) cannot be applied here. Here, a new proof is demonstrated.

Let $K := |\mathcal{Q}|$ denote the cardinality of $\mathcal{Q}$.

Under the null hypothesis, both $\mathbb{P}_T$ and $\mathbb{P}_C$ equal some unknown $\mathbb{P}_0 \in \mathcal{P}$, where $\mathcal{P}$ denotes the collection of all possible probability distributions of covariates $\mathbf{X}$. The attained level $L_n(T) = \varphi_n(T)$, where $\varphi_n(x) = \sup_{\mathbb{P}_0 \in \mathcal{P}} \mathbb{P}_0(T' > x)$ and T' is the statistic when the data are drawn under H_0.

For any $t > 0$,

$$\begin{aligned} L_n(t) &= \sup_{\mathbb{P}_0 \in \mathcal{P}} \mathbb{P}_0 \left(\sqrt{\frac{nm}{n+m}} \max_{Q \in \mathcal{Q}} |\mathbb{P}_n(Q) - \mathbb{P}_m(Q)| > t \right) \\ &\leq \sup_{\mathbb{P}_0 \in \mathcal{P}} \sum_{Q \in \mathcal{Q}} \mathbb{P}_0 \left(\sqrt{\frac{nm}{n+m}} |\mathbb{P}_n(Q) - \mathbb{P}_m(Q)| > t \right) \\ &\leq K \sup_{\mathbb{P}_0 \in \mathcal{P}} \max_{Q \in \mathcal{Q}} \mathbb{P}_0 \left(\sqrt{\frac{nm}{n+m}} |\mathbb{P}_n(Q) - \mathbb{P}_m(Q)| > t \right). \end{aligned}$$

Let $Z_Q \sim N(0, \mathbb{P}_0(Q)(1 - \mathbb{P}_0(Q)))$ and Φ denote the CDF of $N(0,1)$. We have

$$\begin{aligned} \lim_{n \to \infty} L_n(t) &\leq \lim_{n \to \infty} K \sup_{\mathbb{P}_0 \in \mathcal{P}} \max_{Q \in \mathcal{Q}} \{\mathbb{P}_0(|Z_Q| > t)\} \\ &= 2K \sup_{\mathbb{P}_0 \in \mathcal{P}} \max_{Q \in \mathcal{Q}} \Phi \left(- \frac{t}{\sqrt{\mathbb{P}_0(Q)(1 - \mathbb{P}_0(Q))}} \right) \end{aligned}$$

Since $\mathbb{P}_0(Q)(1 - \mathbb{P}_0(Q)) \leq 1/4$, we have $\lim_{n \to \infty} L_n(t) \leq 2K\Phi(-2t)$.

This implies that

$$\lim_{n \to \infty} - \frac{\log L_n(T)}{n+m} \geq \lim_{n \to \infty} - \frac{1}{n+m} \log\{2K\Phi(-2T)\}.$$

With the fact from Duembgen (2010) that $\forall x < 0, \Phi(x) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{|x|} \exp(-\frac{x^2}{2})$,

$$\begin{aligned} &\lim_{n \to \infty} - \frac{\log L_n(T)}{n+m} \\ &\geq \lim_{n \to \infty} - \frac{1}{n+m} \log \left\{ \frac{2K}{\sqrt{2\pi}} \frac{1}{2T} \exp\left(- \frac{4T^2}{2} \right) \right\} \\ &= \lim_{n \to \infty} \frac{2T^2}{n+m}. \end{aligned}$$

By Glivenko-Cantelli Theorem,

$$\lim_{n \to \infty} - \frac{\log L_n(T)}{n+m} \geq \lim_{n \to \infty} \frac{2nm}{(n+m)^2} \max_{Q \in \mathcal{Q}} (\mathbb{P}_T(Q) - \mathbb{P}_C(Q))^2.$$

Since $m/n = \tau$,

$$\lim_{n\to\infty} -\frac{\log L_n(T)}{n+m} \geq \frac{2\tau}{(\tau+1)^2} \max_{Q\in\mathcal{Q}} (\mathbb{P}_T(Q) - \mathbb{P}_C(Q))^2. \tag{A.3}$$

Next, we focus on a lower bound for $\lim_{n\to\infty} L_n(t)$, which can give an upper bound for the Bahadur slope. With a similar technique as before,

$$\begin{aligned}
\lim_{n\to\infty} L_n(t) &\geq \lim_{n\to\infty} \sup_{\mathbb{P}_0\in\mathcal{P}} \max_{Q\in\mathcal{Q}} \mathbb{P}_0 \left(\sqrt{\frac{nm}{n+m}} |\mathbb{P}_n(Q) - \mathbb{P}_m(Q)| > t \right) \\
&= 2 \sup_{\mathbb{P}_0\in\mathcal{P}} \max_{Q\in\mathcal{Q}} \Phi \left(-\frac{t}{\sqrt{\mathbb{P}_0(Q)(1-\mathbb{P}_0(Q))}} \right) \\
&= 2\Phi(-2t)
\end{aligned}$$

Using the fact that $\forall x < 0, \Phi(x) \geq \frac{1}{\sqrt{2\pi}} (\frac{|x|}{x^2+1}) \exp(-\frac{x^2}{2})$ (Duembgen, 2010), we have

$$\begin{aligned}
&\lim_{n\to\infty} -\frac{\log L_n(T)}{n+m} \\
\leq &\lim_{n\to\infty} -\frac{1}{n+m} \log\{2\Phi(-2T)\} \\
\leq &\lim_{n\to\infty} -\frac{1}{n+m} \log \left\{ \frac{2}{\sqrt{2\pi}} \left(\frac{2T}{4T^2+1} \right) \exp\left(-\frac{4T^2}{2} \right) \right\} \\
= &\lim_{n\to\infty} \frac{2T^2}{n+m}
\end{aligned}$$

Again by Glivenko-Cantelli Theorem,

$$\lim_{n\to\infty} -\frac{\log L_n(T)}{n+m} \leq \frac{2\tau}{(\tau+1)^2} \max_{Q\in\mathcal{Q}} (\mathbb{P}_T(Q) - \mathbb{P}_C(Q))^2 \tag{A.4}$$

Combining (A.3) and (A.4), we have

$$\lim_{n\to\infty} -\frac{\log L_n(T)}{n+m} = \frac{2\tau}{(\tau+1)^2} \max_{Q\in\mathcal{Q}} (\mathbb{P}_T(Q) - \mathbb{P}_C(Q))^2.$$

Hence, the Bahadur slope of T is

$$\frac{4\tau}{(\tau+1)^2}\max_{Q\in\mathcal{Q}}(\mathbb{P}_T(Q)-\mathbb{P}_C(Q))^2.$$

Proof of Proposition 9

To obtain the Bahadur slope of T^*, we first consider the Bahadur slope of each individual GFKS statistic, which can be derived from the Lemma 1. The Bahadur slope of a GFKS statistic on $\mathbf{X}_I$ is a special case of Lemma 1, with $\mathcal{Q}$ being the collection of orthants $\mathcal{Q}_I$. With $d_I = \max_{Q\in\mathcal{Q}_I}|\mathbb{P}_T(Q)-\mathbb{P}_C(Q)|$, the Bahadur slope of T_I is

$$\frac{4\tau}{(\tau+1)^2}d_I^2.$$

Combining these results with the minimum p-value argument (Berk and Jones, 1978), the Bahadur slope of the summary statistic T^* is

$$\frac{4\tau}{(\tau+1)^2}\max_{I\in\mathcal{I}}d_I^2.$$

Proof of Proposition 10

Since $\tau = m/n$, for any $I^*\in\mathcal{J}^*$, an application of Lemma 1 suggests that

$$-\frac{\log p_{I^*}}{n+m}\to\frac{2\tau}{(\tau+1)^2}d_{I^*}^2,\quad\text{almost surely.}$$

Similarly, for any $I\in\mathcal{J}$, we also have

$$-\frac{\log p_I}{n+m}\to\frac{2\tau}{(\tau+1)^2}d_I^2,\quad\text{almost surely.}$$

Thus, for any $I^* \in \mathcal{J}^*$ and $I \in \mathcal{J}$,

$$\frac{\log p_I}{n+m} - \frac{\log p_{I^*}}{n+m} \to \frac{2\tau}{(\tau+1)^2}\left(d_{I^*}^2 - d_I^2\right) > 0, \quad \text{almost surely.}$$

That is, for any $I^* \in \mathcal{J}^*$ and $I \in \mathcal{J}$,

$$\mathbb{P}\left(\lim_{n\to\infty}\left(\frac{\log p_I}{n+m} - \frac{\log p_{I^*}}{n+m}\right) > 0\right) = 1.$$

So we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\log p_I}{n+m} - \frac{\log p_{I^*}}{n+m} > 0\right) = 1, \quad \forall I^* \in \mathcal{J}^*, I \in \mathcal{J}.$$

Equivalently,

$$\lim_{n\to\infty} \mathbb{P}\left(p_{I^*} - p_I < 0\right) = 1, \quad \forall I^* \in \mathcal{J}^*, I \in \mathcal{J}.$$

Then, we have

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\left(\max_{I^*\in\mathcal{J}^*} p_{I^*} < \min_{I\in\mathcal{J}} p_I\right) =& 1 - \lim_{n\to\infty} \mathbb{P}\left(p_{I^*} - p_I \geq 0, \exists I^* \in \mathcal{J}^*, I \in \mathcal{J}\right) \\
\geq& 1 - \sum_{I^*\in\mathcal{J}^*}\sum_{I\in\mathcal{J}} \lim_{n\to\infty} \mathbb{P}\left(p_{I^*} - p_I \geq 0\right) \\
=& 1.
\end{aligned}$$

A.2.3. Using GFKS to Evaluate and Improve the Matched Samples

An iterative algorithm of using GFKS to evaluate and improve the matched samples

One specific implementation of the proposed method in Sections 2 and 3 is to use the minimum p-value of GFKS tests to evaluate the balance of matched samples. To improve a match, we use GFKS statistics to identify a worst-balanced covariate and a worst-case cut of that covariate at every iteration, until a satisfactory match is obtained. In the following, an iterative algorithm with GFKS tests to evaluate and improve the matched samples is formalized.

1. Choose a pre-specified level for type I error, e.g., $\alpha = 0.05$ or 0.1. For the the continuous covariates, find the quantiles of the distribution in the treated group based on a pre-specified vector of probabilities.

2. Set iteration index $l = 1$.

3. Perform GFKS tests for all covariates and their joint distributions (for discrete covariates, evaluate GFKS based on the orthants at their original values; for continuous covariates, evaluate GFKS based on the orthants at the quantiles determined in Step 1), use the minimum p-value T^* to combine p-values of all these tests, and denote the p-value for T^* by p_l.

4. If $p_l \leq \alpha$, reject the null hypothesis. Set $l = l + 1$. Construct a binary variable u_l on the worst-case cut of the most imbalanced variable and go to Step 5. Otherwise, stop and a satisfactory match has been obtained.

5. Construct a new match which nearly finely balance the interaction of $u_1, ..., u_l$, and go back to Step 3.

In step 1, one can use the suggested quantiles in Cochran (1968) of the distribution of the continuous variables in the treated group. For example, in the real data application in § 3.6, GFKS tests consider the orthants at $(11\%, 35\%, 65\%, 89\%)$ quantiles in the treated group for the continuous variables. GFKS tests should be performed for all individual variables in step 3, bivariate distributions are also recommended. As more joint moments are considered, computation time will increase.

Properties of combining p-values of GFKS tests in a more general way

The proposed algorithm treats the smallest p-value as the primary focus, and adjusts for that at every iteration. The following proposition considers a more general form of combining

p-values. Specifically, let

$$\mathcal{H} = \{h(x_1, ..., x_k) : h \text{ is symmetric and non-decreasing in each coordinate } x_i,\ i = 1, ..., k\}$$

(see Benjamini and Heller, 2008), and $\mathcal{L}$ denote the collection of any arbitrary k p-values from $\{p_I : I \in \mathcal{I}\}$. For any $h \in \mathcal{H}$, define

$$T^* := \min_{(x_1,...,x_k)\in\mathcal{L}} h(x_1, ..., x_k) = h(p_{(1)}, ...p_{(k)}),$$

where $p_{(i)}$'s are the order statistics of $\{p_I : I \in \mathcal{I}\}$. A special case of the following proposition suggests that T^* aggregates the k largest Bahadur slopes with probability approaching to one as the sample size goes to infinity, which provides guidance on adjusting k most imbalanced problems simultaneously. It is easiest to understand Proposition 12 in the case that the M Bahadur slopes are distinct, but Proposition 12 is also correct if some slopes are tied.

Proposition 12 *For any $d^* \in \mathbb{R}$, let $\widetilde{\mathcal{J}}^* = \{I \in \mathcal{I} : d_I \geq d^*\}$ denote the collection of I's for which T_I has Bahadur slopes at least $\frac{4\tau}{(\tau+1)^2}d^{*2}$ and $\widetilde{\mathcal{J}} = \{I \in \mathcal{I} : d_I < d^*\}$ denotes the rest of I's. If $\widetilde{\mathcal{J}}^*$ and $\widetilde{\mathcal{J}}$ are not empty, we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{I^*\in\widetilde{\mathcal{J}}^*} p_{I^*} < \min_{I\in\widetilde{\mathcal{J}}} p_I\right) = 1.$$

Let $d_{(i)}$ be the ith ordered value of $\{d_I : I \in \mathcal{I}\}$. With $d^* = d_{(M-k+1)}$ in Proposition 12, the set $\widetilde{\mathcal{J}}^*$ becomes the collection of I's for which T_I has the k largest Bahadur slopes, and $\widetilde{\mathcal{J}}$ becomes the rest of I's, such that $|\widetilde{\mathcal{J}}^*| \geq k$ and $|\widetilde{\mathcal{J}}| \leq M - k$. Therefore, it says that in sufficiently large samples, the k largest Bahadur slopes are likely to lead to the k smallest p-values. Proof of Proposition 12, is similar to that of Proposition 10, and is omitted.

A.2.4. Simulation: Comparison with the Classification Permutation Test

Consider the following simulation experiment: $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_{10})$ in the treated group (with sample size 500) and $\mathbf{X} \sim N((\delta, \delta, 0, ..., 0), \mathbf{I}_{10})$ in the control group (with sample size 500). The listed methods in §5.1 are compared with the classification permutation test (Gagnon-Bartsch and Shem-Tov, 2019) using one popular machine learning method, random forests, in terms of power and type I error. We use the `R` package `cpt` with 500 permutations to estimate its p-value. Due to computation limits, the empirical probability of rejecting the null hypothesis is estimated with 500 replications. Results are summarized in the following figure. Similar as results in §5.1, the unadjusted two-sample t-test and KS test inflates the type I error (when $\delta = 0$). In addition, it can be concluded from Figure 8 that the four methods with the proposed framework have a larger power than random forests based permutation test in this simulation.

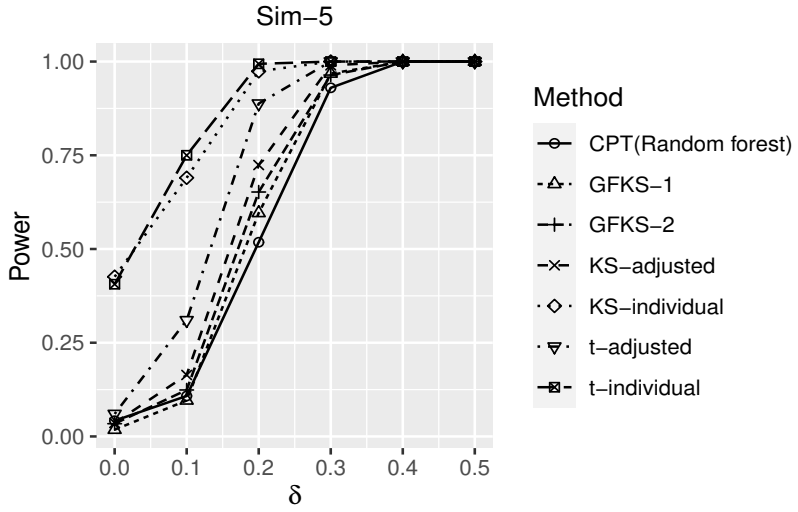


Figure 8: Estimated power and type I error of two-sample t test on individual covariates (t-individual), Kolmogorov-Smirnov test on individual covariates (KS-individual), adjusted t test (t-adjusted), adjusted Kolmogorov-Smirnov test (KS-adjusted), univariate GFKS (GFKS-1), bivariate GFKS (GFKS-2), and random forest based permuatation test (CPT (Random forest)).

A.2.5. Real Data Application Details

Definitions of covariates

- age: age in years

- female: indicator for female
- black: indicator for black (race)
- Hispanic: indicator for Hiapanic (race)
- education: education in 5-point scale, with 1 for less than 9th grade, 3 for high school or equivalent, and 5 for college graduate
- poverty: the ratio of family income to the poverty level, with missing values imputed as mean
- povertyNA: an indicator for missingness of poverty
- dietsup: whether has taken any dietary supplements in the past month
- height: height in cm
- weight: weight in kg
- BMI: body mass index
- cotinine: serum cotinine level (ng/mL)
- vitaminD: serum 25-hydroxyvitamin D2 + D3 level (nmol/L)
- diabetes: whether diagnosed with diabetes or borderline diabetes
- insurance: whether covered by insurance
- weighmore: whether would like to weigh more
- weighless: whether would like to weigh less
- weightchange: weight change in the recent 1 year of finishing the questionnaire (kg)
- physicalact: whether has moderate and vigorous activity in the past 30 days.

Effects of the iterative algorithm on estimating ATT

We compare the estimated average treatment effects on the treated (ATT) with the naive estimate – difference of the average of the treated group and the average of the control group and using the four matched samples. As in §6.3 in the thesis, all four outcomes, femur bone mineral density, femur bone mineral content, femoral neck bone mineral density and femoral neck bone mineral content, are analyzed in a log scale. Results are summarized in the following table. From Table 10, we can see that the matching estimates are smaller in absolute value than the naive estimate. Estimates from the four matched samples differ slightly, which may not be the case in other examples.

Table 10: Estimated average treatment effects on the treated comparison for four outcomes: femur bone mineral density, femur bone mineral content, femoral neck bone mineral density and femoral neck bone mineral content.

	Naive	M0	M1	M2	M3
Femur BMD	-0.031	-0.002	-0.004	-0.003	-0.003
Femur BMC	-0.050	-0.006	-0.011	-0.013	-0.010
Fneck BMD	-0.038	-0.000	0.003	0.002	0.002
Fneck BMC	-0.045	0.001	0.002	-0.000	-0.000

BIBLIOGRAPHY

I. G. Abrahamson. Exact bahadur efficiencies for the kolmogorov-smirnov and kuiper one- and two-sample statistics. *Annals of Mathematical Statistics*, 38(5):1475–1490, 1967.

T. W. Anderson and D. A. Darling. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, pages 193–212, 1952.

P. C. Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25):3083–3107, 2009.

R. R. Bahadur. Rates of convergence of estimates and test statistics. *Annals of Mathematical Statistics*, 38(2):303–324, 1967.

L. A. Bazzano, J. He, P. Muntner, S. Vupputuri, and P. K. Whelton. Relationship between cigarette smoking and novel risk factors for cardiovascular disease in the united states. *Annals of Internal Medicine*, 138(11):891–897, 2003.

R. H. Berk and D. H. Jones. Relatively optimal combinations of test statistics. *Scandinavian Journal of Statistics*, 5:158–162, 1978.

D. P. Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21(1):152–171, 1981.

D. P. Bertsekas. *Network Optimization*. Belmont, MA: Athena Scientific, 1998.

D. P. Bertsekas and P. Tseng. The relax codes for linear minimum cost network flow problems. *Annals of Operations Research*, 13(1):125–190, 1988.

P. J. Bickel. A distribution free version of the smirnov two sample test in the p-variate case. *Annals of Mathematical Statistics*, 40(1):1–23, 1969.

K. C. G. Chan, S. C. P. Yam, and Z. Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 78(3):673–700, 2016.

H. Chen and D. S. Small. New multivariate tests for assessing covariate balance in matched observational studies. *Biometrics*, https://doi.org/10.1111/biom.13395, 2020.

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

W. G. Cochran and D. B. Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446, 1973.

X. de Luna, I. Waernbaum, and T. S. Richardson. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875, 2011.

S. J. Diem, T. L. Blackwell, K. L. Stone, K. Yaffe, E. M. Haney, M. M. Bliziotes, and K. E. Ensrud. Use of antidepressants and rates of hip bone loss in older women: the study of osteoporotic fractures. *Archives of Internal Medicine*, 167(12):1240–1245, 2007.

F. Dudbridge and B. P. C. Koeleman. Rank truncated product of p-values, with application to genomewide association scans. *Genetic Epidemiology*, 25(4):360–366, 2003.

L. Duembgen. Bounding standard gaussian tail probabilities. *arXiv preprint arXiv:1012.2063*, 2010.

J. Fan, K. Imai, H. Liu, Y. Ning, and X. Yang. Improving covariate balancing propensity score: A doubly robust and efficient approach. Technical report, Technical report, Princeton University, 2016.

A. J. Feuer, R. T. Demmer, A. Thai, and M. G. Vogiatzi. Use of selective serotonin reuptake inhibitors and bone mass in adolescents: An nhanes study. *Bone*, 78:28–33, 2015.

M. L. Fisher. The lagrangian relaxation method for solving integer programming problems. *Management Science*, 27(1):1–18, 1981.

R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

M. A. Fligner and D. A. Wolfe. Some applications of sample analogues to the probability integral transformation and a coverage property. *American Statistician*, 30(2):78–85, 1976.

C. B. Fogarty, M. E. Mikkelsen, D. F. Gaieski, and D. S. Small. Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *Journal of the American Statistical Association*, 111(514): 447–458, 2016.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*. New York: Springer, 2001.

J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *Annals of Statistics*, 7:697–717, 1979.

J. Gagnon-Bartsch and Y. Shem-Tov. The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies. *Annals of Applied Statistics*, 13(3):1464–1483, 2019.

F. Glover. Maximum matching in a convex bipartite graph. *Naval Research Logistics Quarterly*, 14(3):313–316, 1967.

J. J. Goeman, A. Solari, and T. Stijnen. Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Statistics in Medicine*, 29(20):2117–2125, 2010.

J. Hahn. Functional restriction and efficiency in causal inference. *Review of Economics and Statistics*, 86(1):73–76, 2004.

E. M. Haney, B. K. S. Chan, S. J. Diem, K. E. Ensrud, J. A. Cauley, E. Barrett-Connor, E. Orwoll, and M. M. Bliziotes. Association of low bone mineral density with selective serotonin reuptake inhibitor use by older men. *Archives of Internal Medicine*, 167(12): 1246–1251, 2007.

G. J. Hankey and J. W. Eikelboom. Homocysteine and vascular disease. *Lancet*, 354(9176): 407–413, 1999.

B. B. Hansen. Optmatch: Flexible, optimal matching for observational studies. *R News*, 7 (2):18–24, 2007.

B. B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008.

B. B. Hansen and J. Bowers. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, pages 219–236, 2008.

B. B. Hansen and S. O. Klopfer. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627, 2006.

J. J. Heckman, H. Ichimura, and P. E. Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654, 1997.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 243–263, 2014.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

C. Kilcioglu and J. R. Zubizarreta. Maximizing the information content of a balanced matched sample in a study of the economic performance of green buildings. *Annals of Applied Statistics*, 10(4):1997–2020, 2016.

B. Korte and J. Vygen. *Combinatorial Optimization*. New York: Springer, 2012.

E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

W. Lipski and F. P. Preparata. Efficient algorithms for finding maximum matchings in convex bipartite graphs and related problems. *Acta Informatica*, 15(4):329–346, 1981.

B. Lu, R. Greevy, X. Xu, and C. Beck. Optimal nonbipartite matching and its statistical applications. *American Statistician*, 65(1):21–30, 2011.

J. S. Maritz. A note on exact robust confidence intervals for location. *Biometrika*, 66(1): 163–170, 1979.

J. Pearl. On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint arXiv:1203.3503*, 2012.

S. D. Pimentel. Large, sparse optimal matching with r package rcbalance. *Observational Studies*, 2:4–23, 2016.

S. D. Pimentel, R. R. Kelz, J. H. Silber, and P. R. Rosenbaum. Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *Journal of the American Statistical Association*, 110(510): 515–527, 2015a.

S. D. Pimentel, F. Yoon, and L. Keele. Variable-ratio matching with fine balance in a study of the peer health exchange. *Statistics in Medicine*, 34(30):4070–4082, 2015b.

S. D. Pimentel, D. S. Small, and P. R. Rosenbaum. Constructed second control groups and attenuation of unmeasured biases. *Journal of the American Statistical Association*, 111 (515):1157–1167, 2016.

P. A. Pirraglia, R. S. Stafford, and D. E. Singer. Trends in prescribing of selective serotonin reuptake inhibitors and other newer antidepressant agents in adult primary care. *Primary Care Companion to the Journal of Clinical Psychiatry*, 5(4):153–157, 2003.

S. Roman, S. Axler, and F. W. Gehring. *Advanced Linear Algebra*, volume 3. Springer, 2005.

P. R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987.

P. R. Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032, 1989.

P. R. Rosenbaum. Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association*, 97(457):183–192, 2002.

P. R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530, 2005.

P. R. Rosenbaum. Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics*, 63(2):456–464, 2007.

P. R. Rosenbaum. Testing hypotheses in order. *Biometrika*, 95(1):248–252, 2008.

P. R. Rosenbaum. *Design of Observational Studies*. New York: Springer, 2010.

P. R. Rosenbaum. Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics*, 69(1):118–127, 2013.

P. R. Rosenbaum. Bahadur efficiency of sensitivity analyses in observational studies. *Journal of the American Statistical Association*, 110(509):205–217, 2015.

P. R. Rosenbaum. Using Scheffé projections for multiple outcomes in an observational study of smoking and periodontal disease. *Annals of Applied Statistics*, 10(3):1447–1471, 2016.

P. R. Rosenbaum. Imposing minimax and quantile constraints on optimal matching in observational studies. *Journal of Computational and Graphical Statistics*, 26(1):66–78, 2017a.

P. R. Rosenbaum. *Observation and Experiment*. Cambridge, MA: Harvard University Press, 2017b.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

P. R. Rosenbaum and D. B. Rubin. The bias due to incomplete matching. *Biometrics*, pages 103–116, 1985a.

P. R. Rosenbaum and D. B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, 39(1): 33–38, 1985b.

P. R. Rosenbaum and J. H. Silber. Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association*, 104(488):1398–1405, 2009.

P. R. Rosenbaum, R. N. Ross, and J. H. Silber. Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association*, 102(477):75–83, 2007.

R. Rosenthal. An evaluation of procedures and results. In K. Wachter and M. Straf, editors, *The Future of the Meta-Analysis*. New York: Russell Sage Foundation, 1990.

D. B. Rubin. Matching to remove bias in observational studies. *Biometrics*, 29:159–183, 1973.

D. B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74:318–328, 1979.

D. B. Rubin. Bias reduction using mahalanobis metric matching. *Biometrics*, 36, 1980.

D. B. Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36, 2007.

A. Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, 1998.

S. Seshadri, A. Beiser, J. Selhub, P. F. Jacques, I. H. Rosenberg, R. B. D'Agostino, P. W. F. Wilson, and P. A. Wolf. Plasma homocysteine as a risk factor for dementia and alzheimer's disease. *New England Journal of Medicine*, 346(7):476–483, 2002.

J. H. Silber, P. R. Rosenbaum, M. D. McHugh, J. M. Ludwig, H. L. Smith, B. A. Niknam, O. Even-Shoshan, L. A. Fleisher, R. R. Kelz, and L. H. Aiken. Comparison of the value of nursing work environments in hospitals across different levels of patient risk. *JAMA Surgery*, 151(6):527–536, 2016.

J. H. Silber, P. R. Rosenbaum, W. Wang, S. R. Calhoun, J. G. Reiter, O. Even-Shoshan, and W. J. Greeley. Practice style variation in medicaid and non-medicaid children with complex chronic conditions undergoing surgery. *Annals of Surgery*, 267(2):392–400, 2018.

E. A. Stuart. Matching methods for causal inference. *Statistical Science*, 25(1):1–21, 2010.

Y. Wang and J. R. Zubizarreta. Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika*, 107(1):93–105, 2020.

G. N. Welch and J. Loscalzo. Homocysteine and atherothrombosis. *New England Journal of Medicine*, 338(15):1042–1050, 1998.

H. S. Wieand. A condition under which the pitman and bahadur approaches to efficiency coincide. *Annals of Statistics*, 4(5):1003–1011, 1976.

L. A. Wolsey. *Integer Programming*, volume 42. Wiley Online Library, 1998.

D. Yang, D. S. Small, J. H. Silber, and P. R. Rosenbaum. Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics*, 68 (2):628–636, 2012.

R. Yu and P. R. Rosenbaum. Directional penalties for optimal matching in observational studies. *Biometrics*, 75(4):1380–1390, 2019.

R. Yu, J. H. Silber, and P. R. Rosenbaum. Matching methods for observational studies derived from large administrative databases. *Statistical Science*, 35(3):338–355, 2020.

D. V. Zaykin, L. A. Zhivotovsky, P. H. Westfall, and B. S. Weir. Truncated product method for combining p-values. *Genetic Epidemiology*, 22(2):170–185, 2002.

J. R. Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500): 1360–1371, 2012.

J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

J. R. Zubizarreta, C. E. Reinke, R. R. Kelz, J. H. Silber, and P. R. Rosenbaum. Matching for several sparse nominal variables in a case-control study of readmission following surgery. *American Statistician*, 65(4):229–238, 2011.