Statistical Science
2020, Vol. 35, No. 3, 338–355
https://doi.org/10.1214/19-STS699

Matching Methods for Observational Studies Derived from Large Administrative Databases[1]

Ruoqi Yu, Jeffrey H. Silber and Paul R. Rosenbaum

Abstract. We propose new optimal matching techniques for large administrative data sets. In current practice, very large matched samples are constructed by subdividing the population and solving a series of smaller problems, for instance, matching men to men and separately matching women to women. Without simplification of some kind, the time required to optimally match T treated individuals to T controls selected from $C \geq T$ potential controls grows much faster than linearly with the number of people to be matched—the required time is of order $O\{(T+C)^3\}$—so splitting one large problem into many small problems greatly accelerates the computations. This common practice has several disadvantages that we describe. In its place, we propose a single match, using everyone, that 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. In this graph, a modified form of near-fine balance can be used on a much larger scale, improving its effectiveness. We illustrate the method using data 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, controlling for 29 covariates plus 463 Principal Surgical Procedures, plus 973 Principal Diagnoses. The method is implemented in an `R` package `bigmatch` available from `CRAN`.

Key words and phrases: Causal inference, fine balance, Glover's algorithm, observational study, optimal caliper, optimal matching, propensity score.

1. INTRODUCTION: THE PROBLEM; AN EXAMPLE; OUTLINE

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 chil-

Ruoqi Yu is a PhD student, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: ruoqiyu@wharton.upenn.edu). Jeffrey H. Silber is Professor, Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: silber@email.chop.edu; URL: https://cor.research.chop.edu/). Paul R. Rosenbaum is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: rosenbaum@wharton.upenn.edu).

[1]Discussed in 10.1214/19-STS739, 10.1214/19-STS740, 10.1214/19-STS741; rejoinder at 10.1214/20-STS790.

dren 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, 1985), externally estimated prognostic or risk scores (Hansen, 2008), covariate distances (Rubin, 1980), fine balance constraints (Rosenbaum, Ross and Silber, 2007, Yang et al., 2012, Zubizarreta, 2012, Pimentel et al., 2015), 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, Ross and Silber, 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.

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 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 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 (1985).

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 standardard deviation before matching. Standardized differences above* 0.2 *standard deviations are in* ***bold***

	Covariate mean			Standardized difference	
		Controls			
Covariate	Treated	Matched	All	Matched	All
Sample size	38,841	38,841	159,527	38,841	159,527
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

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

1.3 Outline of the Paper: Concepts in Pictures, Formal Results, Application

Section 2 uses a toy example with 30 individuals and a few drawings to indicate the changes we suggest for matching in large administrative data bases. 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 Section 2.1 removing edges from a graph, and with Section 2.2 bringing in the key concept of fine balance. The practical aspects of creating a match are briefly sketched in Section 3. Then, Section 4 develops the topic formally and in greater generality. A goal in Section 4 is to quantify the reduction in computational effort produced by the ideas informally introduced in Section 2. We illustrate the technique in Section 5 using the Medicaid data mentioned in Section 1.2. Proofs are given in the Appendix.

2. AN INFORMAL DISCUSSION OF OPTIMAL MATCHING FOR LARGE DATA BASES

2.1 A Motivating Picture Illustrating Some Issues and Methods

2.1.1 *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.

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 Section 2.1, we seek a pair matching that minimizes the total cost over the 7 chosen edges, a standard combinatorial optimization problem; however, in Section 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 Section 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 2019. The difficulty of optimal matching grows much faster than linearly with

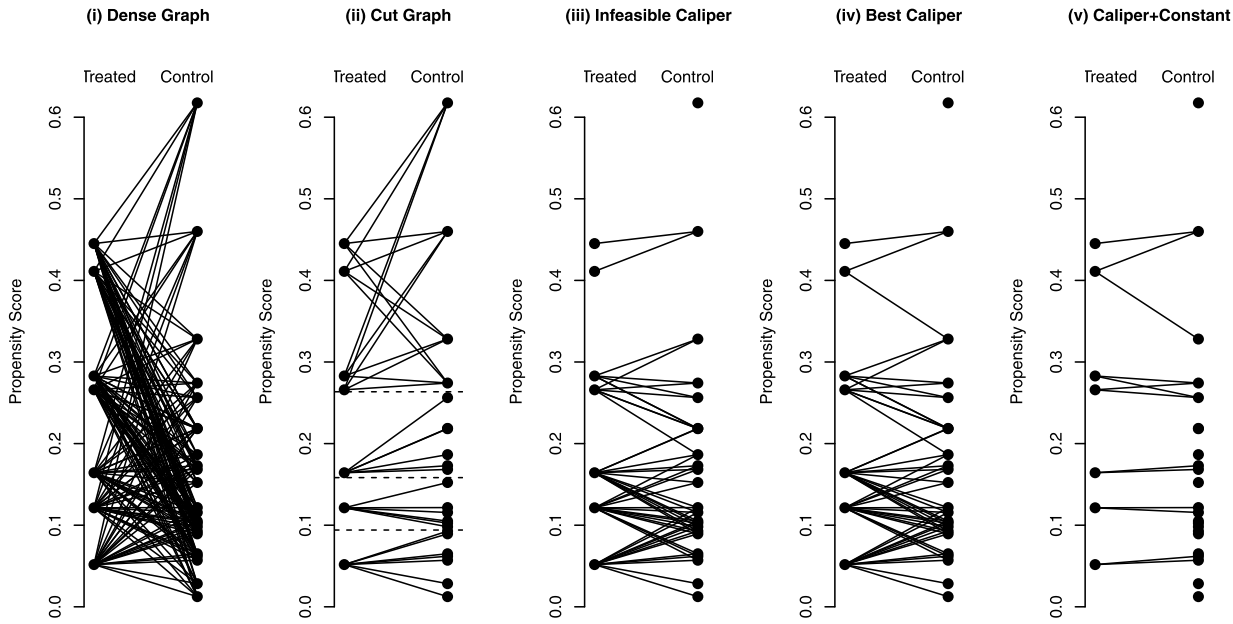

FIG. 1. *Five bipartite graphs, where the vertical axis is the propensity score. There is a decision variable for each potential pairing of a treated subject and a potential control, that is, a decision variable for each edge. Graph* (*i*) *has all possible pairings. Graph* (*ii*) *has reduced the number 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.*

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, for example, Section 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.

2.1.2 *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(b) 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.

2.1.3 *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 control, then we would have imposed a caliper of $x \pm 2$. Cochran and Rubin (1973) discussed caliper matching. Rosenbaum and Rubin (1985) 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's (1967) 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).

2.1.4 *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 A Second Motivating Figure Incorporating Other Matching Techniques

2.2.1 *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

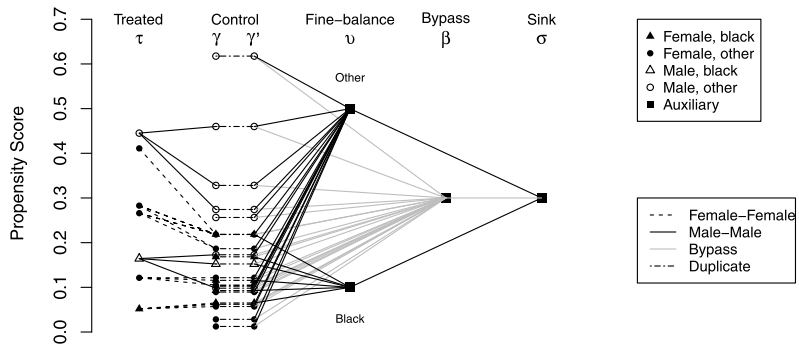

FIG. 2. *A bipartite graph matching exactly for gender, expanded for near fine balance of race, υ, black or other. The optimal caliper is now* 0.1925, *and with this caliper the minimum number of neighbors is* $\nu = 3$. *The duplicate edges connect γ to γ', with capacity* 1, *so they insist that a control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

feature, exact matching for a nominal covariate, is discussed in Section 2.2.1, while the second feature, near-fine balance, is discussed in Section 2.2.2.

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

2.2.2 *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, Ross and Silber, 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, |treated − 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. (2015) 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. (2015) can be applied in conjunction with the new methods described informally in Section 2.1 and formally in Section 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 Section 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. (2015). A soft constraint is necessary here because the bipartite graph is thinned by calipers and near neighbors.

2.2.3 *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 $O(P^2)$ arithmetic operations for P covariates. This can be reduced to $O(P)$ 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 Section 5 where $\Xi = 463$ principal procedures are exactly matched.

3. PRACTICAL ASPECTS OF MATCHING IN LARGE DATABASES

Section 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 Section 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 Section 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.

4. NETWORK STRUCTURE: A SPARSE BIPARTITE GRAPH EXPANDED FOR NEAR-FINE BALANCE

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 $\mathcal{T} \cap \mathcal{C} = \varnothing$, 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 Section 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, Chapters 8 and 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 (1985).

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 Section 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 $\mathcal{M}$,

$$|\{\tau \in \mathcal{T} : \upsilon(\tau) = k\}| = |\{\gamma \in \mathcal{M} : \upsilon(\gamma) = k\}| \quad \text{for } k = 1, \ldots, \Upsilon. \tag{4.1}$$

Notice that (4.1) 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 (4.1) 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 (4.1) is not feasible; it cannot be done. Write

$$d_k = \big|\{\tau \in \mathcal{T} : \upsilon(\tau) = k\}\big| - \big|\{\gamma \in \mathcal{M} : \upsilon(\gamma) = k\}\big|, \tag{4.2}$$

so $d_k = 0$ for $k = 1, \ldots, \Upsilon$ when (4.1) holds. Near-fine balance means that we minimize $\sum_{k=1}^{K} |d_k|$ when fine balance (4.1) 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*

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

All proofs are in the Appendix.

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 $\gamma_{i+1} \in \phi(\tau)$, $\ldots$, $\gamma_{j-1} \in \phi(\tau)$. 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 Section 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 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 glover$(\varkappa) = 0$. If glover$(0) = 1$, stop; pair matching is feasible with caliper $\varkappa = 0$. If 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 glover$(\overline{\varkappa}) = 1$, set $\varkappa_{\max} \leftarrow \overline{\varkappa}$ and go to step 1, but if 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 $o_\nu(\tau)$ 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 $\xi(\tau) = \xi(\gamma)$, $|\rho(\tau) - \rho(\gamma)| \leq \min\{\varkappa, o_\nu(\tau)\}$. 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 Section 4.4.

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 $\mathcal{C}' = \{\gamma'_1, \ldots, \gamma'_C\}$, adding the γ' to the set of nodes and adding C edges (γ_j, γ'_j) 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

$$\begin{aligned}\mathcal{N} &= \mathcal{T} \cup \mathcal{C} \cup \mathcal{C}' \cup \{1, \ldots, \Upsilon\} \cup \{\beta, \sigma\},\\ \mathcal{E} &= \mathcal{B} \cup \mathcal{O} \cup \big\{(\gamma', \upsilon(\gamma')) : \gamma' \in \mathcal{C}'\big\}\\ &\quad \cup \big\{(\upsilon, \sigma) : \upsilon \in \{1, \ldots, \Upsilon\}\big\}\\ &\quad \cup \big\{(\gamma', \beta) : \gamma' \in \mathcal{C}'\big\} \cup \big\{(\beta, \sigma)\big\}.\end{aligned} \tag{4.3}$$

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 Section 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.

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 $e \in \mathcal{E}$,

$$\begin{aligned}&0 \leq f(e) \leq \text{cap}(e)\\ &\quad \text{with } f(e) \in \{0, 1, 2, \ldots\} \text{ for each } e \in \mathcal{E}\end{aligned} \tag{4.4}$$

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

$$\begin{aligned}\text{div}(n) = &\sum_{n'':(n,n'')\in\mathcal{E}} f\big\{(n, n'')\big\}\\ &- \sum_{n':(n',n)\in\mathcal{E}} f\big\{(n', n)\big\}\\ &\text{for each } n \in \mathcal{N}.\end{aligned} \tag{4.5}$$

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 (4.3) or in Figure 2, set

$$\begin{aligned}\text{div}(\tau) &= 1 \quad \text{for } \tau \in \mathcal{T},\\ \text{div}(\sigma) &= -T,\\ \text{div}(n) &= 0 \quad \text{for } n \notin \mathcal{T} \cup \{\sigma\}\end{aligned} \tag{4.6}$$

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

$$\begin{aligned}\text{cap}\big\{(\tau, \gamma)\big\} &= 1 \quad \text{for } (\tau, \gamma) \in \mathcal{B},\\ \text{cap}\big\{(\gamma, \gamma')\big\} &= 1 \quad \text{for } (\gamma, \gamma') \in \mathcal{O},\\ \text{cap}\big\{(\gamma', \upsilon(\gamma'))\big\} &= 1 \quad \text{for } \gamma' \in \mathcal{C}',\\ \text{cap}\big\{(\gamma', \beta)\big\} &= 1 \quad \text{for } \gamma' \in \mathcal{C}',\\ \text{cap}\big\{(k, \sigma)\big\} &= \big|\{\tau \in \mathcal{T} : \upsilon(\tau) = k\}\big|\\ &\qquad \text{for } k \in \{1, \ldots, \Upsilon\},\\ \text{cap}\big\{(\beta, \sigma)\big\} &= T.\end{aligned} \tag{4.7}$$

Combining (4.4), (4.5), (4.6), and (4.7) says the following about a feasible flow, $f(\cdot)$. No control $\gamma \in \mathcal{C}$ can receive more than one unit of flow, because it must transfer all its flow to γ', and $\text{cap}\{(\gamma, \gamma')\} = 1$. Because $f(\cdot)$ takes on nonnegative integer values, $f\{(\tau, \gamma)\} = 1$ for at most

T nonoverlapping pairs, (τ, γ). The pair match will be defined by $\mu(\tau) = \gamma$ if and only if $f\{(\tau, \gamma)\} = 1$.

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

$$\begin{aligned} &\text{cost}(e) = \delta(\tau, \gamma) \geq 0 \quad \text{for } e = (\tau, \gamma) \in \mathcal{B}, \\ &\text{cost}(e) = \Psi > 0 \quad \text{for } e = (\beta, \sigma), \\ &\text{cost}(e) = 0 \quad \text{for } e \notin \mathcal{B} \cup \{(\beta, \sigma)\}. \end{aligned} \tag{4.8}$$

DEFINITION 3. The matching network refers to nodes $\mathcal{N}$ and directed edges $\mathcal{E}$ given by (4.3), divergences given by (4.6), capacities given by (4.7) and costs given by (4.8). A flow in the matching network is feasible if it satisfies (4.4), (4.5), (4.6), and (4.7). A minimum cost flow is a feasible flow that minimizes $\text{cost}(f) = \sum_{e \in \mathcal{E}} f(e)\,\text{cost}(e)$ 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 $f(\cdot)$ in the matching network $(\mathcal{N}, \mathcal{E})$ defines a feasible pair matching in $\mathcal{B}$ as follows: $\mu(\tau) = \gamma$ if and only if $f\{(\tau, \gamma)\} = 1$.*

Recall the definition of the deviation d_k from fine balance in (4.2). 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 (4.8) 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. (2015) 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 $\sum_{k=1}^{K} |d_k|$. Moreover, among pair matches in $\mathcal{B}$ that minimize $\sum_{k=1}^{K} |d_k|$, a match obtained from a minimum cost feasible flow minimizes $\sum_{\tau \in \mathcal{T}} \delta\{\tau, \mu(\tau)\}$.*

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 $O\{\nu C^2 + C^2 \log(C)\}$. In particular, if ν is uniformly bounded, $\nu \leq \overline{\nu}$, then the time required is bounded by $O\{C^2 \log(C)\}$.*

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 $O\{|\mathcal{N}| \cdot |\mathcal{E}| + |\mathcal{N}|^2 \cdot \log(|\mathcal{N}|)\} = O(C^3)$, a much larger bound than $O\{C^2 \log(C)\}$. Proposition 6 indicates that even with growing values of ν we have a time bound of $O\{C^2 \log(C)\}$ 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's (2016) `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.

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 $\mathcal{T}^* = \{\tau_1, \tau_1', \ldots, \tau_T, \tau_T'\}$, so $|\mathcal{T}^*| = 2T$, and then apply the method for pair matching described earlier. Because Glover's algorithm in Section 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 $\text{div}(\tau_t) = \omega$ for $t = 1, \ldots, T$ in (4.6) to match with $\omega \geq 2$ controls. This is implemented in the `R` package `bigmatch`.

5. CONSTRUCTING THE MATCHED SAMPLE IN THE MEDICAID EXAMPLE

5.1 Finding the Minimal Caliper $\varkappa$ and Number of Neighbors ν

In the Medicaid example in Section 1.2, the first step is to construct the bipartite graph $(\mathcal{T} \cup \mathcal{C}, \mathcal{B})$ analogous

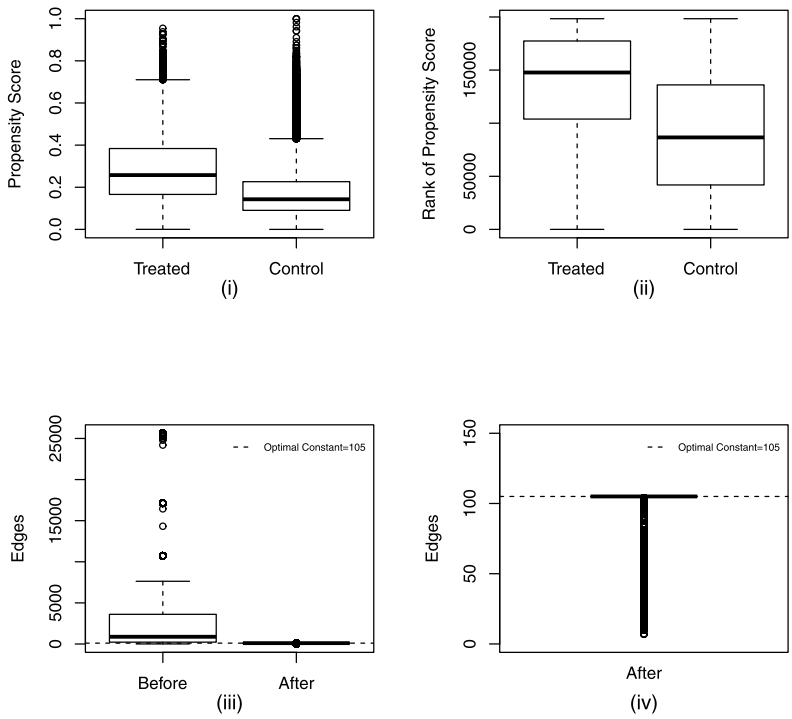

FIG. 3. *Creating the bipartite graph by exact matching for* 463 *Principal Procedures with an optimal caliper on the rank of the propensity score. There are* 38,841 *treated nodes and* 159,527 *control nodes.* (*i*) *The propensity score before matching.* (*ii*) *Ranks of the propensity score before 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.*

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 Section 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; that is, this is the smallest feasible caliper in the sense that distinguished Figure 1(iii) and Figure 1(iv). Note that $\varkappa/(T + C) = 170{,}925.1/198{,}368 = 0.86$, 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.

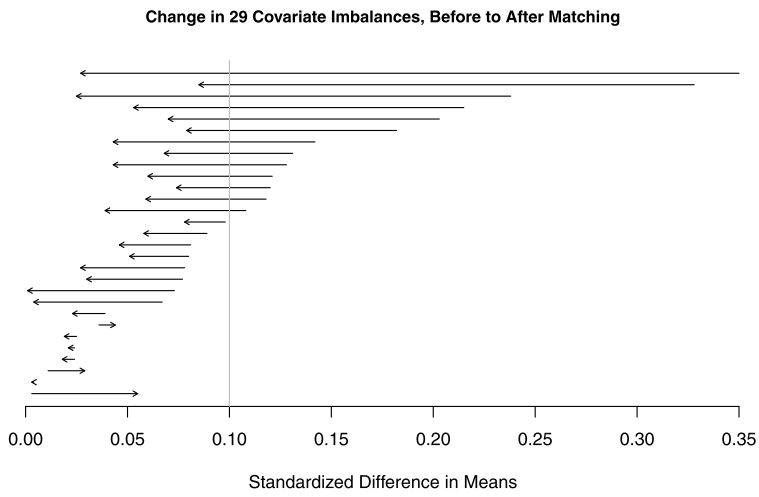

FIG. 4. *Change in covariate imbalance from before matching to after matching for the* 29 *covariates in Table* 1. *The point of the arrow is after matching. A vertical line is at* 0.1. *After matching, all standardized differences are less than* 0.1, *and all large imbalances before matching have been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

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 4(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.

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), Section 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 Section 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.

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's (1979) 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, $\theta : \mathcal{T} \cup \mathcal{C} \to \{1, 2, \ldots, \Theta\}$, 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 (4.2) 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. 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. (2015), we compare our matched sample to 10,000 randomized experiments 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

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 × Diagnosis
Categories	463	973	463×973
Imbalance	0	846	11,704
Minimum imbalance	2880	3354	13,122
Mean imbalance	3479	3930	13,802
Chi-squared statistic	0	307	8092
Miniumum chi-squared statistic	366	646	8446
Mean chi-squared statistic	462	776	8734

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.

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.

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.

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 children's hospital	Dead	16	94	110
	Alive	95	38,636	38,731
	Total	111	38,730	38,841

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, Solari and Stijnen (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. (2015), Section 5, using the particular test in Rosenbaum (2002), Section 6, combined with the general method of Goeman, Solari and Stijnen (2010).

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.

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.

APPENDIX: PROOFS

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 (4.4), (4.5), (4.6), and (4.7), 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 $f\{(\gamma, \gamma')\} \le 1$, so that $\sum_{\tau:(\tau,\gamma)\in\mathcal{B}} f\{(\tau, \gamma)\} \le 1$; 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 $\text{cost}(f) \le \text{cost}(g)$. First, we show that the bypass flow is smaller for $f(\cdot)$, or more precisely, we show $f\{(\beta, \sigma)\} \le g\{(\beta, \sigma)\}$. Let $h(\cdot)$ be any feasible flow. Using (4.8) 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

$$\begin{aligned} & h\{(\beta, \sigma)\} \cdot \Psi \\ & \quad \le \text{cost}(h) \\ & \quad \le h\{(\beta, \sigma)\} \cdot \Psi + \sum_{(\tau,\gamma)\in\mathcal{B}} \delta(\tau, \gamma) \\ & \quad < \big[h\{(\beta, \sigma)\} + 1\big] \cdot \Psi, \end{aligned} \tag{A.1}$$

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 (4.4). If $g\{(\beta, \sigma)\} < f\{(\beta, \sigma)\}$ then $f\{(\beta, \sigma)\} \ge g\{(\beta, \sigma)\} + 1$, and using (A.1),

$$\begin{aligned} \text{cost}(g) & < \big[g\{(\beta, \sigma)\} + 1\big] \cdot \Psi \\ & \le f\{(\beta, \sigma)\} \cdot \Psi \\ & \le \text{cost}(f), \end{aligned}$$

which is impossible because $f(\cdot)$ is a minimum cost feasible flow; so, we conclude $f\{(\beta, \sigma)\} \le g\{(\beta, \sigma)\}$. 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 (4.2) and (4.7), that $d_k = |\{\tau \in \mathcal{T} : \upsilon(\tau) = k\}| - |\{\gamma \in \mathcal{M} : \upsilon(\gamma) = k\}|$ may be written

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

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 $d_k \le 0$. 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 $\text{cost}(f) \le \text{cost}(g)$, then it follows that $\sum_{(\tau,\gamma)\in\mathcal{B}} \delta(\tau, \gamma) f\{(\tau, \gamma)\} \le \sum_{(\tau,\gamma)\in\mathcal{B}} \delta(\tau, \gamma) g\{(\tau, \gamma)\}$, 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 $O\{|\mathcal{N}| \cdot |\mathcal{E}| + |\mathcal{N}|^2 \cdot \log(|\mathcal{N}|)\}$; 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 $T \le C$, 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 $\{\gamma', \upsilon(\gamma')\}$, $\Upsilon \le C$ 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 $T \le C$, so the result follows. □

REFERENCES

BERTSEKAS, D. P. (1981). A new algorithm for the assignment problem. *Math. Program.* **21** 152–171. MR0623835 https://doi.org/10.1007/BF01584237

BERTSEKAS, D. P. (1998). *Network Optimization*. Athena Scientific, Belmont, MA.

BERTSEKAS, D. P. and TSENG, P. (1988). The RELAX codes for linear minimum cost network flow problems. *Ann. Oper. Res.* **13** 125–190. MR0950991 https://doi.org/10.1007/BF02288322

COCHRAN, W. G. and RUBIN, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā, A* **35** 417–446.

GLOVER, F. (1967). Maximum matching in a convex bipartite graph. *Naval Res. Logist.* **14** 313–316.

GOEMAN, J. J., SOLARI, A. and STIJNEN, T. (2010). Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Stat. Med.* **29** 2117–2125. MR2756559 https://doi.org/10.1002/sim.4002

HANSEN, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika* **95** 481–488. MR2521594 https://doi.org/10.1093/biomet/asn004

HANSEN, B. B. and KLOPFER, S. O. (2006). Optimal full matching and related designs via network flows. *J. Comput. Graph. Statist.* **15** 609–627. MR2280151 https://doi.org/10.1198/106186006X137047

KORTE, B. and VYGEN, J. (2012). *Combinatorial Optimization*: *Theory and Algorithms*, 5th ed. *Algorithms and Combinatorics* **21**. Springer, Heidelberg. MR2850465 https://doi.org/10.1007/978-3-642-24488-9

LIPSKI, W. JR. and PREPARATA, F. P. (1981). Efficient algorithms for finding maximum matchings in convex bipartite graphs and related problems. *Acta Inform.* **15** 329–346. MR0632418 https://doi.org/10.1007/BF00264533

LU, B., GREEVY, R., XU, X. and BECK, C. (2011). Optimal nonbipartite matching and its statistical applications. *Amer. Statist.* **65** 21–30. MR2899649 https://doi.org/10.1198/tast.2011.08294

PIMENTEL, S. D. (2016). Large, sparse optimal matching with R package rcbalance. *Obs. Stud.* **2** 4–23.

PIMENTEL, S. D., KELZ, R. R., SILBER, J. H. and ROSENBAUM, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *J. Amer. Statist. Assoc.* **110** 515–527. MR3367244 https://doi.org/10.1080/01621459.2014.997879

ROSENBAUM, P. R. (1989). Optimal matching for observational studies. *J. Amer. Statist. Assoc.* **84** 1024–1032.

ROSENBAUM, P. R. (2002). Attributing effects to treatment in matched observational studies. *J. Amer. Statist. Assoc.* **97** 183–192. MR1963391 https://doi.org/10.1198/016214502753479329

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. *Springer Series in Statistics*. Springer, New York. MR2561612 https://doi.org/10.1007/978-1-4419-1213-8

ROSENBAUM, P. R., ROSS, R. N. and SILBER, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Amer. Statist. Assoc.* **102** 75–83. MR2345534 https://doi.org/10.1198/016214506000001059

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

ROSENBAUM, P. R. and SILBER, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *J. Amer. Statist. Assoc.* **104** 1398–1405. MR2750570 https://doi.org/10.1198/jasa.2009.tm08470

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–328.

RUBIN, D. B. (1980). Bias reduction using Mahalanobis metric matching. *Biometrics* **36** 293–298.

SILBER, J. H., ROSENBAUM, P. R., MCHUGH, M. D., LUDWIG, J. M., SMITH, H. L., NIKNAM, B. A., EVEN-SHOSHAN, O., FLEISHER, L. A., KELZ, R. R. et al. (2016). Comparison of the value of nursing work environments in hospitals across different levels of patient risk. *JAMA Surg.* **151** 527–536.

SILBER, J. H., ROSENBAUM, P. R., WANG, W., CALHOUN, S. R., REITER, J. G., EVEN-SHOSHAN, O. and GREELEY, W. J. (2018). Practice style variation in medicaid and non-medicaid children with complex chronic conditions undergoing surgery. *Ann. Surg.* **267** 392–400.

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313

YANG, D., SMALL, D. S., SILBER, J. H. and ROSENBAUM, P. R. (2012). Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics* **68** 628–636. MR2959630 https://doi.org/10.1111/j.1541-0420.2011.01691.x

ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* **107** 1360–1371. MR3036400 https://doi.org/10.1080/01621459.2012.703874

ZUBIZARRETA, J. R., REINKE, C. E., KELZ, R. R., SILBER, J. H. and ROSENBAUM, P. R. (2011). Matching for several sparse nominal variables in a case-control study of readmission following surgery. *Amer. Statist.* **65** 229–238. MR2867507 https://doi.org/10.1198/tas.2011.11072