Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

Optimal Matching of an Optimally Chosen Subset in Observational Studies

Paul R. ROSENBAUM

An algorithm is proposed for optimally matching to controls an optimally chosen subset of treated subjects. The algorithm makes three optimal decisions at once: (i) the number of treated subjects to match, (ii) the identity of the treated subjects to match, and (iii) the identity of the controls with whom they are paired. The algorithm finds an optimal assignment for an augmented distance matrix. An example from critical care medicine is considered in detail. An R-workspace is available under "supplemental materials"; it can reproduce the matches in the example.

Key Words: Assignment algorithm; Matched sample; Network optimization.

1. INTRODUCTION: MATCHING WHEN SOME TREATED SUBJECTS ARE FAR FROM ALL POTENTIAL CONTROLS

When the goal is to estimate the effect of a treatment on a well-defined population, there is little choice but to study that specific population. For instance, if one wishes to estimate the effect of a treatment on the type of person who typically receives the treatment, then a matching algorithm that alters the population may remove one bias due to covariate imbalance while introducing another bias due to incomplete matching, where the latter can be substantial (Rosenbaum and Rubin 1985a). Often, however, the available data are simply that, available data, and the data do not represent a natural population, so there is no compelling reason to estimate the effect of the treatment on all people recorded in this source of data. For instance, in a medical context, virtually all patients of certain types may receive a particular treatment and virtually no patient of certain other types may receive this treatment, but there may be a marginal group of patients who may or may not receive the treatment depending upon circumstances such as availability, preference, or heterogeneous opinion on the part of the medical profession. In such an instance, one might wish to estimate the effect of a treatment on marginal patients, that is, patients who might or might not

Paul R. Rosenbaum is Robert G. Putzel Professor, Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *rosenbaum@wharton.upenn.edu*).

Journal of Computational and Graphical Statistics, Volume 21, Number 1, Pages 57–71
DOI: 10.1198/jcgs.2011.09219

receive this treatment. Study of the marginal patient may guide treatment of the marginal patient, and it may also serve to initiate questions about a consensus of opinion concerning the treatment for patients not deemed marginal. That is, study of the marginal patient may shift where that margin is, so that over a period of time, a sequence of studies may gradually shift a consensus. Study of the marginal patient is attractive in the specific but not inconsequential sense that there is the realistic hope of locating patients who look similar in terms of observed covariates yet who received different, competing treatments. Crump et al. (2009) discussed related issues.

An example of this sort is found in an excellent study by Connors et al. (1996) of the effect of right heart catheterization (RHC) as a diagnostic procedure for critically ill patients newly arrived in an intensive care unit. This diagnostic procedure is intended to guide the therapy for some groups of seriously ill patients, but as with any form of cardiac catheterization, the procedure has risks of its own. Because RHC is highly relevant to certain forms of critical care and largely irrelevant to others, there are groups of patients highly likely to receive it and others highly unlikely to receive it, and estimating the effect of RHC on these groups is difficult, because most patients of a given type received the same treatment. There is, however, a group of marginal patients who may or may not receive RHC, and for this group there are, at least, both treated and control patients to compare. Beginning with 2184 patients who received RHC and 3551 patients who did not, Connors et al. (1996) formed 1008 matched pairs consisting of one RHC patient and one control, matching for several dozen covariates using a propensity score. The propensity score estimated the probability of RHC given the covariates using a logit model. In effect, they looked at $1008/2184 = 46\%$ of the RHC patients, excluding the rest because of the difficulty in finding comparable controls. Their study raised the concern that RHC may do more harm than good, increasing the mortality rate, at least for the marginal patient, that is, for the type of patient who might or might not have received RHC. Although this does not affect the conclusions reached by Connors et al., a part of their data will be used here to illustrate a new method for optimally matching an optimally selected portion of a treated group, and in this illustration about 71% of treated subjects are extremely well matched.

The current article develops an algorithm for optimal matching of an optimally selected subset of treated subjects. There are two goals, which are at odds with one another: (i) to match as many treated subjects as possible, recognizing that some treated subjects may be too extreme to match, and (ii) to match as closely as possible on the propensity score and other variables with a view to balancing many covariates. The algorithm must make three simultaneous decisions: (i) how many treated subjects to match, (ii) which specific treated subjects to match, and (iii) which controls to pair with these treated subjects. As is typically true in the solution of combinatorial optimization problems, the algorithm proceeds by transforming this new problem into another standard problem for which a fast algorithm exists. Section 2 discusses the algorithm in general terms while Section 3 presents the illustrative example.

2. OPTIMAL MATCHING OF AN OPTIMALLY CHOSEN SUBSET

2.1 REVIEW OF THE OPTIMAL ASSIGNMENT ALGORITHM

There are T treated subjects, $\mathcal{T} = \{\tau_1, \ldots, \tau_T\}$, and C potential controls, $\mathcal{C} = \{\gamma_1, \ldots, \gamma_C\}$, with $C \geq T$. Each subject has a vector $\mathbf{x}$ of observed covariates, and there is a distance $\delta_{\tau_t,\gamma_c} \geq 0$ between the covariates for τ_t and γ_c, $t = 1, \ldots, T$, $c = 1, \ldots, C$, recorded in a $T \times C$ matrix $\mathbf{\Delta}$. If A is a finite set, write $|A|$ for the number of elements of A, so $|\mathcal{T}| = T$. Define $\delta_{\min}$ to be the minimum δ_{τ_t,γ_c}, $\tau_t \in \mathcal{T}$, $\gamma_c \in \mathcal{C}$.

The optimal assignment algorithm pairs each τ_t with a different γ_c to minimize the sum of the T distances for paired individuals. That is, an assignment $\alpha(\cdot)$ pairs each row of $\mathbf{\Delta}$ with a different column, $\alpha : \mathcal{T} \rightarrow \mathcal{C}$ such that $\tau_t \neq \tau_{t'}$ implies $\alpha(\tau_t) \neq \alpha(\tau_{t'})$; therefore, there are $C(C-1)(C-2)\cdots(C-T+1)$ possible assignments, and an optimal assignment is one that minimizes $\sum_{t=1}^{T} \delta_{\tau_t,\alpha(\tau_t)}$. The problem is not trivial because two or more controls or columns may be closest to the same treated subject or row, so one cannot simply assign each treated subject to the nearest control. See the article by Kuhn (1955) for the first solution to the assignment problem, known as the Hungarian method; see the article by Dell'Amico and Toth (2000) for a recent review and comparison of algorithms for the assignment problem, and see the work of Rosenbaum (1989, 2010) for discussion of optimal matching in observational studies. One very good solution to the optimal assignment problem is due to Bertsekas (1981, 1991), whose Fortran code is accessible inside R (R Development Core Team 2007) using the `pairmatch` function of the `optmatch` package; see the works of Hansen and Klopfer (2006) and Hansen (2007). Carpaneto and Toth (1980) provided Fortran code for the Hungarian method. Bergstralh, Kosanke, and Jacobsen (1996) discussed optimal assignment in SAS. For textbook discussions of the assignment problem, see the works of Bertsekas (1991), Burkard, Dell'Amico, and Martello (2009), Cook et al. (1998), and Papadimitriou and Steiglitz (1982, section 11.2). In what follows, there will be many references to "an assignment" or to "an optimal assignment," perhaps for matrices other than $\mathbf{\Delta}$, and the definitions in this paragraph apply without further mention. For some algorithms, there is a worst-case time bound that is of order $O(C^3)$, which is the same order as the time required to multiply two $C \times C$ matrices in the conventional way; see the book by Papadimitriou and Steiglitz (1982, section 11.2).

Table 1 is a small illustration. The distances are for the first five treated subjects and six potential controls in the data of Connors et al. (1996); see Section 3. The distance is a rank-based Mahalanobis distance using 65 covariates, with a caliper on the estimated propensity score, where the caliper is enforced by a penalty function; see the book by Rosenbaum (2010, section 8) for specifics of this distance function and its rationale, and see the article by Rosenbaum and Rubin (1985b) for a related distance. The propensity score was estimated using a logit model with the same covariates used by Connors et al. (1996) in their propensity score. The optimal assignment of all $T = 5$ treated subjects to $C = 5$ distinct controls in Table 1 is in bold with a total distance of 766. A greedy algorithm (or best-first algorithm or nearest available algorithm) would begin with the smallest distance, $\delta_{\tau_2,\gamma_4} = 66$, then pick the best control from those remaining, $\delta_{\tau_3,\gamma_5} = 83$, ending with a substantially inferior total distance of $66 + 83 + 144 + 209 + 430 = 932$. The

Table 1. A 5 × 6 distance matrix $\mathbf{\Delta}$ and an optimal assigment. The optimal assignment is in bold, with total distance $84+185+143+144+210=766$. Potential control γ_2 is not matched. A greedy or best-first algorithm has total distance $66+83+144+209+430=932$ and is substanially inferior.

Treated subjects	Potential controls					
	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6
τ_1	156	515	380	225	**84**	209
τ_2	85	297	**185**	66	172	77
τ_3	110	469	354	**143**	83	119
τ_4	**144**	518	401	214	100	228
τ_5	198	557	430	239	124	**210**

last control selected by the greedy algorithm contributed distance $\delta_{\tau_5,\gamma_3}=430$ which is $430/766=56\%$ of the total of five distances for the optimal assignment.

2.2 THE PROBLEM STATED: OPTIMAL MATCHING OF AN OPTIMALLY CHOSEN SUBSET

Now it may happen that some treated subjects τ_t are very difficult to match because δ_{τ_t,γ_c} is large for most or all $\gamma_c\in\mathcal{C}$. In light of this, consideration is given to matching a reduced subset $\mathcal{T}_r\subseteq\mathcal{T}$ of the treated subjects, that is, to determining an assignment for a submatrix of $\mathbf{\Delta}$ whose rows are in $\mathcal{T}_r$. Several interrelated decisions are needed: (i) how many treated subjects to match, that is, the number $|\mathcal{T}_r|$, (ii) specifically which treated subjects to match, that is, the set $\mathcal{T}_r$, and finally (iii) which controls to match to these treated subjects, that is, an assignment $\alpha:\mathcal{T}_r\rightarrow\mathcal{C}$ such that $t\neq t'$ implies $\alpha(t)\neq\alpha(t')$. The pair $(\mathcal{T}_r,\alpha)$ defines a match. Write $\mu(\mathcal{T}_r,\alpha)$ for the average distance within pairs for match $(\mathcal{T}_r,\alpha)$, that is, $\mu(\mathcal{T}_r,\alpha)=(1/|\mathcal{T}_r|)\sum_{\tau_t\in\mathcal{T}_r}\delta_{\tau_t,\alpha(\tau_t)}$.

There are two goals pulling in opposite directions. One goal is to match most or all treated subjects, that is, to make $|\mathcal{T}_r|$ large. The other goal is to make the typical distance within pairs small, that is, to make $\mu(\mathcal{T}_r,\alpha)$ small. For instance, one can achieve $\delta_{\min}=\mu(\mathcal{T}_r,\alpha)$ by matching just one treated subject, $|\mathcal{T}_r|=1$. More generally, for any match $(\mathcal{T}_r,\alpha)$ with $|\mathcal{T}_r|>1$ such that the $|\mathcal{T}_r|$ distances $\delta_{\tau_t,\alpha(\tau_t)}$, $\tau_t\in\mathcal{T}_r$, within the $|\mathcal{T}_r|$ pairs are not all equal, it is possible to find another match $(\mathcal{T}_r^\dagger,\alpha^\dagger)$ with one fewer treated subjects, $|\mathcal{T}_r^\dagger|=|\mathcal{T}_r|-1$, such that $\mu(\mathcal{T}_r^\dagger,\alpha^\dagger)<\mu(\mathcal{T}_r,\alpha)$; simply remove from $\mathcal{T}_r$ the treated subject τ_j for whom $\delta_{\tau_t,\alpha(\tau_t)}$ is largest and take $\alpha^\dagger(\tau_t)=\alpha(\tau_t)$ for the remaining elements of $\mathcal{T}_r^\dagger$. So some form of compromise or trade-off between the two goals is needed. It is clear that the only cases in which a trade-off arises are those in which a closer match is obtained by matching fewer treated subjects, that is, those cases in which we have two matches, say $(\mathcal{T}_r,\alpha)$ and $(\mathcal{T}_r^*,\alpha^*)$, with $\mu(\mathcal{T}_r,\alpha)>\mu(\mathcal{T}_r^*,\alpha^*)$ and $|\mathcal{T}_r|>|\mathcal{T}_r^*|$.

A first thought is to select a number $\widetilde{\delta}\geq 0$ and to draw the line at $\widetilde{\delta}$. Specifically, the first thought is to say that a matching $(\mathcal{T}_r,\alpha)$ is unacceptable if either (i) one of the matched pairs has a distance greater than $\widetilde{\delta}$, that is, $\delta_{\tau_t,\alpha(\tau_t)}>\widetilde{\delta}$ for some $\tau_t\in\mathcal{T}_r$, or (ii) there is an unmatched treated subject $\tau_t\in\mathcal{T}-\mathcal{T}_r$ and an unmatched control $\gamma_c\notin\{\alpha(\tau_t):\tau_t\in\mathcal{T}_r\}$ such that $\delta_{\tau_t,\gamma_c}<\widetilde{\delta}$. This first thought is correct in spirit, but it does not quite work in detail.

There are two issues. First, in altering a match, we can not only add or delete a pair, but we can change who is matched to whom, so we must be able to compare matched samples that pair individuals differently. Second, we may want the option of insisting that $|\mathcal{T}_r| \geq n$ for some fixed n even if some distances are therefore large.

A better requirement is that if $|\mathcal{T}_r| < |\mathcal{T}_r^*|$, then we prefer $(\mathcal{T}_r, \alpha)$ to $(\mathcal{T}_r^*, \alpha^*)$, written $(\mathcal{T}_r, \alpha) \succ (\mathcal{T}_r^*, \alpha^*)$, if

$$\frac{\sum_{\tau_t \in \mathcal{T}_r^*} \delta_{\tau_t, \alpha^*(\tau_t)} - \sum_{\tau_t \in \mathcal{T}_r} \delta_{\tau_t, \alpha(\tau_t)}}{|\mathcal{T}_r^*| - |\mathcal{T}_r|} > \widetilde{\delta}, \tag{2.1}$$

we prefer $(\mathcal{T}_r^*, \alpha^*)$ to $(\mathcal{T}_r, \alpha)$, written $(\mathcal{T}_r^*, \alpha^*) \succ (\mathcal{T}_r, \alpha)$, if the inequality in (2.1) is reversed, and we are indifferent if there is equality in (2.1), written $(\mathcal{T}_r^*, \alpha^*) \sim (\mathcal{T}_r, \alpha)$. Also, write $(\mathcal{T}_r, \alpha) \succsim (\mathcal{T}_r^*, \alpha^*)$ if either $(\mathcal{T}_r^*, \alpha^*) \succ (\mathcal{T}_r, \alpha)$ or $(\mathcal{T}_r^*, \alpha^*) \sim (\mathcal{T}_r, \alpha)$. In words, in passing from $(\mathcal{T}_r, \alpha)$ to $(\mathcal{T}_r^*, \alpha^*)$, we are matching $|\mathcal{T}_r^*| - |\mathcal{T}_r|$ more treated subjects, and the total distance is changing by $\sum_{\tau_t \in \mathcal{T}_r^*} \delta_{\tau_t, \alpha^*(\tau_t)} - \sum_{\tau_t \in \mathcal{T}_r} \delta_{\tau_t, \alpha(\tau_t)}$ where the number of terms in the first sum is $|\mathcal{T}_r^*| - |\mathcal{T}_r|$ terms greater than in the second sum, so the left side of (2.1) is the change in total distance per treated subject added. Expressed informally, the ordering $\succ$ prefers more treated subjects if their average increase in distance is less than $\widetilde{\delta}$ and prefers fewer treated subjects if their average increase in distance is more than $\widetilde{\delta}$, so $\widetilde{\delta}$ is the distance at which there is indifference. The "first thought" above referred to the simple case in which $|\mathcal{T}_r^*| = |\mathcal{T}_r| + 1$ and $\mathcal{T}_r \subset \mathcal{T}_r^*$. In contrast to the "first thought," the relation $\succ$ compares any two matches $(\mathcal{T}_r, \alpha)$ and $(\mathcal{T}_r^*, \alpha^*)$ with $|\mathcal{T}_r| \neq |\mathcal{T}_r^*|$.

Fix a number n, $1 \leq n \leq T$, and a number $\widetilde{\delta} \geq 0$, so the relation $\succ$ is well-defined. Problem 1 requires at least n treated subjects to be matched, $|\mathcal{T}_r| \geq n$, and among all such matched samples requires the selected sample to be the best possible in terms of the preference $\succsim$, and among all such matched samples requires the total distance to be a minimum. The solution to Problem 1 may decide to take $|\mathcal{T}_r| = n$ or $|\mathcal{T}_r| > n$ and is guided in this choice by the preference $\succsim$.

Problem 1: For fixed n, $1 \leq n \leq T$, and $\widetilde{\delta} \geq 0$, find a match $(\mathcal{T}_r, \alpha)$ with $|\mathcal{T}_r| \geq n$ such that if $(\mathcal{T}_r^*, \alpha^*)$ is any other match with $|\mathcal{T}_r^*| \geq n$, then:

(i) if $|\mathcal{T}_r| \neq |\mathcal{T}_r^*|$, then $(\mathcal{T}_r, \alpha) \succsim (\mathcal{T}_r^*, \alpha^*)$,

(ii) if $|\mathcal{T}_r| = |\mathcal{T}_r^*|$, then $\sum_{\tau_t \in \mathcal{T}_r} \delta_{\tau_t, \alpha(\tau_t)} \leq \sum_{\tau_t \in \mathcal{T}_r^*} \delta_{\tau_t, \alpha^*(\tau_t)}$.

2.3 SOLUTION USING AN AUGMENTED DISTANCE MATRIX IN THE ASSIGNMENT ALGORITHM

The following simple algorithm solves Problem 1.

Algorithm 1: Augment the distance matrix $\mathbf{\Delta}$ to form $\mathbf{\Upsilon} = (\mathbf{\Delta}, \mathbf{\Psi})$ where $\mathbf{\Psi}$ has T rows and $K = T - n$ columns with every coordinate $\Psi_{tk} = \widetilde{\delta}$, $t = 1, \ldots, T$, $j = 1, \ldots, K$, where the columns of $\mathbf{\Psi}$ are labeled $\kappa_1, \ldots, \kappa_K$, and $\mathcal{K} = \{\kappa_1, \ldots, \kappa_K\}$. Find an optimal assignment $\eta : \mathcal{T} \to \mathcal{C} \cup \mathcal{K}$ for $\mathbf{\Upsilon}$. Let $\mathcal{T}_r \subseteq \mathcal{T}$ be the treated subjects that are paired to controls, $\mathcal{T}_r = \{\tau_t \in \mathcal{T} : \eta(\tau_t) \in \mathcal{C}\}$. Let $\alpha : \mathcal{T}_r \to \mathcal{C}$ be the restriction of $\eta(\cdot)$ to $\mathcal{T}_r$.

Proposition 1 proves that Algorithm 1 solves Problem 1. The intuition is as follows. In Algorithm 1, up to K of the treated subjects $\tau_t \in \mathcal{T}$ may be paired with elements of $\mathcal{K} = \{\kappa_1, \ldots, \kappa_K\}$ and hence removed from $\mathcal{T}$ when forming $\mathcal{T}_r$, but the "price" for deleting τ_t is $\widetilde{\delta}$, so the algorithm will avoid deleting subjects who can find closely matched controls. Although the algorithm is able to discard K treated subjects, it will often decide to discard fewer than K. Changing $\widetilde{\delta}$ will change the number of treated subjects who are matched, $|\mathcal{T}_r|$. In contrast, reducing n will change the current match only if $|\mathcal{T}_r| = n$ in the current match.

Because $\boldsymbol{\Upsilon}$ has $C + n \leq 2C$ columns and $T \leq C$ rows, the time bound for the above algorithm is of the same order as for the optimal assignment algorithm in Section 2.1, namely $O(C^3)$.

Proposition 1. *Algorithm* 1 *solves Problem* 1.

Proof: Let $\eta^* : \mathcal{T} \to \mathcal{C} \cup \mathcal{K}$ be any assignment, not necessarily optimal, of rows to columns of $\boldsymbol{\Upsilon}$, and let $\mathcal{T}_r^* = \{\tau_t \in \mathcal{T} : \eta^*(\tau_t) \in \mathcal{C}\}$. Let η and $\mathcal{T}_r$ be the result of the optimal assignment in Algorithm 1. For any η^*, the total distance between paired rows and columns of $\boldsymbol{\Upsilon}$ is

$$\theta(\eta^*) = \sum_{\tau_t \in \mathcal{T}} \Upsilon_{\tau_t, \eta^*(\tau_t)} = \sum_{\tau_t \in \mathcal{T}_r^*} \delta_{\tau_t, \eta^*(\tau_t)} + (T - |\mathcal{T}_r^*|)\widetilde{\delta}. \tag{2.2}$$

Because η is an optimal assignment for $\boldsymbol{\Upsilon}$, it follows that $\theta(\eta) \leq \theta(\eta^*)$, so

$$0 \leq \theta(\eta^*) - \theta(\eta) = \left\{ \sum_{\tau_t \in \mathcal{T}_r^*} \delta_{\tau_t, \eta^*(\tau_t)} - |\mathcal{T}_r^*|\widetilde{\delta} \right\} - \left\{ \sum_{\tau_t \in \mathcal{T}_r} \delta_{\tau_t, \eta(\tau_t)} - |\mathcal{T}_r|\widetilde{\delta} \right\}$$

or

$$\sum_{\tau_t \in \mathcal{T}_r^*} \delta_{\tau_t, \eta^*(\tau_t)} - \sum_{\tau_t \in \mathcal{T}_r} \delta_{\tau_t, \eta(\tau_t)} \geq (|\mathcal{T}_r^*| - |\mathcal{T}_r|)\widetilde{\delta}. \tag{2.3}$$

Using (2.3),

$$\text{if} \quad |\mathcal{T}_r^*| > |\mathcal{T}_r| \quad \text{then} \quad \frac{\sum_{\tau_t \in \mathcal{T}_r^*} \delta_{\tau_t, \alpha^*(\tau_t)} - \sum_{\tau_t \in \mathcal{T}_r} \delta_{\tau_t, \alpha(\tau_t)}}{|\mathcal{T}_r^*| - |\mathcal{T}_r|} \geq \widetilde{\delta} \tag{2.4}$$

so $(\mathcal{T}_r, \alpha) \succeq (\mathcal{T}_r^*, \alpha^*)$, whereas

$$\text{if} \quad |\mathcal{T}_r| > |\mathcal{T}_r^*| \quad \text{then} \quad \frac{\sum_{\tau_t \in \mathcal{T}_r} \delta_{\tau_t, \eta(\tau_t)} - \sum_{\tau_t \in \mathcal{T}_r^*} \delta_{\tau_t, \eta^*(\tau_t)}}{|\mathcal{T}_r| - |\mathcal{T}_r^*|} \leq \widetilde{\delta} \tag{2.5}$$

so $(\mathcal{T}_r, \alpha) \succeq (\mathcal{T}_r^*, \alpha^*)$, and finally

$$\text{if} \quad |\mathcal{T}_r| = |\mathcal{T}_r^*| \quad \text{then} \quad \sum_{\tau_t \in \mathcal{T}_r^*} \delta_{\tau_t, \eta^*(\tau_t)} - \sum_{\tau_t \in \mathcal{T}_r} \delta_{\tau_t, \eta(\tau_t)} \geq 0. \tag{2.6}$$

□

3. EXAMPLE: COMPARISON OF FOUR MATCHED SAMPLES

3.1 THE MATCHED SAMPLES

To illustrate, patients under the age of 65 from the work of Connors et al. (1996) will be matched by the procedure in Section 2; among these patients there were 1194 patients who received RHC and 1804 who did not. Age 65 is important because Medicare plays a dominant role above age 65 and little role below age 65, and the form of health insurance is a predictor of RHC. In contrast to Table 1, the initial distance matrix $\mathbf{\Delta}$ is 1194×1804 with 2,153,976 distances; this matrix is then expanded by n columns to form $\mathbf{\Upsilon}$, and the assignment algorithm is applied to $\mathbf{\Upsilon}$. (If one were not illustrating the matching algorithm but actually conducting a study analogous to that of Connors et al., then one would separately match the patients over 65, then combine the two groups of matched pairs, so the resulting pairs would be exactly matched for the indicator of age below 65. The complementary match for age 65 and above is smaller in total and easier to match with relatively more controls to choose from, 990 patients with RHC and 1747 without.)

Four matches will be compared. Pair matching uses all 1194 treated subjects and places no limit on the distances, so it is the same as $n = 1194, \widetilde{\delta} = \infty$. The 5% and 20% quantiles of distances in $\mathbf{\Delta}$ are, respectively, 70.0 and 143.7. The three remaining matches have $(n = 1000, \widetilde{\delta} = 144)$, $(n = 800, \widetilde{\delta} = 144)$, and $(n = 800, \widetilde{\delta} = 70)$. Table 2 shows the resulting number of pairs, $|\mathcal{T}_r|$, for each match and the distribution of distances within matched pairs.

In Table 2, when $n = 800$, the algorithm paired more than 800 treated subjects, $|\mathcal{T}_r| > n$. For the $(n = 800, \widetilde{\delta} = 144)$ match, there were $|\mathcal{T}_r| = 897$ pairs, and for the $(n = 800, \widetilde{\delta} = 70)$ match there were $|\mathcal{T}_r| = 849$ pairs. For $(n = 1000, \widetilde{\delta} = 144)$, the algorithm used $n = |\mathcal{T}_r| = 1000$ pairs but was unable to keep the distances within pairs below 144, that is, $\delta_{\tau_j, \alpha(\tau_j)} > \widetilde{\delta}$ for some $j \in \mathcal{T}_r$.

For $(n = 1000, \widetilde{\delta} = 144)$, the selected match is preferable in the sense of $\preceq$ to all other matches with $|\mathcal{T}_r| \geq n = 1000$, but the requirement that at least one thousand pairs be formed forced some distances to be larger than 144. If $\widetilde{\delta} = 144$, any $n \leq 897$ would produce the match produced for $(n = 800, \widetilde{\delta} = 144)$ with 897 pairs.

The pair match $(n = 1194, \widetilde{\delta} = \infty)$ in Table 2 is not attractive. The largest distance among matched pairs, $\max_{j \in \mathcal{T}} \delta_{\tau_j, \alpha(\tau_j)}$, is 717.4, which is well above the upper quartile

Table 2. Distribution of covariate distances before matching and in four matched comparisons. The distribution before matching refers to all $1194 \times 1804 = 2{,}153{,}976$ distances in Δ.

Match			Distances				
n	$\widetilde{\delta}$	Pairs	Min.	Quartile 1	Median	Quartile 3	Maximum
Before matching			11.5	173.4	331.7	515.1	1115.0
1194	∞	1194	17.2	57.4	109.4	194.7	717.4
1000	144	1000	11.5	39.2	51.5	77.4	272.2
800	144	897	11.5	33.6	40.7	48.0	102.9
800	70	849	11.5	31.9	38.4	44.6	66.7

before matching, namely 515.1, and is an order of magnitude larger than the largest distance within pairs for the ($n = 800, \widetilde{\delta} = 70$) match.

3.2 BALANCE FOR A FEW COVARIATES

Table 3 examines three important covariates among the 65 covariates used in matching. These are the APACHE score or "Acute Physiology And Chronic Health Evaluation" score where higher scores predict a higher probability of death, a primary diagnosis of multiple organ system failure with sepsis, and the propensity score. As is commonly done (e.g., Rosenbaum and Rubin 1985b), two informal diagnostics of covariate imbalance are given. The standardized difference is the difference in means divided by a pooled standard deviation *before* matching. The pooled standard deviation is the square root of the unweighted average of the variances in the treated and control groups before matching. Notice that the same standard deviation is used before and after matching and for all of the matched samples; it is only the means that change. The other informal diagnostic is the P-value from the two-sample t-statistic, which acts as a benchmark for covariate imbalance. In a moderately large completely randomized experiment, this P-value for a covariate whose distribution is not extremely long tailed will be approximately uniform, so in Table 3 it serves to compare the balance on the marginal distribution of a covariate achieved by matching with the balance expected in a completely randomized experiment. The proper comparison here is between a matched sample and a completely randomized experiment of the same size, not a comparison of one matched sample to another of a different size, because

Table 3. Three important covariates of the 65 covariates used in matching.

Match			Treated	Control	Standardized	Two-sample
n	$\widetilde{\delta}$	Pairs	mean	mean	difference	P-value
			APACHE score			
	Before matching		61.07	50.51	0.51	0.00000
1194	∞	1194	61.07	54.44	0.32	0.00000
1000	144	1000	57.98	55.58	0.12	0.00915
800	144	897	56.90	55.96	0.04	0.34134
800	70	849	56.66	56.12	0.03	0.59539
			Multiple organ system failure with sepsis (1 = Yes, 0 = No)			
	Before matching		0.31	0.16	0.37	0.00000
1194	∞	1194	0.31	0.22	0.22	0.00000
1000	144	1000	0.28	0.24	0.10	0.02788
800	144	897	0.26	0.24	0.06	0.22968
800	70	849	0.26	0.24	0.03	0.57529
			Propensity score			
	Before matching		0.55	0.30	1.19	0.00000
1194	∞	1194	0.55	0.40	0.75	0.00000
1000	144	1000	0.50	0.43	0.33	0.00000
800	144	897	0.48	0.45	0.13	0.00576
800	70	849	0.47	0.46	0.07	0.14972

P-values tend to be smaller for the same covariate imbalance when the sample size is larger.

In Table 3, although pair matching ($n = 1194, \widetilde{\delta} = \infty$) substantially reduces the imbalance in these three covariates, the imbalance remaining after matching is unacceptably large. For instance, for the propensity score, the bias before matching was about 1.2 standard deviations, and after matching about 0.75 standard deviations. With $n < 1194$, the covariate means in the treated group are changing, because not all treated subjects are matched. In Table 3, only the match with 849 pairs exhibits good balance on all three covariates, so that $849/1194 = 71\%$ of treated subjects were matched. In the match with 849 pairs, the mean estimated propensity scores (or estimated probabilities of RHC) are 0.47 and 0.46 in treated and control groups, respectively.

Figure 1 parallels Table 3, comparing the imbalance in three covariates in the 1194 pairs in the optimal pair match with ($n = 1194, \widetilde{\delta} = \infty$) and the 849 pairs in the optimal subset match with ($n = 800, \widetilde{\delta} = 70$). The covariates are the APACHE score, blood pressure, and the propensity score. The distributions of these three covariates are in much better balance in the optimal subset match.

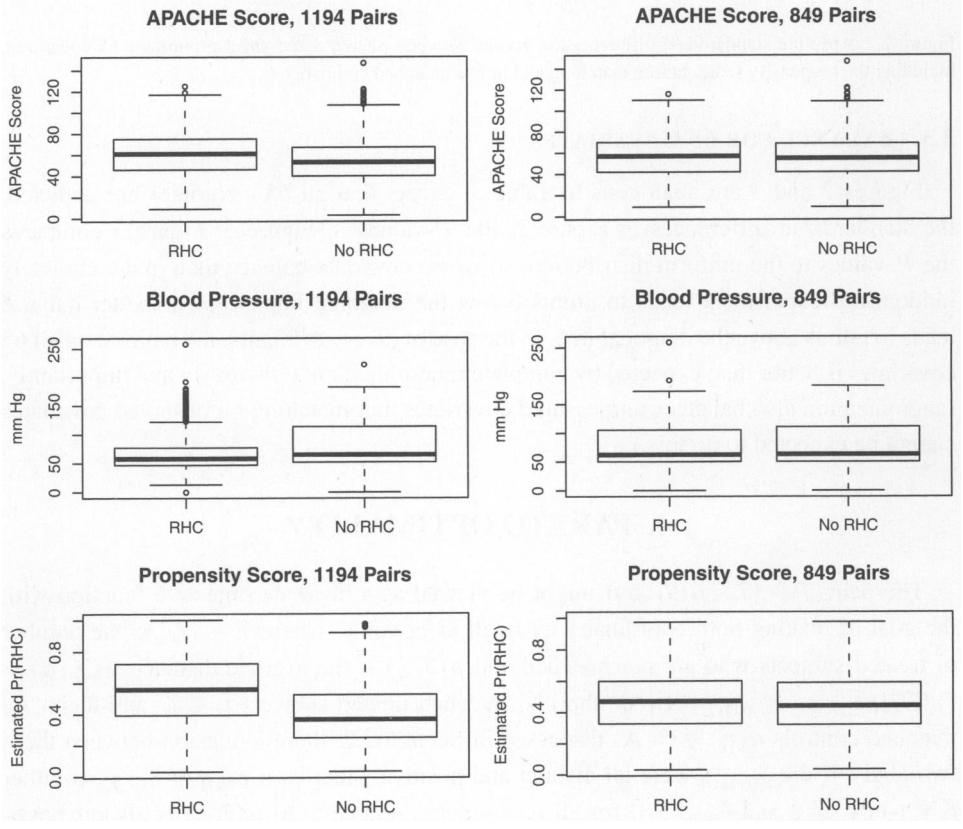

Figure 1. Comparison of the imbalance in three covariates in the best pair match with 1194 pairs and the best subset match with 849 pairs. The covariates are the APACHE score, blood pressure, and the propensity score.

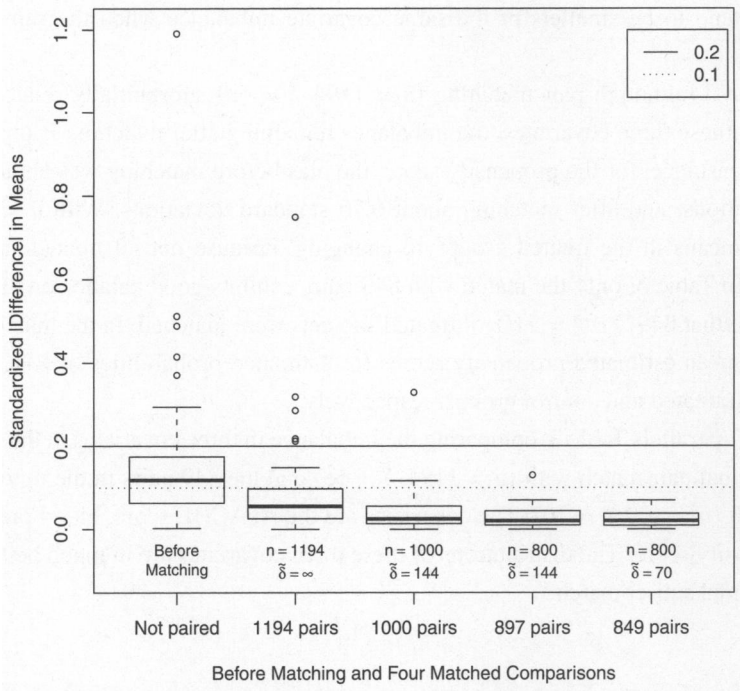

Figure 2. Absolute standardized differences in means between treated and control groups for 65 covariates, including the propensity score, before matching and in four matched comparisons.

3.3 BALANCE FOR 65 COVARIATES

Figures 2 and 3 are analogous to Table 3, except that all 65 covariates are depicted, the standardized differences in Figure 2, the P-values in Figure 3. Figure 3 compares the P-values to the uniform distribution, so worse covariate balance than in a completely randomized experiment leads to points below the diagonal line, whereas better balance leads to points above the diagonal line. In the case of $|\mathcal{T}_r| = 849$ pairs, the balance on all 65 covariates is better than expected by complete randomization. (Obviously and importantly, randomization also balances unmeasured covariates, but matching on observed covariates cannot be expected to do this.)

4. PARETO OPTIMALITY

The pair $\{T - |\mathcal{T}_r|, \mu(\mathcal{T}_r, \alpha)\}$ might be viewed as a bivariate objective function with the goal of making both coordinates as small as possible, where $T - |\mathcal{T}_r|$ is the number of treated subjects who are not matched and $\mu(\mathcal{T}_{r,\alpha})$ is the average distance $\mu(\mathcal{T}_r, \alpha) = (1/|\mathcal{T}_r|)\sum_{\tau_t \in \mathcal{T}_r} \delta_{\tau_t, \alpha(\tau_t)}$ between the $|\mathcal{T}_r|$ matched treated subjects $\tau_t \in \mathcal{T}_r$ and their $|\mathcal{T}_r|$ matched controls $\alpha(\tau_t) \in \mathcal{C}$. As discussed in Section 2.2, there is tension between these two goals: if the δ_{τ_t,γ_c}'s were all distinct and positive—that is, if $\delta_{\tau_t,\gamma_c} \neq \delta_{\tau_{t'},\gamma_{c'}}$ if either $t' \neq t$ or $c' \neq c$ and $\delta_{\tau_t,\gamma_c} > 0$ for all t, c—then a reduction in $\mu(\mathcal{T}_{r,\alpha})$ is always possible by increasing $T - |\mathcal{T}_r|$. For instance, the δ_{τ_t,γ_c}'s would be distinct and positive with probability 1 if they were Mahalanobis distances among covariate vectors with continuous

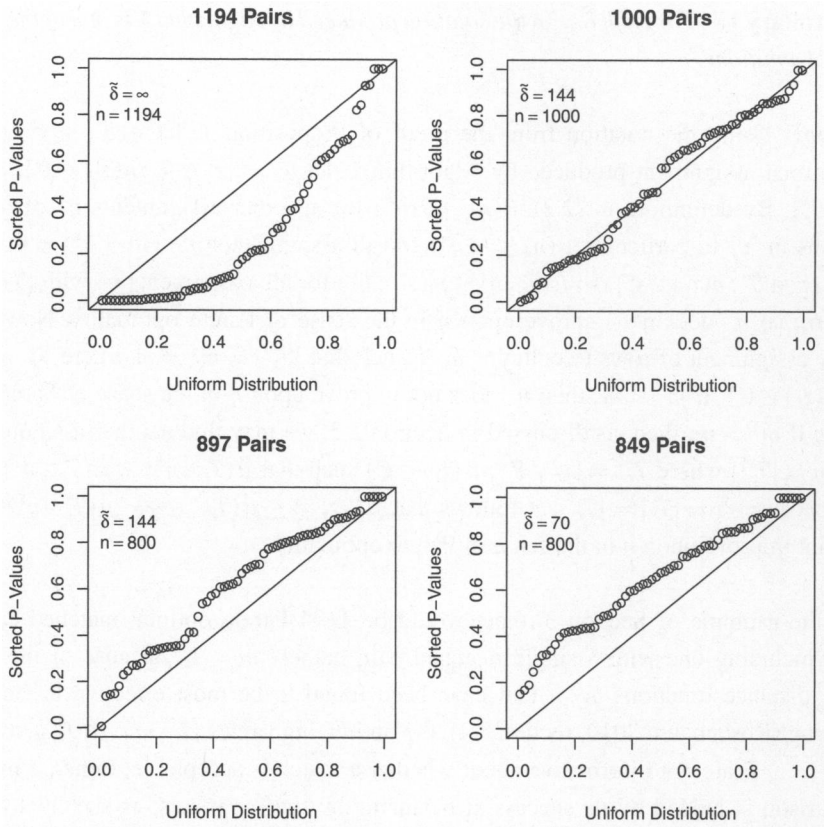

Figure 3. Quantile–quantile plots comparing the two-sample P-values for 65 covariates to the uniform distribution in four matched comparisons.

distributions. For the theoretical point made in the current section, the δ_{τ_t,γ_c}'s are assumed without further mention to be distinct and positive.

A solution $(\mathcal{T}_r, \alpha)$ is Pareto optimal (e.g., Karlin 1959, vol. I, p. 294) if there is no other solution that is uniformly better, that is, no other solution $(\mathcal{T}_r^*, \alpha^*)$ with $T - |\mathcal{T}_r^*| \le T - |\mathcal{T}_r|$ and $\mu(\mathcal{T}_r^*, \alpha^*) < \mu(\mathcal{T}_r, \alpha)$ or with $T - |\mathcal{T}_r^*| < T - |\mathcal{T}_r|$ and $\mu(\mathcal{T}_r^*, \alpha^*) \le \mu(\mathcal{T}_r, \alpha)$. Typically with a bivariate objective function there are many Pareto optimal solutions, and in particular that is true in the case of $\{T - |\mathcal{T}_r|, \mu(\mathcal{T}_r, \alpha)\}$, so Pareto optimality may be one consideration in selecting the one matched sample that will actually be used, but it cannot be the only consideration. Because the δ_{τ_t,γ_c} are distinct and positive, the collection of the T Pareto optimal solutions $(\mathcal{T}_{r,m}, \alpha_m)$, $m = 1, \ldots, T$, has $|\mathcal{T}_{r,m}| = m$ with α_m such that $\mu(\mathcal{T}_{r,m}, \alpha_m)$ is minimized among solutions $(\mathcal{T}_r^*, \alpha^*)$ with $|\mathcal{T}_{r,m}^*| = m$. For a fixed m, Dell'Amico and Martello (1997) discussed finding $(\mathcal{T}_{r,m}, \alpha_m)$ in an efficient manner. One approach to finding $(\mathcal{T}_{r,m}, \alpha_m)$ uses Proposition 1 with $\widetilde{\delta} = 0$: Let $\mathbf{O}_{T-m}$ be the zero matrix with T rows and $T - m$ columns, $1 \le m \le T$; then by Proposition 1, $(\mathcal{T}_{r,m}, \alpha_m)$ may be found by solving the assignment problem for the matrix $(\mathbf{\Delta}, \mathbf{O}_{T-m})$ and discarding the $T - m$ rows or treated subjects τ_t that are paired with columns of $\mathbf{O}_{T-m}$.

Corollary 1. *For each* $\widetilde{\delta} > 0$, *the solution produced by Algorithm* 1 *is one of the Pareto optimal solutions.*

Proof: Using the notation from the proof of Proposition 1, let $\eta : \mathcal{T} \rightarrow \mathcal{C} \cup \mathcal{K}$ be the optimal assignment produced by Algorithm 1, let $\mathcal{T}_r = \{\tau_t \in \mathcal{T} : \eta(\tau_t) \in \mathcal{C}\}$, and let $m = |\mathcal{T}_r|$. By definition, in (2.2), $\theta(\eta) \leq \theta(\eta^*)$ for all other assignments η^* of rows to columns in Υ. In particular, $\theta(\eta) \leq \theta(\eta^*)$ for all assignments η^* with $|\mathcal{T}_r^*| = m$ where $\mathcal{T}_r^* = \{\tau_t \in \mathcal{T} : \eta(\tau_t) \in \mathcal{C}\}$, so $\mu(\mathcal{T}_r, \eta) \leq \mu(\mathcal{T}_r^*, \eta^*)$ for all assignments η^* with $|\mathcal{T}_r^*| = m$ by (2.6), so η^* does not improve upon η in the sense of Pareto optimality. Now let $\eta^\dagger$ be any assignment of rows to columns in Υ such that $|\mathcal{T}_r^\dagger| = m^\dagger \neq m$ where $\mathcal{T}_r^\dagger = \{\tau_t \in \mathcal{T} : \eta^\dagger(\tau_t) \in \mathcal{C}\}$. If $m^\dagger < m$, then $\eta^\dagger$ does not improve upon η in the sense of Pareto optimality. If $m^\dagger > m$, then, as discussed in Section 2.2, we may find another assignment $\eta^\ddagger$ with $m = |\mathcal{T}_r^\ddagger|$ where $\mathcal{T}_r^\ddagger = \{\tau_t \in \mathcal{T} : \eta^\ddagger(\tau_t) \in \mathcal{C}\}$ such that $\mu(\mathcal{T}_r^\ddagger, \eta^\ddagger) < \mu(\mathcal{T}_r^\dagger, \eta^\dagger)$; however, because $m = |\mathcal{T}_r| = |\mathcal{T}_r^\ddagger|$, it follows that $\mu(\mathcal{T}_r, \eta) \leq \mu(\mathcal{T}_r^\ddagger, \eta^\ddagger) < \mu(\mathcal{T}_r^\dagger, \eta^\dagger)$, so $\eta^\dagger$ does not improve upon η in the sense of Pareto optimality. □

In the example of Section 3, there would be 1194 Pareto optimal matched comparisons, including one with a single matched pair, namely $m = 1$. Because of the nature of the distance functions $\delta_{\tau_t, \gamma_c}$ that have been found to be most effective in statistical matching (Rosenbaum 2010, section 8.4), the optimizing values $\{T - m, \mu(\mathcal{T}_{r,m}, \alpha_m)\}$ for $m = 1, \ldots, T$ are not informative about whether a match is acceptable; rather, a matched comparison is judged by its success at balancing all the covariates, as sketched for just three of the 65 covariates in Section 3.2. It is not practical for a clinical research team to examine the balance of 65 covariates in each of 1194 matched comparisons. Nonetheless, Corollary 1 indicates that if one compares a few matched samples produced by Algorithm 1 with different values of $\widetilde{\delta}$, as was done in Section 3, then one is comparing a few matched samples no one of which can be improved in the sense of Pareto optimality.

5. DISCUSSION: SUMMARY, PRACTICALITIES, OTHER APPLICATIONS

If there are relatively standard criteria for using or not using a treatment, there may be many treated subjects unlike all controls and controls unlike all treated subjects, and these may be of little or no value in attempting to estimate the effect of the treatment. Nonetheless, there may be a category of marginal subjects, who might or might not have been treated, and it may not be unreasonable to attempt to estimate the treatment effect in this marginal category. Arguably, this was true in the study of right heart catheterization by Connors et al. (1996), and their match focused on 46% of the patients treated with RHC. The algorithm proposed in the current article tries to find a match that is as close as possible yet matching as many treated subjects as possible, in the sense formalized in Problem 1. The solution entails solving the optimal assignment problem for an augmented distance matrix.

As seen clearly in Table 3, use of Algorithm 1 alters the treated group to focus on the marginal patient who might or might not have received the treatment. For instance, in the match with 849 pairs, the APACHE scores for treated subjects are lower and for the controls are higher than in the unmatched population, but they are similar in the treated and control matched pairs. The same pattern is seen for the other two covariates in Table 3. Because the algorithm itself is defining the marginal patient population, an empirical study would need to describe the covariate distributions in the final matched pairs, so that readers would know the nature of this marginal patient population. For instance, Connors et al. (1996) did this, albeit with a different matching algorithm.

Generally, it will be informative to compare a few matched comparisons derived from different values of $(n, \widetilde{\delta})$, as was done in Section 3. How should $(n, \widetilde{\delta})$ be chosen for this comparison? Typically, one match would retain all of the treated subjects, $n = T$; perhaps this match will be acceptable, or perhaps it will simply serve as a benchmark for comparison. In Section 3, $\widetilde{\delta}$ was set to either the 5% or 20% quantiles of distances in $\mathbf{\Delta}$, in the example 70.0 or 143.7. The use of low quantiles of distances in $\mathbf{\Delta}$ to specify $\widetilde{\delta}$ is practical because we know that substantial numbers of controls are available at these distances. The distance functions used to define $\mathbf{\Delta}$ often combine a Mahalanobis distance with various penalty functions for violations of constraints such as calipers on the propensity score (e.g., Rosenbaum 2010, section 8.4), and in this case one might consider a value of $\widetilde{\delta}$ that suggests the absence of material violations of the constraints. If some version of the Mahalanobis distance (i.e., $\delta_{\tau_t,\gamma_c} = (\mathbf{x}_t - \mathbf{x}_c)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_t - \mathbf{x}_c)$) is being used inside calipers, then it is often useful to note that two independent observations (say $\mathbf{x}_t$ and $\mathbf{x}_c$) drawn from the same P-dimensional multivariate Normal distribution with covariance matrix $\mathbf{\Sigma}$ have expected Mahalanobis distance of $\mathrm{E}(\delta_{\tau_t,\gamma_c}) = 2P$. In other words, if there were no systematic bias from different covariate means in treated and control groups, then the Mahalanobis distance expected by chance in a single multivariate Normal population is $2P$, so for $P = 65$ covariates as in Section 1 and Section 3, we have $2P = 130$. For instance, in Table 2 where $P = 65$, the $(n, \widetilde{\delta}) = (800, 70)$ match with $|\mathcal{T}_r| = 849$ pairs has maximum distance 66.7, suggesting that (i) the constraints imposed by the calipers were not materially violated and (ii) the Mahalanobis distances are in almost all cases less than half the chance expectation of $2P$ for a multivariate Normal distribution. For a Mahalanobis distance applied within propensity score calipers, two plausible candidates for $\widetilde{\delta}$ are $\widetilde{\delta} = P$ and $\widetilde{\delta} = 2P$.

Matching is part of the design of an observational study, which means that a match is selected without looking at outcomes. Just as one cannot perform several parallel statistical analyses and report the one analysis that supports a favored conclusion, one cannot choose one matched comparison having looked at outcomes for several matched comparisons.

Algorithm 1 has other uses. Stuart and Rubin (2008) proposed matching with two sources of controls, a better but limited source and a poorer but extensive source; see also the article by Stuart (2010). They tried to first make the best use possible of the better source, then shifted to match from the poorer source, so they ultimately matched all treated subjects, but not always with controls from the better source. Algorithm 1 could be used in the first step of this two-step process to optimize the selection of controls from the better

but limited source. In a related context, there is sometimes a "seemingly innocuous covariate" strongly related to treatment assignment and hence difficult to match, yet plausibly not much related to the outcomes subjects would exhibit under treatment and control; see the book by Rosenbaum (2010, section 18.2) for discussion of an example from the article of Silber et al. (2009). The worry is that the covariate is not innocuous but only seems so; it is, after all, predictive of treatment assignment, and in an economic or medical context that typically means someone thought it mattered for something. When good controls are abundant, tapered matching (Daniel et al. 2008; Heller, Small, and Rosenbaum 2010) creates two matched control groups, one matched for all covariates, the other matched for all covariates except the "seemingly innocuous covariate," thereby permitting examination of this issue. When controls matched for the "seemingly innocuous covariate" are in short supply, Algorithm 1 may be used to create a subset $\mathcal{T}_r$ of pairs matched for all covariates, with the balance $\mathcal{T} - \mathcal{T}_r$ subsequently matched for all covariates except the seemingly innocuous covariate, so again the issue may be explicitly and transparently examined.

SUPPLEMENTARY MATERIALS

R-workspace optsub: The R-workspace optsub contains what is needed to reproduce the matches in the example. The object README in the workspace explains the steps and the contents of the workspace. (supplement.zip)

ACKNOWLEDGMENTS

This work was supported by SES-0849370 from the US National Science Foundation.

[Received December 2009. Revised November 2010.]

REFERENCES

Bergstralh, E. J., Kosanke, J. L., and Jacobsen, S. L. (1996), "Software for Optimal Matching in Observational Studies," *Epidemiology*, 7, 331–332. [59]

Bertsekas, D. P. (1981), "A New Algorithm for the Assignment Problem," *Mathematical Programming*, 21, 152–171. [59]

——— (1991), *Linear Network Optimization*, Cambridge, MA: MIT Press. [59]

Burkard, R., Dell'Amico, M., and Martello, S. (2009), *Assignment Problems*, Philadelphia: SIAM. [59]

Carpaneto, G., and Toth, P. (1980), "Algorithm 548: Solution of the Assignment Problem [H]," *ACM Transactions on Mathematical Software*, 6, 104–111. [59]

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., Fulkerson, W. J., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., and Knaus, W. A. (1996), "The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients," *Journal of the American Medical Association*, 276, 889–897. Available at *http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets*. [58,59,63,68,69]

Cook, W. J., Cunningham, W. H., Pulleyblank, W. R., and Schrijver, A. (1998), *Combinatorial Optimization*, New York: Wiley. [59]

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), "Dealing With Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199. [58]

Daniel, S., Armstrong, K., Silber, J. H., and Rosenbaum, P. R. (2008), "An Algorithm for Optimal Tapered Matching, With Application to Disparities in Survival," *Journal of Computational and Graphical Statistics*, 17, 914–924. [70]

Dell'Amico, M., and Martello, S. (1997), "The k-Cardinality Assignment Problem," *Discrete Applied Mathematics*, 76, 103–121. [67]

Dell'Amico, M., and Toth, P. (2000), "Algorithms and Codes for Dense Assignment Problems: The State of the Art," *Discrete Applied Mathematics*, 100, 17–48. [59]

Hansen, B. B. (2007), "Optmatch," *R News*, 7, 18–24. [59]

Hansen, B. B., and Klopfer, S. O. (2006), "Optimal Full Matching and Related Designs via Network Flows," *Journal of Computational and Graphical Statistics*, 15, 609–627. [59]

Heller, R., Small, D., and Rosenbaum, P. R. (2010), "Using the Cross-Match Test to Appraise Covariate Balance in Matched Pairs," *The American Statistician*, 64, 299–309. [70]

Karlin, S. (1959), *Mathematical Methods and Theory in Games, Programming and Economics*, Vol. I, Reading, MA: Addison-Wesley. [67]

Kuhn, H. W. (1955), "The Hungarian Method for the Assignment Problem," *Naval Research Logistics Quarterly*, 2, 83–97. [59]

Papadimitriou, C., and Steiglitz, K. (1982), *Combinatorial Optimization*, Englewood Cliffs, NJ: Prentice Hall. [59]

R Development Core Team (2007), *R*, Vienna: R Foundation. Available at *http://www.R-project.org*. [59]

Rosenbaum, P. R. (1989), "Optimal Matching in Observational Studies," *Journal of the American Statistical Association*, 84, 1024–1032. [59]

——— (2010), *Design of Observational Studies*, New York: Springer. [59,68-70]

Rosenbaum, P. R., and Rubin, D. B. (1985a), "The Bias Due to Incomplete Matching," *Biometrics*, 41, 106–116. [57]

——— (1985b), "Constructing a Control Group by Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, 39, 33–38. [59,64]

Silber, J. H., Lorch, S. L., Rosenbaum, P. R., Medoff-Cooper, B., Bakewell-Sachs, S., Millman, A., Mi, L., Even-Shoshan, O., and Escobar, G. E. (2009), "Additional Maturity at Discharge and Subsequent Health Care Costs," *Health Services Research*, 44, 444–463. [70]

Stuart, E. A. (2010), "Matching Methods for Causal Inference," *Statistical Science*, 25, 1–21. [69]

Stuart, E. A., and Rubin, D. B. (2008), "Matching to Multiple Control Groups With Adjustment for Group Differences," *Journal of Educational and Behavioral Statistics*, 33, 279–306. [69]