Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Minimum Distance Matched Sampling With Fine Balance in an Observational Study of Treatment for Ovarian Cancer

Paul R Rosenbaum, Richard N Ross & Jeffrey H Silber

To cite this article: Paul R Rosenbaum, Richard N Ross & Jeffrey H Silber (2007) Minimum Distance Matched Sampling With Fine Balance in an Observational Study of Treatment for Ovarian Cancer, Journal of the American Statistical Association, 102:477, 75-83, DOI: 10.1198/016214506000001059

To link to this article: https://doi.org/10.1198/016214506000001059

Minimum Distance Matched Sampling With Fine Balance in an Observational Study of Treatment for Ovarian Cancer

Paul R. ROSENBAUM, Richard N. ROSS, and Jeffrey H. SILBER

In observational studies of treatment effects, matched samples have traditionally been constructed using two tools, namely close matches on one or two key covariates and close matches on the propensity score to stochastically balance large numbers of covariates. Here we propose a third tool, *fine balance*, obtained using the assignment algorithm in a new way. We use all three tools to construct a matched sample for an ongoing study of provider specialty in the treatment of ovarian cancer. Fine balance refers to exact balance of a nominal covariate, often one with many categories, but it does not require individually matched treated and control subjects for this variable. In the example the nominal variable has $72 = 9 \times 8$ categories formed from 9 possible years of diagnosis and 8 geographic locations or SEER sites. We obtain exact balance on the 72 categories and close individual matches on clinical stage, grade, year of diagnosis, and other variables using a distance, and stochastically balance a total of 61 covariates using a propensity score. Our approach finds an optimal match that minimizes a suitable distance subject to the constraint that fine balance is achieved. This is done by defining a special patterned distance matrix and passing it to a subroutine that solves the optimal assignment problem, which optimally pairs the rows and columns of a matrix using a polynomial time algorithm. In the example we used the function Proc Assign in SAS. A new theorem shows that with our patterned distance matrix, the assignment algorithm returns an optimal, finely balanced matched sample whenever one exists, and otherwise returns an infinite distance, indicating that no such matched sample exists.

KEY WORDS: Assignment algorithm; Matched sampling; Observational study; Optimal matching; Propensity score.

1. EXACT BALANCE WITHOUT EXACT MATCHING

In an observational study of treatment effects, matched sampling is often used to select subjects for study who, before to treatment, appeared similar with respect to observed covariates. As one example, the Surgical Outcomes Study (Silber et al. 2001, 2005) examined the effects of preoperative antibiotics on mortality after surgery in the Pennsylvania Medicare population by matching deaths to survivors using many covariates available from Medicare. For 681 matched pairs, paper hospital charts were located throughout Pennsylvania and individually abstracted to determine, among other things, preoperative antibiotics use. Although performing chart abstraction is too expensive for the entire Medicare population, it is essential if information not recorded by Medicare is to be used, and it is practical for a matched sample.

Traditionally, matched samples have been constructed either by finding close matches on one or two key covariates, by finding close matches on the propensity score to stochastically balance large numbers of covariates, or by doing both at once. For a few key covariates, the Mahalanobis distance is often used (Rubin 1980). The propensity score is the conditional probability of exposure to treatment given the *observed* covariates, and samples matched for the propensity score tend to balance, in a stochastic sense, all of the observed covariates used to define the score (Rosenbaum and Rubin 1983, 1985). To a limited extent, the stochastic balance produced by propensity scores resembles the stochastic balance produced by random assignment of treatments, but of course propensity scores can be expected to balance only the observed covariates used to construct the score, whereas random assignment also tends to balance covariates not observed. Typically, the propensity score is estimated from the data; for instance, in Section 5, a logit model is used. These two tactics are often implemented using optimization algorithms (see Rosenbaum 1989; Bergstralh, Kosanke, and Jacobsen 1996; Ming and Rosenbaum 2001; Hansen 2004 for algorithms and Silber et al. 2001; Dreger et al. 2004; Singh et al. 2004 for a few representative applications). Here we develop a new, quick, easy way of implementing a third approach to matching, called *fine balance*, and use it in conjunction with propensity scores and close individual matches to build a matched sample for an ongoing study of provider specialty in the treatment of ovarian cancer. Fine balance refers to exactly balancing a nominal variable, often one with many categories, without trying to match individuals on this variable. A network optimization algorithm for fine balance was proposed by Rosenbaum (1989, sec. 3.2), but it was rarely used, perhaps because the algorithm was not easy to implement. In contrast, the new quick and easy approach that we develop here entails creating a patterned distance matrix and passing it to a subroutine that optimally pairs the rows and columns of a matrix; indeed, in the example, all of the work is done using Proc Assign in SAS.

The assignment algorithm accepts a distance matrix as input and returns a minimum distance pairing of the rows and columns. It is one of the oldest, most widely studied, and most widely implemented combinatorial optimization algorithms. Several implementations are available for download without charge; see the review of the assignment algorithm in Section 3.1 for specifics. In many implementations, the optimal assignment algorithm runs in polynomial time, and good implementations are often quite a bit faster than the suboptimal greedy algorithms that people have sometimes used for matching. A key result of this article is that if the assignment algorithm receives a suitably patterned distance matrix, then it can be used to find an optimal, finely balanced matched sample; see Section 4 for

Paul R. Rosenbaum is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *rosenbaum@stat.wharton.upenn.edu*). Richard N. Ross is Statistician, Center for Outcomes Research, Children's Hospital of Philadelphia, Philadelphia, PA 19104. Jeffrey H. Silber is Professor, Department of Pediatrics, School of Medicine, University of Pennsylvania, Philadelphia, Philadelphia, PA 19104. This work was supported by the National Cancer Institute (grant R01-CA095664) and the National Science Foundation.

Journal of the American Statistical Association
March 2007, Vol. 102, No. 477, Applications and Case Studies
DOI 10.1198/016214506000001059

development of this result. Before discussing the algorithm, we illustrate its use in the example in Section 2. Details of its implementation in the example are deferred to Section 5.

2. EXAMPLE: PROVIDERS OF TREATMENT FOR OVARIAN CANCER

2.1 Data, Background, and Empirical Questions

Chemotherapy for ovarian cancer is routinely provided by two very different types of specialists. Medical oncologists (MOs) are internists who specialize in the treatment of cancer using chemotherapy; gynecologic oncologists (GOs) are gynecologists with additional training that includes surgical oncology and the use of chemotherapy. Currently, most surgery for ovarian cancer is performed by gynecologists or general surgeons, and most chemotherapy is delivered by medical oncologists, but a growing number of women are being treated by GOs. As part of a project funded by the National Cancer Institute (NCI), we obtained data from Medicare that was merged with data from the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute to compare the care provided by MOs and GOs and the outcomes after that care. The study focused on elderly women diagnosed with ovarian cancer between 1991 and 1999 at one of eight SEER sites listed in Table 1. (Perhaps surprisingly, some SEER sites are cities and others are states.) Among the women who met our criteria for inclusion, 344 were treated with chemotherapy by a GO and 2,011 were given chemotherapy by an MO. As a prelude to expensive detailed chart abstraction, we constructed a matched sample of 344 pairs of women, with one woman given chemotherapy by a GO and the other given chemotherapy by an MO. The matching controlled for 61 covariates using new matching technologies which we describe and illustrate in this article. See also Silber et al. (2007).

Plausibly, either an MO or a GO could claim to have the relevant expertise to provide treatment for ovarian cancer (Junor, Hole, MacNulty, Mason, and Young 1999; Heintz 2002). Because cancer is most lethal when it metastasizes—that is, when it spreads beyond its primary site of origin—adequate chemotherapy is considered essential to reduce the recurrence of cancer. GOs are surgeons, and typically they perform the surgery that precedes chemotherapy, whereas MOs are almost invariably not surgeons, so they provide chemotherapy after surgery has been performed by someone else. Cancer chemotherapy is often accompanied by toxic side effects, but the alternative to toxicity may be metastatic cancer and death. Consequently, some toxicity is typically considered acceptable and appropriate, but excessive toxicity may force the patient off chemotherapy or may be harmful or even lethal. Managing the toxic side effects of chemotherapy is a central aspect of care. Effective chemotherapy for cancer requires extensive skill, experience, and up-to-date knowledge; it is unclear whether MOs and GOs differ in these respects and, if so, how they differ. There were substantial improvements in chemotherapy for ovarian cancer during the period covered by this study, with the introduction of platinum/paclitaxel regimes roughly in 1996 (McGuire et al. 1996; Neijt 1996; Harlan, Clegg, and Trimble 2003), and it is of interest whether MOs and GOs differed in the speed with which they adopted the newer drugs. Similarly, the practice of "second-look" surgery (Chu and Rubin 2001) may or may not be effective, and it may possibly be more common among GOs because they are surgeons. When the cancer does recur, MOs and GOs may or may not differ in how aggressively they treat this recurrence of measurable cancer, and the effectiveness of treatment after recurrence is unclear. These are some of the questions that have motivated the present study.

Table 1. Distributions (percent) Before and After Matching for SEER Site and Year of Diagnosis, Both Exactly Balanced by Matching

Covariate	*GO*	*MO-matched*	*MO-all*
Connecticut	18	18	15
Detroit	26	26	12
Iowa	17	17	17
New Mexico	7	7	3
Seattle	9	9	16
Atlanta	9	9	7
Los Angeles	12	12	19
San Francisco	1	1	9
Total	100	100	100
1991	4	4	9
1992	7	7	14
1993	10	10	14
1994	11	11	12
1995	11	11	12
1996	10	10	12
1997	16	16	10
1998	13	13	9
1999	18	18	9
Total	100	100	100

NOTE: GO, patient of a gynecological oncologist, $n = 344$; MO-matched, matched patient of a medical oncologist, $n = 344$; MO-all, patients of medical oncologists before matching, $n = 2{,}011$.

2.2 The Matched Sample

Most of this article discusses our new matching technology, its implementation and properties, and the proof of its correctness. For motivation, however, it is best to begin by describing the situation before matching and the balance on covariates obtained in our matched sample. Our matching balanced 61 covariates, some of which were substantially out of balance before matching.

As shown in Table 1, treatment by a GO is more common in some SEER sites than in others, and is becoming somewhat more common over time. For instance, 26% of the GO patients were in the Detroit SEER site, whereas only 12% of the MO patients were in Detroit, and GOs were uncommon in San Francisco. (In Tables 1 and 2 the percents add up to 100% exactly before rounding for display.)

Year of diagnosis in Table 1 is an important variable in this study, because there were important improvements in chemotherapy during this time period. In part, because of these improvements, prognosis may have changed from 1991 to 1999. We also wanted to study the adoption of new chemotherapies by GOs and MOs; for that, controlling for year of diagnosis is essential. For these reasons, we matched exactly for year of diagnosis; a GO patient diagnosed in 1994, say, was always matched to an MO patient diagnosed in 1994.

Whether or not the SEER site is important is unclear. On the one hand, one strongly suspects that ovarian cancer is better characterized by stage and grade than by geography, and the health of patients is better characterized by comorbid conditions, such as diabetes or congestive heart failure. On the

Table 2. Covariate Balance Before and After Matching for Selected Covariates

Covariate	*GO*	*MO-matched*	*MO-all*
Stage 1	9	8	9
Stage 2	11	10	9
Stage 3	51	51	47
Stage 4	26	28	31
Stage missing	3	3	3
Total	100	100	100
Grade 1	5	5	4
Grade 2	16	15	17
Grade 3	52	52	47
Grade 4	9	8	11
Grade missing	18	19	21
Total	100	100	100
White	91	91	94
Black	8	7	3
Hypertension	48	46	42
Diabetes	11	10	8
Congestive heart failure	2	2	4
Weight loss in 90 days before diagnosis	1	1	3
Age (mean)	72.2	72.6	72.8
Propensity score (mean)	.23	.21	.14

NOTE: GO, patient of a gynecologic oncologist $n = 344$, MO-matched, matched patient of a medical oncologist, $n = 344$; MO-all; patients of medical oncologists before matching, $n = 2{,}011$. Values are percents, except as noted.

other hand, proximity to hospitals, quality of hospitals, styles of medical practice, and economic and social conditions vary with geography—Detroit, Seattle, and New Mexico differ in these respects—so geography possibly may act as a surrogate for relevant covariates that we do not have. Before matching, the relationships between provider type and site are often quite strong; for example, for Detroit and San Francisco, the GO versus MO odds ratio is $(26/1)/(12/9) = 19.5$, so a woman in the Medicare claims data is about 20 times more likely to be treated by a GO in Detroit than in San Francisco.

Using our new algorithm for fine balance, we exactly balanced the SEER sites, as shown in Table 1, but we did not insist that patients be individually matched for SEER site. That is, a GO patient from Detroit might be matched to an MO patient from Seattle, or vice versa, but the fraction of patients from each site was forced to agree exactly. Actually, the balance is finer than this; there is balance on SEER site within each year of diagnosis. For instance, among matched patients diagnosed in, say 1993, the number of GO patients in Detroit, say, equals the number of matched MO patients in Detroit.

The rest of the 61 variables were balanced approximately using the propensity score in conjunction with minimum distance matching using the Mahalanobis distance. Table 2 gives details about several clinically important covariates, and Figure 1 de-

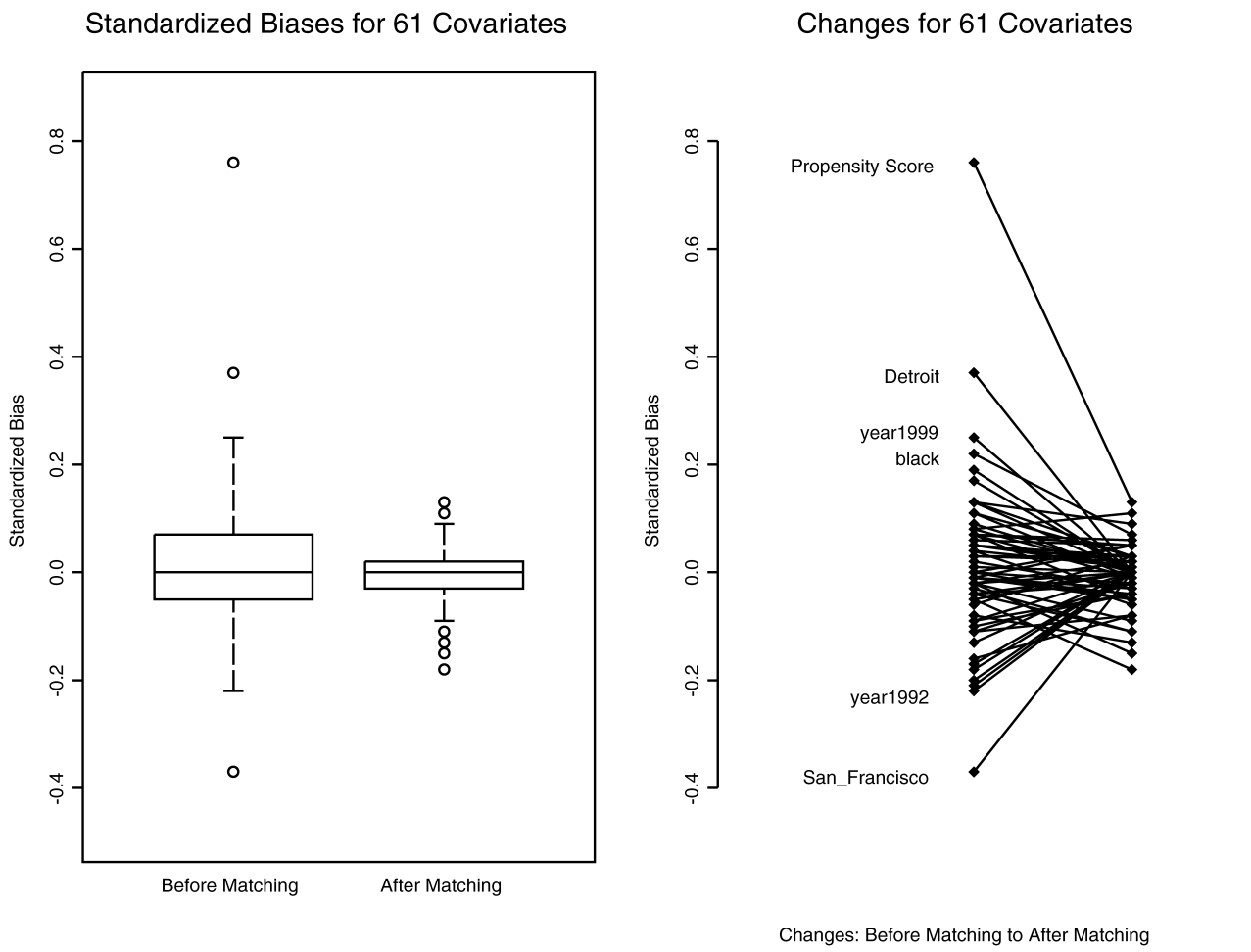

Figure 1. Plots of Imbalances in 61 Covariates, Before and After Matching. Values are differences in covariate means, GO versus MO, before and after matching, divided by the average within group standard deviation before matching. Six covariates with large initial biases are identified by name.

picts all 61 variables. (For this purpose, the five categories of stage in Table 2 are five related binary variables, and the 61st variable—propensity score—is a function of the others.)

In Table 2, consider the situation before matching, comparing the columns "GO" and "MO-all." Before matching, the patients of GOs were more likely than those of MOs to be stage 2 or 3 rather than stage 4, more likely to be grade 3, significantly more likely to be black, and had more hypertension and diabetes at diagnosis but less congestive heart failure. As it turns out, for unknown reasons, blacks have a poorer prognosis for ovarian cancer than whites. In contrast, the matched samples look reasonably similar. In particular, the substantial difference in the propensity score before matching is substantially reduced.

Aside from the covariates in Tables 1 and 2, most of the remaining 61 covariates describe comorbid conditions on diagnosis, including arrhythmia, renal dysfunction, liver dysfunction, asthma, and many others. Comorbid conditions may affect survival directly and may limit the patient's tolerance for chemotherapy, thereby possibly limiting its effectiveness; moreover, GOs and MOs may or may not have differing experience in providing chemotherapy to patients with varied comorbidities. Figure 1 depicts the improvement in covariate balance from matching for all 61 covariates, and Table 3 gives numerical summaries of the same information. The quantities displayed in Figure 1 and Table 3 are differences or absolute differences in means, GO versus MO, before and after matching, in units of the standard deviation *before* matching. More precisely, for a covariate x, write $\bar{x}_G$ and s_G for the mean and standard deviation of x among the 344 patients of GOs, write $\bar{x}_{Ma}$ and s_{Ma} for the mean and standard deviation of x among all 2,011 patients of MOs before matching, and write $\bar{x}_{Mm}$ for the mean of x for the 344 matched patients of MOs. Then the two standardized differences, before and after matching, in Figure 1 are $(\bar{x}_G - \bar{x}_{Ma})/\sqrt{1/2(s_G^2 + s_{Ma}^2)}$ and $(\bar{x}_G - \bar{x}_{Mm})/\sqrt{1/2(s_G^2 + s_{Ma}^2)}$; note that the denominators are the same. The absolute values of the standardized differences in Table 3 are much smaller after matching, with 75% of the covariates having a mean difference of <5% of a standard deviation after matching. Before matching, one of the covariates—propensity score—had a standardized difference of .76 standard deviation, but after matching it was .13. The absolute standardized differences before and after matching are not strongly related, with a slightly negative Kendall's rank correlation of $-.20$. Some other interesting graphical displays of covariate balance have been described by Love (2004).

Tables 1, 2, and 3 and Figure 1 indicate that the 61 covariates are much better balanced after matching. Improvement in balance is a relative standard, but an absolute standard is also needed, because the balance might be greatly improved but still unacceptable. How nearly balanced are the covariates after matching? How does the balance on the 61 covariates obtained by matching compare with the balance obtained by not matching and instead randomly assigning patients to GO or MO? (Obviously, random assignment would confer the important additional benefit of balancing other unobserved covariates, but matching cannot be expected to do this.)

Table 3. Quartiles and Extremes of the 61 Absolute Standardized Differences in Covariate Means, Before and After Matching

	Minimum	*25%*	*Median*	*75%*	*Maximum*
Before	0	.03	.07	.13	.76
After	0	0	.03	.05	.18

Figure 2 and Table 4 compare the balance on the 61 covariates with the balance expected in a completely randomized experiment in which 344 of $688 = 2 \times 344$ patients were randomly picked and assigned to a GO. For each of the 61 covariates, the p value for a two-sample test was computed, comparing GO to MO for the 688 patients; specifically, the chi-squared test was used for binary covariates and Wilcoxon's rank sum test was used for continuous covariates. Had the 688 patients been randomly assigned to GO or MO, each p value would have had the uniform distribution (ignoring the slight effects of discreteness). (Because the covariates are correlated with one another, the 61 p values are not expected to be independent.) Figure 2 and Table 4 contrast the 61 observed p values with the uniform distribution. The empirical distribution of p values is consistently above the diagonal $y = x$ line, so the empirical distribution is stochastically larger than the uniform distribution. There was only one p value $< .05$ (for "deficiency anemia before diagnosis"), whereas with random assignment, $.05 \times 61 \doteq 3$ were expected, and about 75% of the p values were $.496 \doteq \frac{1}{2}$ or larger, and only 50% would be expected to be that large under complete randomization. In short, the balance on the 61 covariates obtained by matching was much greater than the balance expected from random assignment of 344 of the 688 patients to a GO.

3. NOTATION, REVIEW, AND DEFINITION OF FINE BALANCE

3.1 Review: Assignment Algorithms; Implementations

Given an $I \times J$ matrix $\mathbf{\Delta}$ of nonnegative distances, $\Delta_{ij} \geq 0$, the *assignment algorithm* picks for each row i, $i = 1, \ldots, I$, a different column, a_i, the column assigned to row i, to minimize the total distance among all possible assignments; that is, the assignment minimizes $\sum_{i=1}^{I} \Delta_{i,a_i}$ subject to $1 \leq a_i \leq J$, $i = 1, \ldots, I$, and $a_i \neq a_k$ for all $i \neq k$. The problem is not trivial, because there can be—and usually is—competition between different rows for the same column, and an unwise initial choice can result in a very poor pairing of rows and columns. For instance, with

$$\mathbf{\Delta} = \begin{bmatrix} 0 & \varepsilon \\ \varepsilon & \infty \end{bmatrix}, \qquad \varepsilon > 0,$$

a greedy algorithm or best-first algorithm would pair row 1 with column 1 with a distance of $\Delta_{11} = 0$ and then would be left with no alternative to pairing row 2 with column 2 with a distance of $\Delta_{22} = \infty$ for a total distance of $\Delta_{11} + \Delta_{22} = 0 + \infty = \infty$, whereas an optimal assignment has a total distance of $\Delta_{12} + \Delta_{21} = 2\varepsilon$.

The problem of finding an optimal assignment is one of the oldest combinatorial optimization problems, and several fast algorithms exist. For statisticians, one convenient implementation is Proc Assign in SAS/OR, which accepts $\mathbf{\Delta}$ as input and returns an optimal assignment. Kuhn (1955) proposed a solution using the so-called "Hungarian method," which can solve the assignment problem in $O(J^3)$ arithmetic operations (see Pa-

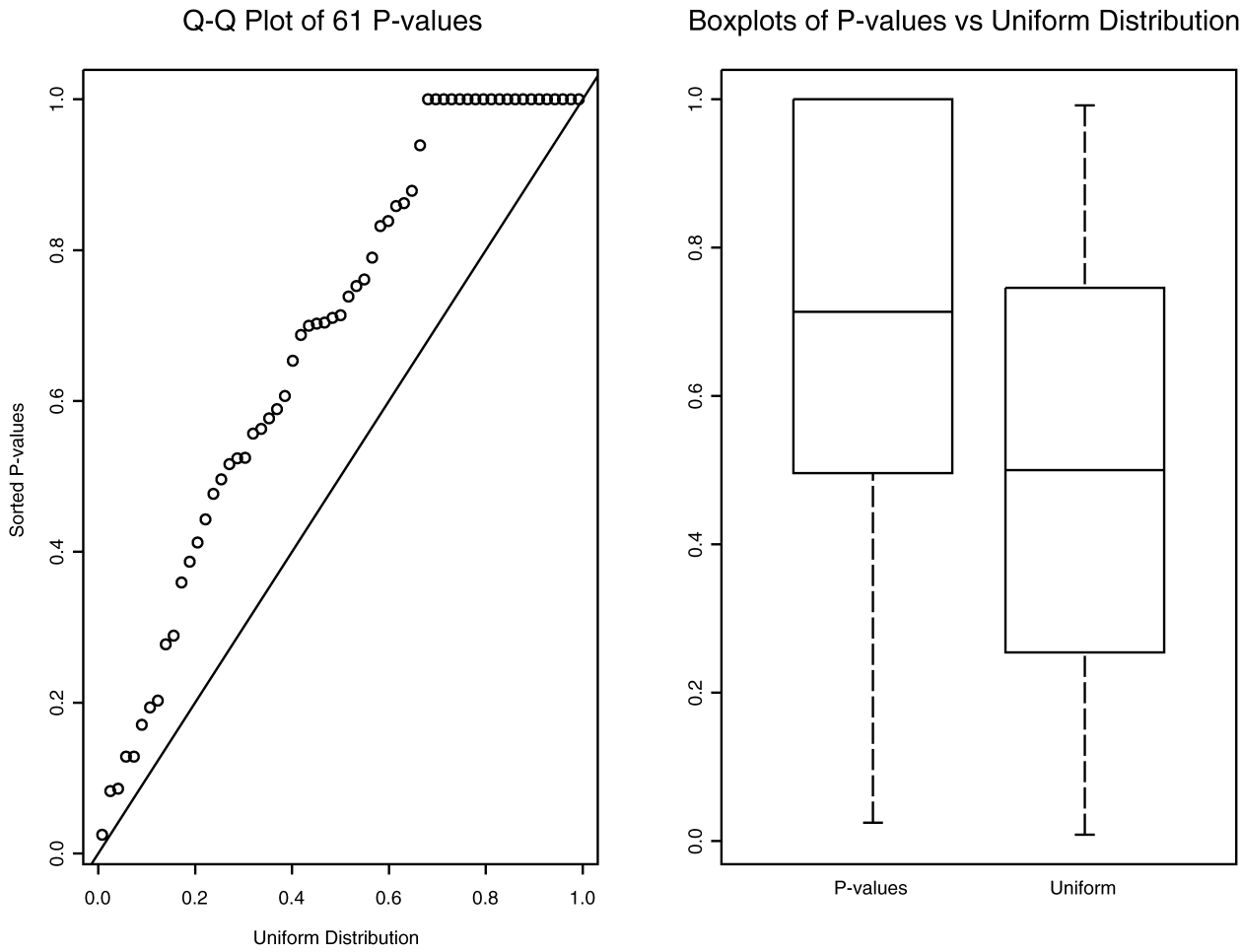

Figure 2. Comparing the Balance on 61 Covariates With the Balance Expected in a Completely Randomized Experiment: 61 p Values From Two-Sample Tests Compared With the Uniform Distribution. The diagonal line is y = x.

padimitriou and Steiglitz 1982, sec. 11.2). For comparison, if one multiplies two $J \times J$ matrices in the usual way, then it takes $O(J^3)$ arithmetic operations. (Fortran code for the Hungarian method is available from Carpaneto and Toth 1980.) An alternative algorithm that is more intuitive and may be faster is the auction algorithm (and Fortran code) of Bertsekas (1981); it literally holds an auction, selling the columns to the rows. Hansen's (2004) optmatch software in R provides convenient access to Bertsekas's very fast Fortran code and can be used to solve the assignment problem inside R. The auction algorithm may be implemented using parallel processing (see Bertsekas and Tsitsiklis 1989, sec. 5.3.1). Dell'Amico and Toth (2000) compared eight implementations of the assignment algorithm.

3.2 Notation and Review: Optimal Matching With a Fixed Number of Controls

There are T treated subjects, $\mathcal{T} = \{\tau_1, \ldots, \tau_T\}$, and $C > T$ potential controls, $\mathcal{C} = \{\gamma_1, \ldots, \gamma_C\}$, with $\mathcal{T} \cap \mathcal{C} = \emptyset$ and a nonnegative, possibly infinite distance $\delta_{tc} \geq 0$ between τ_t and γ_c, for each $\tau_t \in \mathcal{T}$ and $\gamma_c \in \mathcal{C}$. In Section 2 $T = 344$, $\mathcal{T}$ contains the 344 patients of a GO, $C = 2{,}011$, and $\mathcal{C}$ contains the 2,011 patients of an MO. Write $|\mathcal{S}|$ for the number of elements of a finite set $\mathcal{S}$, so $|\mathcal{T}| = T$ and $|\mathcal{C}| = C$. The direct product $\mathcal{S}_1 \times \mathcal{S}_2$ of two sets, $\mathcal{S}_1$ and $\mathcal{S}_2$, is the set of ordered pairs, (s_1, s_2) with $s_1 \in \mathcal{S}_1$ and $s_2 \in \mathcal{S}_2$.

Table 4. p Values From Two-Sample Tests for 61 Covariates Compared With the Uniform Distribution

	Minimum	*Quartile*	*Median*	*Quartile*	*Maximum*
Observed *p* values	.025	.496	.714	1.000	1.000
Uniform distribution		.250	.500	.750	

Using the observed covariates, a nonnegative distance $\delta_{tc} \geq 0$ is defined between the covariates describing τ_t and γ_c. One popular distance is the Mahalanobis distance (Rubin 1980). The distance between τ_t and γ_c may be infinite, $\delta_{tc} = \infty$, if one wishes to forbid matching of τ_t and γ_c. Moreover, the distance may incorporate a sequence of progressively larger but finite penalties to constrain the matching while giving greater priority to certain constraints. See Section 5 for the definition of δ_{tc} used in constructing the match in Section 2.

If $C \geq \kappa T$, where $\kappa \geq 1$ is an integer, then a matched sample with κ controls matches each $\tau_t \in \mathcal{T}$ to κ controls $\gamma_c \in \mathcal{C}$, so that each potential control is used at most once. In Section 2 $\kappa = 1$ and $\kappa T = 344$ matched pairs were formed. Formally, a matched sample $\mathbf{\Lambda}$ with κ controls is a subset of the direct product, $\mathbf{\Lambda} \subset \mathcal{T} \times \mathcal{C}$, such that each $\tau_t \in \mathcal{T}$ appears in κ pairs, $(\tau_t, \gamma_c) \in \mathbf{\Lambda}$, and each potential control $\gamma_c \in \mathcal{C}$ appears in at most one pair, $(\tau_t, \gamma_c) \in \mathbf{\Lambda}$. A matching $\mathbf{\Lambda}$ with κ controls is optimal if, among all matchings with κ controls, $\mathbf{\Lambda}$ minimizes the total distance

between treated subjects and their matched controls,

$$\delta(\mathbf{\Lambda}) = \sum_{(\tau_t,\gamma_c)\in\mathbf{\Lambda}} \delta_{tc}. \tag{1}$$

Write $\mathcal{M} \subseteq \mathcal{C}$ for the subset of controls selected by $\mathbf{\Lambda}$, that is, $\mathcal{M} = \{\gamma_c \in \mathcal{C} : (\tau_t, \gamma_c) \in \mathbf{\Lambda}\}$. In Section 2 $\mathbf{\Lambda}$ refers to the 344 pairs of a GO patient and an MO patient, whereas $\mathcal{M}$ refers to the 344 MO patients included in the matched sample. Issues and trade-offs arising with the use of multiple controls, $\kappa > 1$, have been discussed by Smith (1997) and Ming and Rosenbaum (2000).

If the assignment algorithm is applied to a matrix $\mathbf{\Delta}$ with $I = T$, $J = C$, and $\Delta_{ij} = \delta_{ij}$, then it will return an optimal pair matching $\mathbf{\Lambda}$ in which each $\tau_t \in \mathcal{T}$ is matched to a different $\gamma_c \in \mathcal{C}$, so the total distance within matched pairs is minimized. If $C \geq 2T$ and the assignment algorithm is applied to a matrix $\mathbf{\Delta}$ with $I = 2T$, $J = C$, and $\Delta_{(2i-1),j} = \delta_{ij}$ and $\Delta_{(2i),j} = \delta_{ij}$, $i = 1, \ldots, 2T$, $j = 1, \ldots, C$, then it will return T matched triples in which each $\tau_t \in \mathcal{T}$ is matched to two different $\gamma_c \in \mathcal{C}$ to minimize the total distance between treated subjects and matched controls. Similarly, if $C \geq \kappa T$, then the assignment algorithm applied to a matrix $\mathbf{\Delta}$ with $I = \kappa T$ and $J = C$ and with the row $(\delta_{i1}, \ldots, \delta_{iC})$ repeated κ times yields an optimal match with κ controls.

3.3 Fine Balance: Exact Balance Without Exact Matching

There is a discrete, nominal variable with $B \geq 2$ levels, $b = 1, \ldots, B$, with $n_b \geq 0$ treated subjects at level b, where $T = \sum n_b$. In Section 2 the nominal variable describes the $B = 8 \times 9 = 72$ combinations of the 8 SEER sites and the 9 years. Let $\mathcal{B}_b \subset \mathcal{C}$ be the subset of controls with level b on the nominal variable, $b = 1, \ldots, B$, where $\mathcal{C} = \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_B$. A match $\mathbf{\Lambda}$ with κ controls is *finely balanced* if there are κn_b controls with level b of the nominal variable, that is, if $|\mathcal{M} \cap \mathcal{B}_b| = \kappa n_b$ for $b = 1, \ldots, B$. The match in Section 2 is finely balanced in this sense, as partially described in Table 1. Obviously, fine balance is possible if and only if $|\mathcal{B}_b| \geq \kappa n_b$ for $b = 1, \ldots, B$. If $\mathbf{\Lambda}$ minimizes $\delta(\mathbf{\Lambda})$ among all finely balanced matches with κ controls, then $\mathbf{\Lambda}$ is optimal.

The term "finely balanced" is intended to suggest that the nominal variable, often with many levels, has been balanced exactly at every level, that is, with fine attention to detail. We avoided the term "exactly balanced" because, although it is descriptive, the term "exact matching" is in widespread use for a different concept, and we wanted to avoid any possibility of confusion.

3.4 Fine Balance: Pros and Cons

Fine balance is a matching tool typically used in conjunction with other matching tools, such as propensity scores and minimum distance matching. Consequently, the choice is not which one tool to use, but rather whether or not to apply a particular tool in conjunction with other tools. In this section we discuss circumstances in which fine balance is a useful addition, as well as other circumstances in which it may have little or nothing to add.

With a single or one-dimensional covariate, say age, it is often practical to find close individual matches for age, so each treated subject is matched to a control whose age differs only by a few years. However, it is virtually impossible to obtain very close individual matches on every coordinate of a high-dimensional covariate. In Section 2.2 the covariate was of dimension 61, so if one divided each coordinate into two categories, above and below the univariate median, then one would have divided the 61-dimensional covariate space into $2^{61} = 2.3 \times 10^{18}$ quadrants; so even if there were a million controls to choose from, there would be about a trillion quadrants per control. In this case, it is very unlikely that one could match so that matched treated–control pairs fell on the same side of the median for all 61 covariates. For this reason, when seeking close individual matches, attention must focus on a few important covariates. In ovarian cancer, clinical stage is one such important covariate, and, as discussed in Section 5, it was given more attention than most of the 61 covariates in defining the distance, δ_{tc}.

In contrast, although it is impractical to match exactly on every coordinate of a high-dimensional covariate, it is practical to balance a high-dimensional covariate. Balance refers to the distribution of the covariate in treated and control groups after matching, rather than to close matches in each and every pair. For instance, there is balance on diabetes if the proportion of diabetics is about the same in treated and control groups after matching, even if diabetics are not always matched to other diabetics; see Table 2. Covariate balance may be obtained either stochastically or deterministically.

In experiments, random assignment of subjects to treatment or control tends to stochastically balance both observed and unobserved covariates. Let $\mathcal{X}$ be any one fixed (measurable) subset of the covariate space; thus in Section 2.2, $\mathcal{X}$ is a subset of a 61-dimensional space. What proportions of the time do treated subjects and controls have covariate values in $\mathcal{X}$? In a completely randomized experiment, one would apply the weak law of large numbers to the binary variable indicating whether the covariate fell in $\mathcal{X}$ to see that these two proportions converge in probability to the same value as the sample size increases. Randomization does not ensure covariate balance in very small experiments, but reasonable balance typically is attained in experiments of moderate size. In observational studies, matching on an estimate of the propensity score—the conditional probability of treatment given observed covariates—tends to stochastically balance observed covariates used to construct the score (see Rosenbaum and Rubin 1983).

Two tools have just been described: stochastic balance on many covariates using the propensity score and close matching on a few of the most important covariates using a distance. Used together, these two tools work well in many situations. An important situation in which they do not work adequately concerns a covariate with many nominal levels. In Table 1 such a covariate consists of the $8 \times 9 = 72$ combinations of a SEER site and a year of diagnosis. On average, there are only $344/72 = 4.8$ GO patients per category of this covariate, and one cannot appeal to the law of large numbers to stochastically balance all 72 levels of such a covariate. For such a covariate, it might be more realistic to imagine that as the sample size increases, the number of categories of the covariate increases in direct proportion, so the number of people per category tends to a constant. On the other hand, if one required an exact match on this nominal

variable with its many levels, this would drastically restrict the possible matches, and consequently make it much more difficult to match on other variables. In contrast, fine balance forces exact balance at all levels of the nominal variable but places no restriction on individual matched pairs—any one treated subject can be matched to any one control. Similar considerations apply when the nominal variable takes on only a few values, but of those few values, some are exceedingly rare.

Fine balancing may also be used to enforce balance on a variable that happens to be difficult to balance by other methods. Because the propensity score is constructed to best distinguish the treated and control groups, it may be the most difficult variable to balance in some problems. This is usually solved by caliper matching on the propensity score together with distance matching inside the caliper (Rosenbaum and Rubin 1985), but optimal fine balancing for several categories of the propensity score is a practical alternative. Unlike caliper matching, fine balancing will not insist that subjects with similar propensity scores be matched to each other, but simply that the distribution of the propensity score be balanced in the matched sample.

The principal disadvantage of fine balancing is that it is a constraint on an optimization problem, namely the minimization of the total distance within matched sets, so one can obtain a better or lower minimum total distance by removing the constraint. In practice, providing that one makes no use of information about outcomes, one can construct several matched samples by different methods and select for use the sample that produces the most satisfactory balance on covariates. In a simple case, this was illustrated by Rosenbaum and Rubin (1985), who constructed three matched samples, one of which is clearly best.

4. OBTAINING OPTIMAL MATCHES WITH FINE BALANCE USING THE ASSIGNMENT ALGORITHM

4.1 A Small Illustration

We describe the algorithm in the context of a small illustration. In this illustration, the treated group contains $T = 5$ individuals, $\mathcal{T} = \{\tau_1, \tau_2, \tau_3, \tau_4, \tau_5\}$, divided into $B = 3$ types, where τ_1 is of type $b = 1$, τ_2 and τ_3 are of type $b = 2$, and τ_4 and τ_5 are of type $b = 3$, so $n_1 = 1$, $n_2 = 2$, and $n_3 = 2$. There are $C = 14$ potential controls, $\mathcal{C} = \{\gamma_1, \ldots, \gamma_{14}\}$, where $\mathcal{B}_1 = \{\gamma_1, \gamma_2, \gamma_3\}$ are of type $b = 1$, $\mathcal{B}_2 = \{\gamma_4, \gamma_5, \gamma_6, \gamma_7\}$ are of type $b = 2$, and $\mathcal{B}_3 = \{\gamma_8, \ldots, \gamma_{14}\}$ are of type $b = 3$. There are distances δ_{tc} between treated subject τ_t and control subject γ_c, for $t = 1, \ldots, 5$ and $c = 1, \ldots, 14$. The matching will have $\kappa = 2$ controls for each treated subject. Fine balance is possible because $|\mathcal{B}_1| = 3 \geq \kappa n_1 = 2 \times 1 = 2$, $|\mathcal{B}_2| = 4 \geq \kappa n_2 = 2 \times 2 = 4$, and $|\mathcal{B}_3| = 7 \geq \kappa n_1 = 2 \times 2 = 4$.

The procedure is to construct a certain distance matrix, $\mathbf{\Delta}$, and apply the assignment algorithm to it. Here $\mathbf{\Delta}$ is 14×14 with entries Δ_{ij} which will now be defined. For the illustration, the matrix $\mathbf{\Delta}$ is given in Table 5. In the first 10 rows of $\mathbf{\Delta}$ in Table 5, the $T = 5$ treated subjects, $\tau_1, \ldots, \tau_5$ are repeated $\kappa = 2$ times with the corresponding δ_{tc} also duplicated.

Balance is obtained by adding rows, labeled α_{bk}, to $\mathbf{\Delta}$. These added rows act as placeholders; unneeded controls are matched to α_{bk}'s and are later discarded. To obtain fine balance for $b = 1$, two of the three γ's in $\mathcal{B}_1 = \{\gamma_1, \gamma_2, \gamma_3\}$ must be matched to some $\tau_1, \ldots, \tau_5$, and one γ in $\mathcal{B}_1$ must be discarded. This task is accomplished with the aid of α_{11} in row 11 of Table 5, which has $\Delta_{11,1} = 0$ for γ_1, $\Delta_{11,2} = 0$ for γ_2, $\Delta_{11,3} = 0$ for γ_3, and $\Delta_{11,j} = \infty$ for $j = 4, \ldots, 14$. Recall that the assignment algorithm pairs each row to a column, using every row and column exactly once, in such a way as to minimize the total distance within the matched pairs. To avoid the ∞'s in row $i = 11$, the assignment algorithm will pair α_{11} to γ_1 or γ_2 or γ_3 and will make this choice so that it best contributes to minimizing the total distance.

To obtain fine balance for $b = 2$, all of the controls in $\mathcal{B}_2 = \{\gamma_4, \gamma_5, \gamma_6, \gamma_7\}$ must be used. As a result, there is no row α_{2k} in $\mathbf{\Delta}$ in Table 5, because there is no need to remove a γ from $\mathcal{B}_2$.

To obtain fine balance for $b = 3$, four of the seven potential controls in $\mathcal{B}_3 = \{\gamma_8, \ldots, \gamma_{14}\}$ must be retained. Therefore, $\mathbf{\Delta}$ in Table 5 has three rows α_{3k}, $k = 1, 2, 3$, to remove three of the γ's in $\mathcal{B}_3$. The assignment algorithm will pair these rows with a column to minimize the total distance, and to avoid the ∞'s, it will pair these rows to elements of $\mathcal{B}_3$. The selection of three elements of $\mathcal{B}_3$ to pair with α's will reflect the distances δ_{tc} and the overall task of minimizing the total distance.

4.2 General Procedure

The general procedure is as follows:

Step 1. Check that $|\mathcal{B}_b| \geq \kappa n_b$ for $b = 1, \ldots, B$. If this condition fails, then fine balance is not attainable with this value of κ.

Step 2. Create κ rows in $\mathbf{\Delta}$ for each τ_t, $t = 1, \ldots, T$, duplicating the δ_{tc}.

Table 5. Distance Matrix $\mathbf{\Delta}$ for Fine Balance With $\kappa = 2$ Controls

	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	γ_{10}	γ_{11}	γ_{12}	γ_{13}	γ_{14}
τ_1	δ_{11}	δ_{12}	δ_{13}	δ_{14}	δ_{15}	δ_{16}	δ_{17}	δ_{18}	δ_{19}	$\delta_{1,10}$	$\delta_{1,11}$	$\delta_{1,12}$	$\delta_{1,13}$	$\delta_{1,14}$
τ_1	δ_{11}	δ_{12}	δ_{13}	δ_{14}	δ_{15}	δ_{16}	δ_{17}	δ_{18}	δ_{19}	$\delta_{1,10}$	$\delta_{1,11}$	$\delta_{1,12}$	$\delta_{1,13}$	$\delta_{1,14}$
τ_2	δ_{21}	δ_{22}	δ_{23}	δ_{24}	δ_{25}	δ_{26}	δ_{27}	δ_{28}	δ_{29}	$\delta_{2,10}$	$\delta_{2,11}$	$\delta_{2,12}$	$\delta_{2,13}$	$\delta_{2,14}$
τ_2	δ_{21}	δ_{22}	δ_{23}	δ_{24}	δ_{25}	δ_{26}	δ_{27}	δ_{28}	δ_{29}	$\delta_{2,10}$	$\delta_{2,11}$	$\delta_{2,12}$	$\delta_{2,13}$	$\delta_{2,14}$
⋮			⋮				⋮				⋮			⋮
τ_5	δ_{51}	δ_{52}	δ_{53}	δ_{54}	δ_{55}	δ_{56}	δ_{57}	δ_{58}	δ_{59}	$\delta_{5,10}$	$\delta_{5,11}$	$\delta_{5,12}$	$\delta_{5,13}$	$\delta_{5,14}$
τ_5	δ_{51}	δ_{52}	δ_{53}	δ_{54}	δ_{55}	δ_{56}	δ_{57}	δ_{58}	δ_{59}	$\delta_{5,10}$	$\delta_{5,11}$	$\delta_{5,12}$	$\delta_{5,13}$	$\delta_{5,14}$
α_{11}	0	0	0	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
α_{31}	∞	∞	∞	∞	∞	∞	∞	0	0	0	0	0	0	0
α_{32}	∞	∞	∞	∞	∞	∞	∞	0	0	0	0	0	0	0
α_{33}	∞	∞	∞	∞	∞	∞	∞	0	0	0	0	0	0	0

Step 3. For $b = 1, \ldots, B$, create one row, α_{bk}, in $\mathbf{\Delta}$ for $k = 1, \ldots, |\mathcal{B}_b| - \kappa n_b$, noting carefully that $|\mathcal{B}_b| - \kappa n_b$ may equal 0 for some b, in which case zero rows are added. The distances in $\mathbf{\Delta}$ for α_{bk} are 0 for $\gamma_c \in \mathcal{B}_b$ and ∞ if $\gamma_c \notin \mathcal{B}_b$. Now $\mathbf{\Delta}$ is a $C \times C$ matrix, so $I = J = C$.

Step 4. Apply the assignment algorithm to $\mathbf{\Delta}$, obtaining an optimal assignment, say $\widetilde{\mathbf{a}} = (\widetilde{a}_1, \ldots, \widetilde{a}_C)$, of the C rows of $\mathbf{\Delta}$ to the C columns of $\mathbf{\Delta}$. Let $\mathbf{\Lambda}$ be the resulting set of (τ_t, γ_c) pairs that remain after discarding all pairs that involve an α.

Recall that $\delta_{tc} \geq 0$ and δ_{tc} may be ∞ for some t and c, where $\delta_{tc} = \infty$ means that a match of t to c is forbidden. Although Step 1 checks to make sure a finely balanced match exists, the pattern of infinite δ_{tc}'s may be such that one cannot have both fine balance and finite distance. Proposition 1 shows that the output of Step 4 indicates what is possible.

Proposition 1. Let $\mathbf{\Lambda}$ be the match obtained by an optimal assignment $\widetilde{\mathbf{a}}$ in Step 4. If this optimal assignment has finite distance $\sum_{i=1}^{C} \Delta_{i,\widetilde{a}_i} < \infty$, then $\mathbf{\Lambda}$ is an optimal, finely balanced match with κ controls. If this optimal assignment has infinite distance $\sum_{i=1}^{C} \Delta_{i,\widetilde{a}_i} = \infty$, then there is no finely balanced match $\mathbf{\Lambda}$ with κ controls with finite distance $\delta(\mathbf{\Lambda}) < \infty$.

Proof. Recall that $\delta_{tc} \geq 0$ and δ_{tc} may be ∞ for some t and c. For the distance matrix $\mathbf{\Delta}$, let $\mathbf{a} = (a_1, \ldots, a_C)$ be any assignment of columns to rows, as defined in Section 3.1; there are $C!$ possible assignments $\mathbf{a}$. For assignment $\mathbf{a}$, let $\mathbf{\Upsilon}_\mathbf{a}$ be the set containing the κT pairs (τ_t, γ_c) that remain after discarding all pairs that involve an α_{bk}, that is, the pairs for the first κT rows of $\mathbf{\Delta}$ defined by $(a_1, \ldots, a_{\kappa T})$. In particular, $\mathbf{\Lambda} = \mathbf{\Upsilon}_{\widetilde{\mathbf{a}}}$. By definition (1), for every assignment $\mathbf{a}$,

$$\delta(\mathbf{\Upsilon}_\mathbf{a}) = \sum_{i=1}^{\kappa T} \Delta_{i,a_i} \leq \sum_{i=1}^{C} \Delta_{i,a_i} \tag{2}$$

with equality if and only if

$$\Delta_{i,a_i} = 0 \quad \text{for } i = \kappa T + 1, \ldots, C. \tag{3}$$

Now (3) holds for an assignment $\mathbf{a}$ if and only if for $b = 1, \ldots, B$, the rows α_{bk}, $k = 1, \ldots, |\mathcal{B}_b| - \kappa n_b$, are assigned to columns with $\gamma_c \in \mathcal{B}_b$, in which case there are κn_b controls $\gamma_c \in \mathcal{B}_b$ assigned to rows for treated subjects, τ_t; that is, (3) holds if and only if $\mathbf{\Upsilon}_\mathbf{a}$ is a finely balanced match with κ controls. Moreover, because rows $i = \kappa T + 1, \ldots, C$ of $\mathbf{\Delta}$ have either $\Delta_{ij} = 0$ or $\Delta_{ij} = \infty$, if (3) does not hold, then $\infty = \sum_{i=1}^{C} \Delta_{i,a_i}$. It follows that if $\infty = \sum_{i=1}^{C} \Delta_{i,a_i}$, then either $\mathbf{\Upsilon}_\mathbf{a}$ is not finely balanced or $\delta(\mathbf{\Upsilon}_\mathbf{a}) = \infty$. Hence, if $\sum_{i=1}^{C} \Delta_{i,\widetilde{a}_i} = \infty$ for an optimal assignment $\widetilde{\mathbf{a}}$, then there is no finely balanced matching with κ controls and finite distance. This proves the second assertion in the proposition. So, conversely, assume that $\sum_{i=1}^{C} \Delta_{i,\widetilde{a}_i} < \infty$, which implies that (3) holds and $\mathbf{\Lambda} = \mathbf{\Upsilon}_{\widetilde{\mathbf{a}}}$ is finely balanced with κ controls. Because $\mathbf{\Lambda}$ was obtained by an optimal assignment, $\delta(\mathbf{\Lambda}) = \delta(\mathbf{\Upsilon}_{\widetilde{\mathbf{a}}}) \leq \delta(\mathbf{\Upsilon}_\mathbf{a})$ for all possible assignments $\mathbf{a}$ that satisfy (3), that is, for all finely balanced matchings with κ controls.

5. IMPLEMENTATION IN THE EXAMPLE

The matched sample in Section 2.2 was constructed as follows. We estimated the propensity score—that is, the conditional probability of treatment by a GO rather than an MO given the 61 observed covariates—using a logit model that included most of the covariates but excluded a few very rare binary covariates. Because we wanted exact matches for year of diagnosis, the matching for each of the 9 years in Table 1 was done separately, so there were 9 separate calls to PROC ASSIGN. Within each year, fine balance was required for the 8 SEER sites in Table 1. This implies, for instance, that GO patients diagnosed in 1993 were matched to MO patients diagnosed in 1993, and the number of GO patients from Detroit in 1993 exactly equals the number of MO patients from Detroit in 1993, but GO patients from Detroit in 1993 were not typically matched to MO patients from Detroit in 1993.

The distance δ_{tc} used in matching was the sum of two "penalties" and a Mahalanobis distance. A penalty is a large number—in this case, 300—added to the distance if some condition (or "constraint") did not hold, so the optimization algorithm worked very hard to avoid violating a constraint. Because we wanted close matches on the propensity score, there was a penalty of 300 added to the distance δ_{tc} if GO patient t and MO patient c differed in the propensity score by more than .2 times the standard deviation of the propensity score. This is similar to using a caliper on the propensity score, as suggested by Rosenbaum and Rubin (1985), but it permits small violations of a few of the constraints if there is no solution that satisfies all of the constraints. Clinical stage is a key variable for treatment decisions, and it was missing for 3% of GO patients and 3% of MO patients. We did not want to discard the patients for whom stage was missing, but it would be hard to compare the treatment given to a Stage 3 patient and a patient with a missing stage. For this reason, we wanted to match GO patients with a missing stage to MO patients with a missing stage. We added a penalty of 300 to δ_{tc} if stage was missing for t or for c but not for both. The Mahalanobis distance included the following covariates, some of which were categorical and were represented by one or more coded variables: propensity score; individual clinical stages; individual clinical grades, including grade missing; black or other; substantial weight loss in the 90 days before diagnosis (yes/no); symptoms of congestive heart failure in the 90 days before diagnosis (yes/no); and diabetes noted within 90 days before diagnosis (yes/no). The ∞'s in Table 5 were coded as 1,000 and so were substantially larger than the penalties. The Mahalanobis distance used 12 linearly independent covariates.

Table 6 summarizes the distances δ_{tc} within the 344 matched pairs (t, c). As a basis for comparison, if one drew two independent observations from a 12-dimensional multivariate normal distribution, the expected Mahalanobis distance between them would be $2 \times 12 = 24$. As shown in Table 6, most of the 344 Mahalanobis distances for matched patients were quite small with an upper quartile of 1.61, and in all but two pairs, the constraints imposed by the penalties were respected.

Table 6. Distances δ_{tc} for 344 Matched Pairs

Median			.13		
Quartiles		.01		1.61	
Extreme 10%		0		11.16	
	5	*4*	*3*	*2*	*1*
Smallest 5	0	0	0	0	0
Largest 5	41.50	55.52	60.00	322.78	336.49

6. SUMMARY

By creating a suitably patterned distance matrix, a minimum distance, finely balanced matched sample may be constructed using the fast, widely implemented assignment algorithm. In the example, we exactly balanced a nominal variable with 72 categories, stochastically balanced 61 covariates using the propensity score, and obtained individually close matches for several key variables, including clinical stage and grade, all with a few calls to Proc Assign in SAS.

[Received September 2005. Revised May 2006.]

REFERENCES

Bergstralh, E. J., Kosanke, J. L., and Jacobsen, S. L. (1996), "Software for Optimal Matching in Observational Studies," *Epidemiology*, 7, 331–332; SAS code available at *http://www.mayo.edu/hsr/sasmac.html*.

Bertsekas, D. P. (1981), "A New Algorithm for the Assignment Problem," *Mathematical Programming*, 21, 152–171; Fortran code available at *http://www.mit.edu:8001/people/dimitrib/home.html*.

Bertsekas, D. P., and Tsitsiklis, J. N. (1989), *Parallel and Distributed Computation: Numerical Methods*, Englewood Cliffs, NJ: Prentice-Hall.

Carpaneto, G., and Toth, P. (1980), "Algorithm 548: Solution of the Assignment Problem [H]," *ACM Transactions on Mathematical Software*, 6, 104–111; Fortran code available at *http://www.netlib.org/toms/548* and *http://portal.acm.org*.

Chu, C. S., and Rubin, S. C. (2001), "Second-Look Laparotomy for Epithelial Ovarian Cancer: A Reappraisal," *Current Oncology Reports*, 3, 11–18.

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.

Dreger, P., Stilgenbauer, S., Benner, A., Ritgen, M., Krober, A., Kneba, M., Schmitz, N., and Dohner, H. (2004), "The Prognostic Impact of Autologous Stem Cell Transplantation in Patients With Chronic Lymphocytic Leukemia: A Risk-Matched Analysis Based on the V–H Gene Mutational Status," *Blood*, 103, 2850–2858.

Hansen, B. B. (2004), "Full Matching in an Observational Study of Coaching for the SAT," *Journal of the American Statistical Association*, 99, 609–618; R-code available at *http://www.stat.lsa.umich.edu/~bbh/*.

Harlan, L. C., Clegg, L. X., and Trimble, E. L. (2003), "Trends in Surgery and Chemotherapy for Women Diagnosed With Ovarian Cancer in the United States," *Journal of Clinical Oncology*, 21, 3488–3494.

Heintz, A. P. M. (2002), "Gynecologic Oncology: A Discipline or a Profession?" *International Journal of Gynecologic Cancer*, 12, 133–134.

Junor, E. J., Hole, D. J., McNulty, L., Mason, M., and Young, J. (1999), "Specialist Gynaecologists and Survival Outcome in Ovarian Cancer: A Scottish National Study of 1866 Patients," *British Journal of Obstetrics and Gynaecology*, 106, 1130–1136.

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

Love, T. (2004), "Graphical Display of Covariate Balance," *http://chrp.org/love/JSM2004RoundTableHandout.pdf*.

McGuire, W. P., Hoskins, W. J., Brady, M. F., Kucera, P. R., Partridge, E. E., Look, K. Y., Clarke-Pearson, D. L., and Davidson, M. (1996), "Cyclophosphamide and Cisplatin Compared With Paclitaxel and Cisplatin in Patients With Stage III and Stage IV Ovarian Cancer," *New England Journal of Medicine*, 334, 1–6.

Ming, K., and Rosenbaum, P. R. (2000), "Substantial Gains in Bias Reduction From Matching With a Variable Number of Controls," *Biometrics*, 56, 118–124.

——— (2001), "A Note on Optimal Matching With Variable Controls Using the Assignment Algorithm," *Journal of Computational and Graphical Statistics*, 10, 455–463.

Neijt, J. P. (1996), "New Therapy for Ovarian Cancer" (Editorial), *New England Journal of Medicine*, 334, 50–51.

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

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

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

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

Rubin, D. B. (1980), "Bias Reduction Using Mahalanobis Metric Matching," *Biometrics*, 36, 293–298.

Silber, J. H., Rosenbaum, P. R., Polsky, D., Ross, R. N., Even-Shoshan, O., Schwartz, J. S., Armstrong, K. A., and Randall, T. C. (2007), "Does Ovarian Cancer Treatment and Survival Differ by the Specialty Providing Chemotherapy?" *Journal of Clinical Oncology*, to appear.

Silber, J. H., Rosenbaum, P. R., Trudeau, M. E., Even-Shoshan, O., Chen, W., Zhang, X., and Mosher, R. E. (2001), "Multivariate Matching and Bias Reduction in the Surgical Outcomes Study," *Medical Care*, 39, 1048–1064.

Silber, J. H., Rosenbaum, P. R., Trudeau, M. E., Chen, W., Zhang, X., Lorch, S., Rapaport-Kelz, R., Mosher, R. E., and Even-Shoshan, O. (2005), "Preoperative Antibiotics and Mortality in the Elderly," *The Annals of Surgery*, 242, 107–114.

Singh, M., Ting, H. H., Gersh, B. J., Berger, P. B., Lennon, R. J., Holmes, D. R., and Garratt, K. N. (2004), "Percutaneous Coronary Intervention for ST-Segment and Non-ST Segment Elevation Myocardial Infarction at Hospitals With and Without On-Site Cardiac Surgical Capability," *Mayo Clinic Proceedings*, 79, 738–744.

Smith, H. (1997), "Matching With Multiple Controls to Estimate Treatment Effects in Observational Studies," *Sociological Methodology*, 27, 325–353.