Published in final edited form as:
Am Stat. 2011 October 1; 65(4): 229–238. doi:10.1198/tas.2011.11072.

Matching for Several Sparse Nominal Variables in a Case-Control Study of Readmission Following Surgery

José R. Zubizarreta, **Caroline E. Reinke**, **Rachel R. Kelz**, **Jeffrey H. Silber**, and **Paul R. Rosenbaum**

José Zubizarreta is a Doctoral Student, and Paul Rosenbaum is a Professor, Department of Statistics, The Wharton School, University of Pennsylvania, 473 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340. Caroline E. Reinke is an Instructor of Surgery, Rachel R. Kelz is an Assistant Professor of Surgery, and Jeffrey H. Silber is a Professor of Pediatrics at the University of Pennsylvania School of Medicine, Philadelphia, PA 19104.

Abstract

Matching for several nominal covariates with many levels has usually been thought to be difficult because these covariates combine to form an enormous number of interaction categories with few if any people in most such categories. Moreover, because nominal variables are not ordered, there is often no notion of a "close substitute" when an exact match is unavailable. In a case-control study of the risk factors for read-mission within 30 days of surgery in the Medicare population, we wished to match for 47 hospitals, 15 surgical procedures grouped or nested within 5 procedure groups, two genders, or $47 \times 15 \times 2 = 1410$ categories. In addition, we wished to match as closely as possible for the continuous variable age (65–80 years). There were 1380 readmitted patients or cases. A fractional factorial experiment may balance main effects and low-order interactions without achieving balance for high-order interactions. In an analogous fashion, we balance certain main effects and low-order interactions among the covariates; moreover, we use as many exactly matched pairs as possible. This is done by creating a match that is exact for several variables, with a close match for age, and both a "near-exact match" and a "finely balanced match" for another nominal variable, in this case a $47 \times 5 = 235$ category variable representing the interaction of the 47 hospitals and the five surgical procedure groups. The method is easily implemented in R.

Keywords

Assignment algorithm; Caliper match; Finely balanced match; Near-exact match

1. A CASE-CONTROL STUDY OF READMISSION FOLLOWING SURGERY IN MEDICARE

Readmission to an acute care hospital within 30 days of discharge for a surgical procedure is not an entirely unambiguous event, but it often reflects some complication of surgery such as a wound infection. Building upon the earlier Surgical Outcome Study (Silber et al. 2001,

correspondence: josezubi@wharton.upenn.edu.

2005), the Obesity and Surgical Outcomes Study or OBSOS (Silber et al. 2011a, 2011b) abstracted charts for Medicare patients aged 65–80 undergoing surgery at 47 hospitals in Illinois, New York, and Texas. Medicare is the U.S. government program that funds health care for people over the age of 65, and Medicare claims were merged with the data obtained by chart abstraction. The patients had one of five groups of surgical procedures—colectomy with cancer, colectomy without cancer, knee surgery, hip surgery without a fracture, or thoracotomy—and these five groups were subdivided into 15 specific procedures, such as knee replacement (ICD-9 8154) or knee repair (ICD-9 8155). We are currently engaged in a nested case-control study of risk factors for readmission following surgery in the OBSOS data, and here we illustrate the techniques used to simultaneously control for several nominal variables with many categories.

A nested or synthetic case-control study (Mantel 1973) is built from a single cohort, here the entire OBSOS study, by taking all of the cases in this cohort, here all of the readmitted patients, and comparing these cases to a suitable comparison group of noncases from the same cohort, commonly called “controls.” Nested case control studies avoid one of the many problems of case-control studies, namely selection bias due to selecting cases and controls from somewhat different cohorts. Because our current interest concerns characteristics of patients that place them at increased risk of readmission, we wished to control for certain other characteristics that were not part of our current interest. In particular, a patient undergoing a colectomy for colon cancer or a thoracotomy for lung cancer may be in the process of dying from cancer, so readmission would mean something quite different from a readmission for an elective knee replacement in a comparatively healthy patient. For this reason, we wished to compare cases and controls undergoing the same surgical procedure. Hospitals vary in both their ability and their inclination to either readmit patients or to provide outpatient care instead. An uncrowded hospital with limited resources for outpatient care may readmit a patient for a wound infection where a crowded hospital might have treated the very same patient at an out-patient clinic without readmission and with several home visits by a nurse. Although the role hospitals play in readmission is an important aspect of health economics and health services research, our case-control study is intended to provide information useful to surgeons at or near the time of surgery. Are there types of patients at high risk of readmission?

For matching, there were 47 hospitals, 15 surgical procedures, and 2 genders, making $47 \times 15 \times 2 = 1410$ categories, plus the continuous variable age, 65–80. If one rounds age to the nearest year, then among the 1380 readmitted patients in OBSOS, 500 readmitted patients (500/1380 = 36%) cannot be matched exactly for hospital, procedure, gender, and age. If instead one uses five 3-year age groups, such as $65 \leq \text{age} < 68$, then 247 readmitted patients (247/1380 = 18%) cannot be matched exactly for hospital, procedure, gender, and 3-year age category. Our goal in this article is to propose a new approach to matching for several sparse nominal variables. Described informally, our algorithm pairs identical people as often as it can, pairs similar people when identical people are unavailable, and exactly balances marginal and pairwise-joint distributions of covariates in case and matched control groups. In other words, unavoidable mismatches in some individual pairs nonetheless combine to produce marginal distributions that are exactly the same; here, of course, the marginal distributions ignore who is paired with whom. The match produced by our algorithm uses all

1380 readmitted patients, is an exact match for the 15 surgical procedures and for gender, exactly balances the 47 hospitals, exactly balances the interaction of the 47 hospitals and the 5 procedure groups (47 × 5 = 235 interaction categories), the 99.5% quantile of the 1380 absolute matched pair case-minus-control differences in ages is 1.15 years; moreover, 1075/1380 = 78% of the readmitted patients are exactly matched for the 47×15×2 = 1410 categories of hospital, procedure and gender.

Matching can usefully be combined with model-based adjustments, such as covariance adjustment or conditional logit regression. In a simulation study, Rubin (1979) compared unmatched covariance adjustment of random samples, matched analyses without model-based adjustments, and covariance adjustment of matched pairs. He concluded that unmatched co-variance adjustment was not robust to model misspecification, sometimes increasing rather than reducing bias from covariates. In contrast, he found model-based adjustments of matched pair differences were robust to model misspecification. See also Rubin (1973).

Used alone without matching, a model-based analysis of readmission, say by logit regression, would need to address the many categories of hospital, procedure, gender and their interactions and do so with 1380 readmitted patients. For instance, one cannot fit a logit model predicting 1380 readmissions with 1410 categories coded as indicator variables, but one could fit a logit model with fewer indicators if one knew that interactions are zero. In the example, interactions are not unlikely: in most hospitals, the hip and knee surgeries are performed by orthopedic surgeons as elective surgery on relatively healthy patients, whereas the colectomies and thoracotomies are performed by general surgeons, often on severely ill patients who may have entered through the emergency room. A stratified analysis for hospital, procedure, gender, and 3-year age categories using either the Mantel–Haenszel test or conditional logit regression to control the strata would largely ignore those readmissions that occur in strata without controls, so that 247/1380 = 18% of readmitted patients would not contribute to such an analysis. In contrast, the matched analysis addresses interactions and uses all 1380 readmissions.

In modern practice, a multivariate match is the solution to a combinatorial optimization problem subject to various constraints. For instance, Hansen's (2007) optmatch package in R matches by combinatorial optimization; see Bergstralh, Kosanke, and Jacobsen (1996) for a related approach in SAS. One such constraint, called "fine balance," requires that a nominal variable has exactly the same distribution in case and control groups, without constraining who is matched to whom; see Rosenbaum (1989, sec. 3.2) and Rosenbaum, Ross, and Silber (2007). This nominal variable may be formed as the interaction or direct product of several other nominal variables, so it may have many levels. In our example, the nominal variable has 235 categories. Fine balance is a hard constraint: the algorithm is required to produce a solution that satisfies a hard constraint. It is also possible to impose a soft or elastic constraint, one that is satisfied as often or as nearly as is possible among solutions that satisfy all of the hard constraints. "Near-exact" matching for a nominal variable is one such soft constraint: the pairs have identical values of the nominal variable as often as is possible for the data at hand. A caliper on a continuous variable is another soft constraint: we may require cases and controls to differ in their exact ages by at most 1 year

whenever this is possible, tolerating a few differences larger than 1 year. Soft constraints, such as near-exact matching or calipers, are imposed using a standard optimization technique called a penalty function, as described precisely for matching in the Appendix. The procedure we propose here for several sparse nominal variables combines fine balance and near-exact matching for the same nominal interaction variable, so the marginal distributions are identical and the individual pairs as exactly matched as they can be. For brevity, a match that is both exactly finely balanced and nearly exact for the same variable will be called "x-fine." In the example, the $47 \times 5 = 235$ categories of hospital and procedure group will be x-fine. For example, each hospital will have exactly the same number of knee-surgery cases as it has knee surgery controls, and these cases and controls will be paired with each other whenever feasible.

Case-control studies increasingly use genetic information as risk factors (Khoury et al. 2004). As a method of adjustment for covariates, matching has the advantage that a single match for covariates may be used to study thousands of genetic risk factors; see Heller et al. (2009).

Section 2 discusses why sparse nominal variables present special difficulties for matching, while Section 3 describes in detail the covariate balance obtained in the match mentioned in the previous paragraph. Section 4 compares the x-fine match in Section 3 to a conventional match which uses some exact matching and a Mahalanobis distance (Rubin 1980). As an illustration, Section 5 looks at one specific risk factor for read-mission, namely BMI. Section 6 discusses issues and options that deserve consideration when the method is used in other applications. After a brief summary in Section 7, the Appendix describes the implementation of the matching: it is a minimum distance, near-exact, exactly finely balanced match, with a 1-year caliper on age implemented using a penalty function, that is, standard techniques assembled in a new way to address sparse nominal covariates. (It is easy to do in R.) Technical terms are defined precisely in the Appendix.

2. WHY DO SPARSE NOMINAL VARIABLES PRESENT DIFFICULTIES FOR MATCHING?

Using five 3-year age categories for ages 65–80, the $47 \times 15 \times 2 = 1410$ categories of hospital, procedure, and gender would become $5 \times 1410 = 7050$ categories, containing 1380 cases. Exact matching with coarse categories (such as 3-year age categories) is possible when there are an enormous number of potential controls and not too many categories, but in other situations some categories will contain too few controls to permit an exact match.

In randomized experiments, covariate balance is achieved by flipping coins, but even in randomized experiments, when there are many categories of pretreatment covariates, some device, such as Efron's (1971) sequential biased coin design, is needed if one wishes to ensure that no categories are substantially out of balance. For instance, with 1000 categories, four subjects per category, and complete randomization, one expects 12.5%, or 125 categories, to contain only treated subjects or only controls. In observational cohort studies, propensity scores balance observed covariates with the aid of probability (Rosenbaum and Rubin 1985; Stuart 2010), but as in randomized experiments, the laws of chance produce

their expected effects only if those laws act repeatedly. With many small categories, the laws of chance acting alone are expected to leave frequent imbalances. Sometimes chance needs a push from an optimization algorithm that has a definite goal in mind.

If an exact match for several sparse nominal variables is not an attainable goal, then what attainable goal should be used instead? In the design of experiments, an incomplete form of balance for nominal variables plays an important role in fractional factorial designs or more specifically in orthogonal arrays [e.g., Rao (1947), Hedayat, Sloane, and Stufken (1999)]. In an orthogonal array with f factors and strength $s < f$, there is a balanced factorial in any subset of s of the f factors without joint balance for all f factors simultaneously. By a limited analogy with this notion of incomplete balance, Tables 2, 3, 4, and 5 will exhibit case-control balance with respect to each nominal variable alone and with respect to certain pair-wise interactions while lacking perfect balance on the entire set of $47 \times 15 \times 2 = 1410$ categories. Moreover, subject to this constraint on the distributions of nominal variables and their low-order interactions in case and control groups, the individual pairs are exactly matched as often as is possible.

For example, hospital 47 had two readmissions among 44 surgical patients. One was a male hip replacement aged 71.9 years, and there was no male hip replacement control in hospital 47 in the same 3-year age category. The other case of readmission was a female right colectomy with cancer aged 78.8 years, and again there was no female right-colectomy-with-cancer control in hospital 47 within 5 years of age.

Table 1 shows how our match handled the two readmissions from hospital 47. Recall that there are 15 specific surgical procedures nested within 5 groups of surgical procedures. Both cases from hospital 47 were matched to controls with the same surgical procedure and gender while differing by less than 1 year in age. The two cases from hospital 47 were not matched to controls from hospital 47, but two controls with the same procedure group from hospital 47 were used as controls for two other cases. Among the 1380 matched pairs, every hospital appears in the case group with a given surgical group exactly the same number of times it appears in the control group with that same surgical group. Moreover, the typical situation is better than in Table 1: specifically, unlike Table 1, in 1075/1380 = 78% of the pairs, the matching is exact for surgical procedure, gender, and hospital, and in 99.5% of pairs the absolute difference in age is at most 1.15 years. That is, the overwhelming majority of pairs are exactly matched for procedure, gender, and hospital, very closely matched for age, and where an exact match was not feasible, a strong form of balance was obtained. This match is described in greater detail in Section 3.

3. DESCRIPTION OF COVARIATE BALANCE IN THE MATCHED COMPARISON

Tables 2, 3, 4, and 5 describe the covariate balance in the matched comparison of 1380 readmitted cases and 1380 controls who were not readmitted. Table 2 counts pairs, not people, and indicates that men were always matched to men, women to women. Table 3 is larger but has the same format: it shows that the 15 surgical procedures were exactly matched.

Table 4 is again of the same format, but it shows a near-exact, finely balanced match, rather than an exact match, for hospitals. Table 4 would be 47 × 47 for the 47 hospitals, but it is abbreviated for display; it shows hospitals 1, 2, …, 10 and 47, plus the marginal totals over all 47 hospitals. The algorithm matched exactly for hospital whenever it could, and it succeeded in 1075/1380 = 78% of the pairs; that is, 78% of pairs fall on the diagonal of Table 4. For instance, 18 cases from hospital 1 were matched to controls from hospital 1, but one case from hospital 1 was matched to a control from hospital 2. As discussed in Section 2, neither case in hospital 47 was matched to a control from hospital 47.

Table 4 shows not only a near-exact match for hospital but also an exactly finely balanced match for hospital. That is to say, the marginal row totals exactly equal the marginal column totals: every hospital is represented in the case group with exactly the same frequency that it is represented in the control group.

Table 5 shows that fine balance extends not just to the 47 hospitals, but to the interaction of the five surgical procedure groups and the 47 hospitals, that is to 47 × 5 = 235 interaction categories. Unlike Table 2–4, Table 5 counts patients, not pairs, so the total count is 2 × 1380 = 2760 patients, not 1380 pairs. Like Table 4, Table 5 is abbreviated for display: it would have 47 rows for the 47 hospitals, but only hospitals 1, 2, …, 10 and 47 and the column totals are shown. In the upper left corner, Table 5 shows that there were eight readmitted cases of colectomy with cancer from hospital 1 among the 1380 pairs and also eight controls of colectomy with cancer from hospital 1, and there is similar fine balance for all of the 47 × 5 = 235 interaction categories.

There is another table worth considering, but it is large and difficult to display, so we describe rather than display it. The table resembles, indeed expands, Table 4, counting pairs, with the case described by the row and the control described by the column. Exact matches appear along the diagonal, and the marginal row and column totals describe marginal distributions in the case and control groups. The table is 235 × 235 where there are 235 = 47× 5 combinations of the 47 hospitals and the five groups of surgical procedures. As in Table 4, the marginal row and column totals in this table are identical, as has already been seen in Table 5. The match is not exact: some pairs are not on the diagonal. However, the total count on the diagonal is as large as possible; specifically, 1075/1380 = 78% of pairs are on this diagonal.

In addition, not seen in the tables, in 99.5% of the 1380 matched pairs, the absolute difference in age was at most 1.15 years.

4. COMPARISON WITH A MATCH BASED ON THE MAHALANOBIS DISTANCE

We now contrast the x-fine match in Section 3 with a conventional approach to multivariate matching in case-control studies. The conventional match was exact for the five procedure groups and minimized a Mahalanobis distance computed from age and indicators for gender, the hospitals, and the procedure subcategories. Table 6 describes this conventional match and is parallel to Table 5 for the x-fine match in Section 3; however, unlike Table 5, Table 6

exhibits substantial imbalances for the interaction of hospital and procedure group. For instance, in Table 6, in hospital 7 there were three readmitted cases of colectomy with cancer but no controls, whereas in Table 5 there were three cases and three controls. Similarly, in Table 6, in hospital 2 there was one readmitted case of colectomy with cancer but there were seven controls.

For a nominal variable, the x-fine match in Section 3 and the conventional match are compared in two ways: (i) η= the number of exactly matched pairs, out of 1380 pairs, and (ii) λ = the sum of the absolute differences in the counts of their marginal distributions. If the marginal distributions are the same, then $\lambda = 0$. Both the x-fine match and the conventional match are exact for the five surgical procedure groups, for gender, and for the interaction of procedure group and gender, so for these variables η= 1380 pairs (100%) and λ=0. The x-fine match is also exact for the 15 specific surgical procedures, whereas the conventional match has a few mismatches, η = 1352 pairs (98%) and λ= 20. For the 47 hospitals, the x-fine match has η= 1075 exact pairs (78%) and exactly the same marginals $\lambda = 0$, while the conventional match has η= 853 exact pairs (62%) and substantially different marginals λ = 242. Because the five procedure groups are matched exactly but hospitals are not, the same values apply to the interaction of procedure groups with hospitals, η= 1075 pairs (78%) and $\lambda = 0$ for the x-fine match and η= 853 pairs (62%) and λ = 242 for the conventional match. The interaction of the 15 specific surgical procedures with the 47 hospitals has η = 1075 pairs (78%) and λ = 306 for the x-fine match and η= 842 pairs (61%) and λ= 516 for the conventional match. The 99.5% quantile of the absolute pair difference in ages is 1.15 years for the x-fine match and is 1.65 years for the conventional match. For hospitals and for the interaction of hospitals with other variables, the conventional match has substantially fewer exactly matched pairs than the x-fine match, and also, ignoring who is matched to whom, the marginal distributions are further apart. Moreover, the conventional match confers no benefit to offset its two disadvantages, namely fewer exact pairs and a larger difference in the marginal distributions.

5. A QUICK LOOK AT ONE RISK FACTOR FOR READMISSION

In this section, we take a brief look at one possible risk factor for readmission, namely body-mass-index or BMI obtained by chart abstraction. The BMI is a measure of obesity: it is the ratio of mass in kilograms to the square of height in meters. The U.S. National Heart, Lung and Blood Institute (http://www.nhlbisupport.com/bmi/) describes a BMI below 18.5 as underweight, 18.5 to 25 as normal weight, 25 to 30 as overweight, and 30 or more as obese; moreover, others describe 35–40 as severe obesity and sometimes 40 or more as morbid obesity. Among patients discharged alive in OBSOS, the median BMI was 28.2. Is BMI a risk factor for readmission following surgery?

For the 1380 readmitted patients and their 1380 matched controls, Figure 1 is a quantile-quantile plot or qq-plot of BMI; see Wilk and Gnanadesikan (1968) or Cleveland (1994, p. 143–149) for discussion of qq-plots. If the distribution of BMIs were the same for readmitted patients and controls, the 1380 points would tend to fall along the 45° line, $x = y$, and in the middle range, perhaps between 22 and 33, the points are close to the line of equality. However, the severely obese and the underweight are both overrepresented among

cases of readmission. Table 7 classifies the 1380 pairs by underweight (BMI < 18.5), severely obese (BMI ≥ 35) and others. In Table 7, both the underweight and the severely obese are overrepresented among the cases of readmission, and the hypothesis of symmetry in Table 7 is rejected at the 0.0024 level by the test of symmetry that generalizes McNemar's test (Agresti 2002, sec. 10.4.1).

The categories in Table 7 were selected after examining the plot. Would similar results be obtained with other categories or without categories? With six categories formed by five cuts at the conventional places, 18.5, 25, 30, 35, and 40, the McNemar *P*-value is 0.017. In Table 7, there are 366 = (2×40)+ 166 + 111 + 3 + 6 patients with BMIs of 35 or more. If BMI categories are not used, but rather the BMIs of these 366 patients are ranked from 1 to 366, with zero ranks for other patients, as suggested by Rosenbaum and Silber (2008), and if a paired permutation test is applied to these ranks in 1380 pairs, the two-sided *P*-value is 0.0081. [See Mehrotra et al. (2006) for discussion of a related procedure.] Doing this same test but cutting at 30 or at 40 instead of 35 both yield *P*-values of 0.027. At the opposite end, in Table 7 there are 80 = (2×1)+41+6+28+3 patients with BMIs below 18.5. Ranking these 80 BMIs from 1 to 80 with rank 80 given to the patient with the smallest BMI of 10.98, rank 0 given to patients with BMIs above 18.5, the two-sided *P*-value is 0.030. If the cut were made at 20 or 22, the analogous *P*-values are 0.018 and 0.029. In short, the impression that the underweight and severely obese are at increased risk of readmission does not depend on the specific category boundaries in Table 7.

If one had compared, without matching, all 1380 readmitted patients to the remaining 14,286 patients discharged alive in OBSOS, a higher body mass index would have (misleadingly) seemed to be associated with a reduced risk of readmission. If the Wilcoxon–Mann–Whitney two-sample test is applied to compare BMIs in these two unmatched groups, the two-sided *P*-value testing no difference in BMI is 1.1×10^{-6}, with a Hodges–Lehmann point estimate of a shift of −0.82 and a 95% confidence interval of [−1.15, −0.49]: that is, it appears that the readmitted have BMIs that are typically 0.82 lower. At least in part, this is because severe obesity is hard on the knees, so the severely obese are substantially overrepresented among knee surgeries—the odds ratio is 4.0 linking a BMI of 35 or more with knee surgery rather than thoracotomy—and readmission was less than half as common among knee surgeries as among thoracotomies and colectomies. The unmatched comparison is comparing patients with very different surgical procedures. A logit model fitted to all 15666 = 1380+14286 patients, not just the matched patients, predicting the 1380 read-missions from BMI as a continuous variable, indicators for the 47 hospitals, indicators for the 15 surgical procedures, an indicator for gender and age, with additive terms on the logit scale, finds that the estimated coefficient of BMI is small and not significantly different from zero (*P* = 0.24). Perhaps this reflects the pattern seen in Figure 1, in which BMI's between 22 and 33 seem unrelated to readmission and both the severely obese and underweight are at increased risk of readmission. However, if the continuous BMI is replaced by two indicators for the two categories in Table 7, then the model does indicate that both the underweight and severely obese are at increased risk of read-mission. Presumably, the careful user of log it regression would discover the inadequacy of the logit

model with continuous BMI aided by examination of logit regression diagnostics (Pregibon 1981).

6. TAILORING THE METHOD FOR USE IN OTHER APPLICATIONS

In the Medicare readmission example, the 47 × 5 = 235 categories of hospital and procedure group were finely balanced, meaning their marginal distributions were the same in case and matched control groups. In addition, an attempt was made whenever possible to pair individuals with the same values of all 15 surgical procedures, gender, and hospital with at most a 1-year difference in age; however, this was often but not always possible. As is discussed with greater precision and with technical detail in the Appendix, this attempt at close individual pairing used a distance matrix with one row for each case and one column for each potential control. The distance between a case and a potential control was zero if they differed in age by at most one year, were of the same gender, and had the same surgical procedure at the same hospital. Each mismatch on a nominal variable caused the distance to increase. Because we thought that the surgical procedure was the most important covariate, a mismatch on surgical procedure counted most with the biggest increase in the distance. A mismatch on gender incurred a smaller but still very large increase in distance, and a mismatch on hospital incurred the smallest but still large increase in distance. Although we wanted at most a 1-year difference in age, a 2-year difference in age increased the distance by the same amount as a difference in gender, rising linearly with excess age difference beyond 1 year. These differences in magnitudes of the increments in the distances were widely spaced, so that a single mismatch on surgical procedure would be avoided if at all possible, even at the price of mismatching many pairs on gender or age or hospital. These priorities were set by the clinicians. As seen in Tables 2 and 3, surgical procedure and gender were exactly matched in every pair.

In general, the user of this technique will make two decisions and then turn the matter over to an algorithm which will find the best match subject to those two decisions. One decision is to select a nominal covariate for fine balance. In the example, the finely balanced covariate was the 47×5 = 235 categories of hospital and procedure group. The only requirement here is that the number of potential controls in each category must at least equal the number of cases in that category. The second decision is to define the distance matrix. In defining the distance matrix, the user has a great deal of latitude, and the remainder of this section is devoted to discussing some of the available options and considerations.

The categories of a sparse nominal variable are sometimes nested within category groups. For instance, knee replacement and knee repair were two specific surgical procedures within the group of knee surgeries. The 47 hospitals were nested within three states, Illinois, New York, and Texas. The distance between categories of a nominal variable may be coded as larger if there is a crossing of category groups, with knee replacement coded as closer to knee repair than to a right colectomy with cancer. The algorithm would then prefer mismatches that stayed within a category group to mismatches that cross category groups. We would have used this device if we had been unable to match exactly for the 15 surgical procedures.

Our distances were zero if there was agreement on surgical procedure, gender, and hospital and an age difference of at most 1 year. Instead, if there are additional continuous or binary covariates, the initial distances may be some form of Mahalanobis distance (Rubin 1980) computed from the continuous covariates. Two independent observations from the same *P*-dimensional Normal distribution are expected to differ by a Mahalanobis distance of 2*P*. If the distance increments for nominal mismatches are large compared to 2*P*, then the distance matrix will handle nominal covariates as in Section 3 but will also try when possible to pair individuals who are close in terms of the continuous covariates. The Mahalanobis distance is suitable for the multivariate Normal distribution, but it can behave oddly with long-tailed distributions or rare binary covariates, so it is safest to use—as was, in fact, done in Section 4—a slightly modified Mahalanobis distance which can be computed in a few lines of R code (Rosenbaum 2010, described in sec. 8.3 with R code smahal in §13.11).

In a cohort study comparing treated and control groups, one may balance many covariates stochastically by matching on a single variable, namely an estimate of the propensity score, which is the conditional probability of the treatment given the covariates. When estimating a treatment effect in a cohort study, in large samples, if it suffices to adjust for a vector **x** of covariates then it also suffices to balance **x** by adjusting for the scalar propensity score given **x**; see Rosenbaum and Rubin (1983, Theorem 3). As noted in Section 2, stochastic balance is typically inadequate with sparse nominal variables, essentially because the sample size is not large within each sparse category. However, the method illustrated in Section 3 may be combined with matching for the propensity score by placing a caliper on the propensity score similar to the caliper on age in Section 3. A propensity score needs to condition on all of **x**, so if there is near-exact matching for a sparse nominal covariate as in Section 3, one may consider estimating the weights for other covariates using conditional log it regression given the nominal covariate.

The three techniques in the previous three paragraphs may be used singly or in combination. Rosenbaum and Rubin (1985) matched using calipers on a propensity score and a Mahalanobis distance within calipers, albeit without sparse nominal covariates.

7. SUMMARY: BALANCING INTERACTIONS WITH PAIRS OF COVARIATES WHILE USING THE MAXIMUM NUMBER OF EXACTLY MATCHED PAIRS

When several nominal variables have many categories, there are an enormous number of interaction categories, and an exact match for all of the nominal variables is not possible for many cases. Borrowing a notion from fractional factorial designs, the match in Section 3 obtains perfect marginal balance on certain two-covariate interactions while matching as many cases exactly as is possible. The Appendix describes the required calculations, which are straightforward in R: essentially, a certain matrix is created and this matrix is passed to an optimization algorithm which returns the matched sample.

Acknowledgments

This research was supported in part through a grant from the U.S. National Institute of Diabetes, Digestive and Kidney Diseases (NIDDK R01 DK 073671) and a grant from the U.S. National Science Foundation.

APPENDIX: HOW THE MATCHING WAS DONE

The matching combined five standard techniques in a new way to address sparse nominal covariates; see Rosenbaum (2010, secs. 8.4, 9.2, 9.3, and 10) for separate discussion of these techniques. Essentially, the goal was to match exactly for certain covariates (e.g., the 15 surgical procedures, gender), perfectly balance certain marginal distributions (e.g., the 47 × 5 = 235 interactions of hospital and procedure group), obtain a close match for age and, subject to these several requirements, match exactly for all the nominal variables as often as possible.

First, the five surgical procedure groups were matched separately, one at a time, ensuring an exact match for the five procedure groups, while replacing one large matching problem by five somewhat smaller problems. The second technique is the "optimal assignment algorithm" which begins with a distance matrix $\mathbf{\Delta}$ with distance δ_{ij} between case i in row i and potential control j in column j, $i = 1,\ldots,I$, $j = 1,\ldots,J$, $I \leq J$, and it pairs each row with a different column to minimize the sum of the distances within matched pairs; see, for instance, Papadimitriou and Steiglitz (1982, sec. 11.2). Bertsekas' (1981) algorithm for optimal assignment is available in R (R Development Core Team 2007) as the pairmatch function of Hansen's (2007) optmatch package. See Bergstralh, Kosanke, and Jacobsen (1996) for similar devices in SAS.

The remaining three techniques define $\mathbf{\Delta}$, which begins as an $I \times J$ matrix of zeros and ends as a $J \times J$ matrix. The third technique is "near-exact matching." For case i and potential control j the distance δ_{ij} was increased in row i and column j of distance matrix $\mathbf{\Delta}$ if case i and control j differed in terms of the 15 surgical procedures ($\delta_{ij} \leftarrow \delta_{ij} + 10^5$), if they had different genders ($\delta_{ij} \leftarrow \delta_{ij} + 1.5 \times 10^4$), or if they had operations in different hospitals ($\delta_{ij} \leftarrow \delta_{ij} + 500$); this is known as "near exact" matching (or "almost exact" matching), because an enormous price is paid for specific forms of mismatch, but unlike exact matching the algorithm may return some such mismatches if there is no alternative. The hierarchy of penalty sizes (10^5, 1.5×10^4, and 500) meant that an exact match for surgical procedure was vastly more important than an exact match for gender which was vastly more important than an exact match for hospital. In Tables 2 and 3, near exact matching yielded an exact match for gender and surgical procedure, whereas in Table 4, it yielded an exact match for hospital in 78% of pairs.

Fourth, a close match on age was ensured using a caliper imposed with the aid of a penalty function. A caliper of one year does not further increase δ_{ij} if i and j have ages that differ by one year or less. As noted by Cochran and Rubin (1973), calipers are better than age categories, because categories prevent the matching of individuals who are close in age but fall on opposite sides of the category boundary. Rather than add a constant to δ_{ij} when i and j differ by more than one year in age, a penalty function (Luenberger 1984, sec 12.1) is used, so δ_{ij} increases slightly for small violations of the caliper constraint but increases dramatically for large violations. Specifically, if case i and potential control j had a difference in age of a_{ij}, then $\delta_{ij} \leftarrow \delta_{ij} + \max[0, \min\{\beta, \beta(|a_{ij}| - 1)\}]$ with $\beta = 1.5 \times 10^4$, so $|aij| \leq 1$ yields no increment in δ_{ij} and $|a_{ij}| \geq 2$ yields the full increment of 1.5×104, with linear interpolation on $1 < |aij| < 2$. In $\mathbf{\Delta}$, a two-year difference in age, $|a_{ij}| = 2$, increases δ_{ij} by the

same amount as a mismatch for gender, whereas a one-year difference in age, $|a_{ij}| = 1$, does not increase δ_{ij} In 84 pairs, the one-year age caliper was violated with an absolute difference in age greater than one year, but the penalty function meant that the algorithm kept searching for small a_{ij}'s subject to the other constraints, and the upper 99.5% point of the absolute difference in ages in matched pairs was 1.15 years, a tiny violation of the caliper.

Exact equality of the marginal distributions in Tables 4 and 5 is known as "fine balance," and this is the fifth technique. Fine balance is obtained by adding $J - I$ patterned rows to the distance matrix, $\mathbf{\Delta}$, making it a square $J \times J$ matrix, in which the added rows force the removal of specific numbers of controls from specific groups; see Rosenbaum, Ross, and Silber (2007) for easy steps required to create these patterned additional rows. The optimal assignment algorithm was applied to this enlarged $J \times J$ matrix $\mathbf{\Delta}$. Fine balance may alternatively be obtained using network optimization, and this may be more efficient in its use of computer memory than storing the $J - I$ patterned rows of $\mathbf{\Delta}$; see Rosenbaum (1989, sec. 3.2) for a description and see Dan Yang's finebalance package in R for an implementation. The superimposition of matching with fine balance and near-exact matching for the same sparse nominal covariates is what produced the balance on marginal distributions in Tables 4 and 5 together with the substantial diagonal counts in Table 4. Indeed, had Table 5 been arranged in the format of Table 4, with $47 \times 5 = 235$ rows and 235 columns, then the total count on its diagonal would be as large as possible subject to the other constraints.

REFERENCES

Agresti, A. Categorical Data Analysis. New York: John Wiley; 2002.

Bergstralh EJ, Kosanke JL, Jacobsen SL. Software for Optimal Matching in Observational Studies. Epidemiology. 1996; 7:331–332. [PubMed: 8728456]

Bertsekas DP. A New Algorithm for the Assignment Problem. Mathematical Programming. 1981; 21:152–171.

Cleveland, WS. The Elements of Graphing Data. Summit, NJ: Hobart Press; 1994.

Cochran WG, Rubin DB. Controlling Bias in Observational Studies: A Review. Sankhya. 1973; 35:417–446.

Efron B. Forcing a Sequential Experiment to be Balanced. Biometrika. 1971; 58:403–417.

Hansen BB. Optmatch. R News. 2007; 7:18–24.

Hedayat, AS.; Sloane, NJA.; Stufken, J. Orthogonal Arrays: Theory and Applications. New York: Springer; 1999.

Heller R, Manduchi E, Small DS. Matching Methods for Observational Microarray Studies. Bioinformatics. 2009; 25:904–909. [PubMed: 19098026]

Khoury, MJ.; Little, J.; Burke, W. Human Genome Epidemiology. New York: Oxford; 2004.

Luenberger, DG. Linear and Nonlinear Programming. Reading, MA: Addison-Wesley; 1984.

Mantel N. Synthetic Retrospective Studies and Related Topics. Biometrics. 1973; 29:479–486. [PubMed: 4793136]

Mehrotra DV, Li X, Gilbert PB. A Comparison of Eight Methods for the Dual-Endpoint Evaluation of Efficiency in a Proof-of-Concept HIV Vaccine Trial. Biometrics. 2006; 62:893–900. [PubMed: 16984333]

R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation; 2007. http://www.R-project.org.

Papadimitriou, CH.; Steiglitz, K. Combinatorial Optimization: Algorithms and Complexity. Englewood Cliffs, NJ: Prentice Hall; 1982. Reprinted: New York: Dover

Pregibon D. Logistic Regression Diagnostics. The Annals of Statistics. 1981; 9:705–724.

Rao CR. Factorial Experiments Derivable from Combinatorial Arrangements of Arrays. Journal of the Royal Statistical Society. 1947; 9(Supplement):128–139.

Rosenbaum PR. Optimal Matching in Observational Studies. Journal of American Statistical Association. 1989; 84:1024–1032.

Rosenbaum, PR. Design of Observational Studies. New York: Springer; 2010.

Rosenbaum PR, Ross RN, Silber JH. Minimum Distance Matched Sampling with Fine Balance in an Observational Study of Treatment for Ovarian Cancer. Journal of American Statistical Association. 2007; 102:75–83.

Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies of Causal Effects. Biometrika. 1983; 70:41–55.

Rosenbaum PR, Rubin DB. Constructing a Control Group by Multivariate Matched Sampling Methods that Incorporate the Propensity Score. The American Statistician. 1985; 39:33–38.

Rosenbaum PR, Silber JH. Aberrant Effects of Treatment. Journal of American Statistical Association. 2008; 103:240–247.

Rubin DB. Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. Biometrics. 1973; 29:185–203.

Rubin DB. Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. Journal of the American Statistical Association. 1979; 74:318–328.

Rubin DB. Bias Reduction using Mahalanobis Metric Matching. Biometrics. 1980; 36:293–298.

Silber JH, Rosenbaum PR, Even-Shoshan O, Mi LY, Kyle FA, Teng Y, Bratzler DW, Fleisher LA. Estimating Anesthesia Time using the Medicare Claim: A Validation Study. Anesthesiology. 2011a; 115:322–333. [PubMed: 21720242]

Silber JH, Rosenbaum PR, Ross RN, Even-Shoshan O, Kelz RR, Neuman MD, Reinke CE, David G, Saynisch PA, Kyle F, Bratzler DW, Fleisher LA. Medical and Financial Risks Associated with Surgery in the Elderly Obese. Annals of Surgery. 2011b to appear.

Silber JH, Rosenbaum PR, Trudeau ME, Chen W, Zhang X, Lorch S, Rapaport-Kelz R, Mosher RE, Even-Shoshan O. Pre-operative Antibiotics and Mortality in the Elderly. Annals of Surgery. 2005; 242:107–114. [PubMed: 15973108]

Silber JH, Rosenbaum PR, Trudeau ME, Even-Shoshan O, Chen W, Zhang X, Mosher RE. Multivariate Matching and Bias Reduction in the Surgical Outcomes Study. Medical Care. 2001; 39:1048–1064. [PubMed: 11567168]

Stuart EA. Matching Methods for Causal Inference. Statistical Science. 2010; 25:1–21. [PubMed: 20871802]

Wilk MB, Gnanadesikan R. Probability Plotting Methods for the Analysis of Data. Biometrika. 1968; 55:1–17. [PubMed: 5661047]

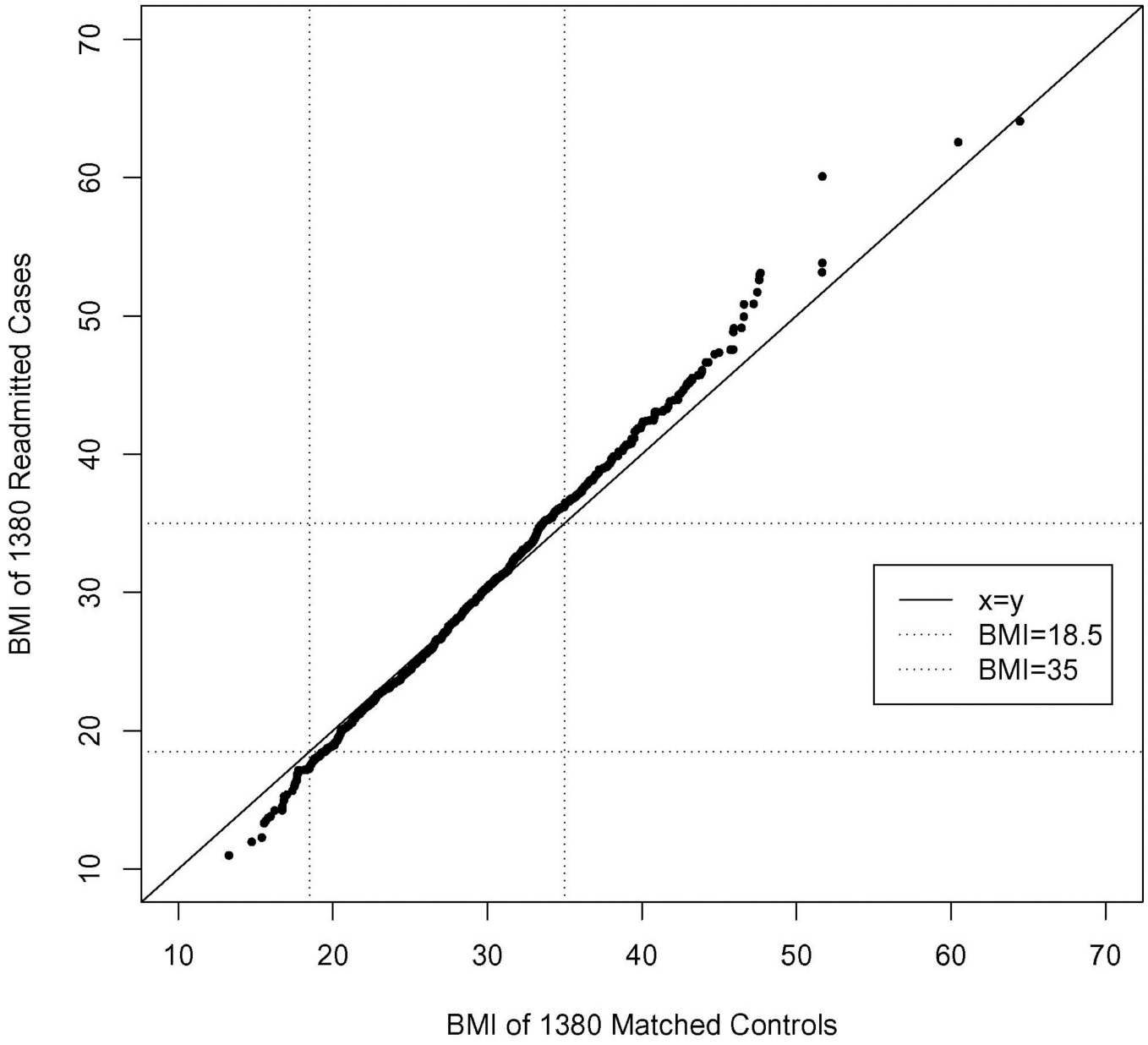

Figure 1.
qq-Plot of BMI for cases and controls.

Table 1

The four case-control pairs that involve patients from hospital 47. In this table, cases and controls are exactly matched for procedure and gender, and matched within one year for age, and the number of cases of hip replacement from hospital 47 equals the number of controls (one of each), and the number of cases of right colectomy with cancer from hospital 47 equals the number of controls (one of each). Hospitals 25, 31, and 39 are similarly balanced, but not in these four pairs.

Pair	Case/control	Surgical procedure	Age	Gender	Hospital
1	Readmitted case	Hip replacement ICD-9 8151	71.9	Male	47
	Control	Hip replacement ICD-9 8151	71.0	Male	31
2	Readmitted case	Hip replacement ICD-9 8151	65.8	Female	39
	Control	Hip replacement ICD-9 8151	66.5	Female	47
3	Readmitted case	Right colectomy with cancer	78.8	Female	47
	Control	Right colectomy with cancer	78.9	Female	25
4	Readmitted case	Right colectomy with cancer	73.1	Female	25
	Control	Right colectomy with cancer	73.8	Female	47

Table 2

For 1380 case-control pairs, the table displays an exact match for gender. The table counts 1380 pairs, where the row refers to the case, the column to the control, and because all counts are along the diagonal, the matching is exact.

	Control, not readmitted		
Readmitted Case	**Female**	**Male**	**Total**
Female	703	0	703
Male	0	677	677
Total	703	677	1380

Table 3

For 1380 case-control pairs, the table displays an exact match for the 15 categories of surgical procedure. The table counts 1380 pairs, where the row refers to the case, the column to the control, and because all counts are along the diagonal, the matching is exact. For colectomy, “wc” means with cancer, “woc’ means without cancer, “L”, “R” and “T” signify, respectively, left, right or total/other colectomy. For Hip surgery without fracture, the categories refer to ICD-9 codes, “H1” is 8151, “H2” is 8152, and “H3” is 8153. For knee replacements and repairs, the categories again refer to ICD-9 codes, “K4” is 8154 and “K5” 8155. For Thoracotomy, “L” is a lobectomy, “P” is a pneumonectomy, “W” is a wedge resection, and “O” is other thoracotomy.

	Control														
	Colectomy						Hip			Knee		Thoracotomy			
	W cancer			Wo cancer											
Case	L	R	T	L	R	T	H1	H2	H3	K4	K5	L	P	W	O
Colectomy wcL	97	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Colectomy wcR	0	204	0	0	0	0	0	0	0	0	0	0	0	0	0
Colectomy wcT	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0
Colectomy wocL	0	0	0	119	0	0	0	0	0	0	0	0	0	0	0
Colectomy wocR	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0
Colectomy wocT	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
Hip, H1	0	0	0	0	0	0	199	0	0	0	0	0	0	0	0
Hip, H2	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0
Hip, H3	0	0	0	0	0	0	0	0	42	0	0	0	0	0	0
Knee, K4	0	0	0	0	0	0	0	0	0	301	0	0	0	0	0
Knee, K5	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0
Thoracotomy-L	0	0	0	0	0	0	0	0	0	0	0	182	0	0	0
Thoracotomy-P	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0
Thoracotomy-W	0	0	0	0	0	0	0	0	0	0	0	0	0	116	0
Thoracotomy-O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 4

Distribution over 47 hospitals of 1380 readmission cases and 1380 paired controls. Only hospitals 1, 2, …, 10 and 47 are shown. The table exhibits exact fine balance, in that the row and column margins are identical, and near-exact matching, in that the total on the diagonal is as large as possible.

	Control's Hospital												
Case's Hospital	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**…**	**47**	**Total**
1	18	1	0	0	0	0	0	0	0	0	…	0	32
2	0	23	0	0	0	0	1	0	0	0	…	0	28
3	0	0	29	0	0	0	0	0	0	0	…	0	33
4	0	0	0	19	0	0	0	0	1	0	…	0	24
5	0	0	0	0	22	0	0	0	0	0	…	0	27
6	0	0	0	0	0	9	0	0	0	0	…	0	10
7	0	1	0	0	0	0	27	0	1	0	…	0	34
8	0	0	0	0	0	0	0	13	0	0	…	0	18
9	0	0	0	0	0	0	0	0	24	1	…	0	32
10	0	0	0	1	0	0	0	1	0	15	…	0	19
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	…	⋮	⋮
47	0	0	0	0	0	0	0	0	0	0	…	0	2
Total	32	28	33	24	27	10	34	18	32	19	…	2	1380

Table 5

Frequency of five procedure-categories by hospital for readmitted cases (R) and controls (C) who were not readmitted. The 47 × 5 interaction of hospital and procedure-category is exactly balanced. Only hospitals 1, 2, …, 10 and 47 are displayed. For instance, there were 8 readmitted colectomies with cancer and 8 control colectomies with cancer from hospital 1.

	Readmitted cases (R) or controls (C) not readmitted									
	Colectomy									
	With cancer		Wo Cancer		Hip		Knee		Thoracotomy	
Hospital	R	C	R	C	R	C	R	C	R	C
1	8	8	4	4	2	2	10	10	8	8
2	5	5	1	1	5	5	16	16	1	1
3	6	6	3	3	8	8	12	12	4	4
4	4	4	0	0	10	10	10	10	0	0
5	7	7	5	5	5	5	3	3	7	7
6	3	3	2	2	1	1	4	4	0	0
7	10	10	3	3	3	3	10	10	8	8
8	3	3	1	1	5	5	4	4	5	5
9	11	11	4	4	8	8	5	5	4	4
10	3	3	0	0	3	3	10	10	3	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
47	1	1	0	0	1	1	0	0	0	0
Total	329	329	150	150	264	264	323	323	314	314

Table 6

Balance obtained by matching exactly for the five procedure groups and using the Mahalanobis distance for other variables, in place of the proposed combination of fine-balance and near-exact matching that produced Table 5. Unlike Table 5, this table displays substantial imbalances for this interaction, although the column totals are the same.

	Readmitted cases (R) or controls (C) not readmitted									
	Colectomy									
	With cancer		Wo cancer		Hip		Knee		Thoracotomy	
Hospital	R	C	R	C	R	C	R	C	R	C
1	8	9	4	2	2	2	10	10	8	3
2	5	8	1	7	5	8	16	16	1	4
3	6	5	3	1	8	6	12	12	4	3
4	4	2	0	1	10	8	10	10	0	0
5	7	6	5	5	5	5	3	4	7	12
6	3	5	2	2	1	2	4	4	0	0
7	10	14	3	0	3	3	10	8	8	6
8	3	4	1	2	5	4	4	5	5	4
9	11	7	4	2	8	10	5	5	4	4
10	3	5	0	2	3	2	10	10	3	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
47	1	0	0	0	1	0	0	0	0	0
Total	329	329	150	150	264	264	323	323	314	314

Table 7

Readmitted cases and matched controls by underweight, BMI < 18.5, and morbid obesity, BMI ≥ 35. Readmission is more common among both the underweight and the morbidly obese.

	Matched Control			
Readmitted case	**BMI < 18.5**	**18.5 ≤ BMI < 35**	**BMI ≥ 35**	**Total**
BMI < 18.5	1	41	6	48
18.5 ≤ BMI < 35	28	984	111	1123
BMI ≥ 35	3	166	40	209
Total	32	1191	157	1380