Contrasting evidence within and between institutions that supply treatment in an observational study of alternative forms of anesthesia

José R. Zubizarreta, Mark Neuman, Jeffrey H. Silber, Paul R. Rosenbaum[1]

University of Pennsylvania, Philadelphia

Abstract. In a randomized trial, subjects are assigned to treatment or control by the flip of a fair coin. In many nonrandomized or observational studies, subjects find their way to treatment or control in two steps, either or both of which may lead to biased comparisons. By a vague process perhaps affected by proximity or sociodemographic issues, subjects find their way to institutions that provide treatment. Once at such an institution, a second process, perhaps thoughtful and deliberate, assigns individuals to treatment or control. In the current paper, the institutions are hospitals, and the treatment under study is the use of general anesthesia alone versus some use of regional anesthesia during surgery. For a specific operation, the use of regional anesthesia may be typical in one hospital and atypical in another. A new matched design is proposed for studies of this sort, one that creates two types of nonoverlapping matched pairs. Using a new extension of optimal matching with fine balance, pairs of the first type exactly balance treatment assignment across institutions, so each institution appears in the treated group with the same frequency that it appears in the control group; hence, differences between institutions that affect everyone in the same way cannot bias this comparison. Pairs of the second type compare institutions that assign most subjects to treatment and other institutions that assign most subjects to control, so each institution is represented in the treated group if it typically assigns subjects to treatment or alternatively in the control group if it typically assigns subjects to control, and no institution appears in both groups. By and large, in the second type of matched pair, subjects became treated subjects or controls by choosing an

[1] *Address for correspondence:* Department of Statistics, The Wharton School, University of Pennsylvania, 473 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 USA. Supported by a grant from the U.S. National Science Foundation and grant RO1-DK07-3671 from the US National Institute of Diabetes, Digestive and Kidney Diseases. E-mail: josezubi@wharton.upenn.edu. 21 December 2011.

institution, not by a thoughtful and deliberate process of selecting subjects for treatment within institutions. The design provides two evidence factors, that is, two tests of the null hypothesis of no treatment effect that are independent when the null hypothesis is true, where each factor is largely unaffected by certain unmeasured biases that could readily invalidate the other factor. The two factors permit separate and combined sensitivity analyses, where the magnitude of bias affecting the two factors may differ. The case of knee surgery in the study of regional versus general anesthesia is considered in detail.
Keywords: Evidence factor; fine balance; optimal subset matching; sensitivity analysis.

1 Introduction: Regional or General Anesthesia; Outline

1.1 Does the method of anesthesia affect outcomes after surgery?

Anesthesia for surgery is intended to temporarily interrupt (i) the sensation of pain, (ii) awareness of the procedure, and (iii) movement by the patient that may interfere with surgery (Wiklund and Rosenbaum 1997). For procedures involving the extremities, regional anesthesia, most often involving injection of local anesthetic at specific sites within the spinal column (i.e. epidural or intrathecal injection) provides analgesia and immobility, and may be used with intravenous or inhaled medications to blunt or eliminate awareness. This approach is contrasted with techniques—which we refer to collectively as "general anesthesia alone"—that omit regional anesthesia and rely on any of several combinations of inhaled and intravenous medications. While use of regional anesthesia has been suggested to reduce the incidence of postoperative pain (Block et al. 2003), complications and mortality (Rogers et al. 2000) compared to general anesthesia alone, the small sample sizes of most studies have not yielded firm conclusions regarding patient outcomes or health care costs.

The Obesity and Surgical Outcomes study abstracted charts for nearly 16,000 Medicare

patients at 47 hospitals in Illinois, New York and Texas, undergoing knee and hip surgery, colectomy or thoracotomy. The abstracted charts were combined with administrative data from Medicare, the U.S. government's program that provides health care to the elderly. Chart abstraction provided information about: (i) type of anesthesia, (ii) type of diabetes, (iii) obesity as measured by the body mass index or BMI, (iv) physiological information, such as systolic blood pressure, with physiological information summarized in an approximate Acute Physiology and Chronic Health Evaluation (APACHE) score (Knaus et al. 1991), and (v) the American Society of Anesthesiologists (ASA) physical status classification. Medicare claims provided survival follow-up, and information about comorbid conditions, complications, length of stay, and readmission.

Some hospitals make extensive use of regional anesthesia while other hospitals typically use general anesthesia alone. Within a given hospital, some patients may receive some regional anesthesia while other patients receive only general anesthesia. Presumably these two processes operate in different ways. With the exception of obstetric care, patients do not typically choose a hospital based on anesthesia practices at that hospital; rather, patients typically end up at one hospital rather than another based on proximity, affiliation, or perhaps reputation. Within a given hospital, the choice between some regional anesthesia or just general anesthesia may reflect a mixture of, on the one hand, deliberate decision making in light of patient risk factors and, on the other hand, some haphazard elements such as the usual practices of different anesthesiologists at a single hospital. Unlike random assignment to local or general anesthesia in a clinical trial, both of these nonrandom processes may lead to biased comparisons, but the two processes are nonetheless different, and either could be severely biased when the other is subject to little or no bias. Proximity and affiliation are associated with sociodemographic and other issues that vary from place to place: Austin, Texas is not Houston, and Buffalo, New York is

not Manhattan. Deliberate decision making guided by patient risk factors could severely bias comparisons, particularly when there is accurate and universal agreement about which risk factors matter. In the current study, widely divergent practices at different hospitals suggest that such agreement is less than universal. As is standardly done, adjustments are made for measured covariates, that is for measured pretreatment characteristics of patients. However, if treatments are not randomly assigned to patients, then there is inevitably the concern that adjustments omit some unmeasured covariate and therefore fail to compare comparable patients under competing treatments.

1.2 Outline: Matched pairs unequally susceptible to types of unmeasured bias

We propose a new design for an observational study for use in this and related contexts. The design produces two independent tests of no treatment effect, that is, of no difference in outcomes between general anesthesia alone or some use of regional anesthesia. Moreover, each of these two tests is only slightly affected by unmeasured biases that might strongly affect the other. That is, the design produces two evidence factors (Rosenbaum 2010a, 2011a; Zhang et al. 2011). An optimal matching algorithm yields two nonoverlapping sets of matched pairs. Each pair contains two patients with different treatments — some local anesthesia (say treated) or just general anesthesia (say control) — but with similar covariates, that is, the same operative procedure, similar age, ASA and APACHE scores, similar comorbidities, and so on. In the first set of matched pairs, the 47 hospitals are perfectly balanced: for $h = 1, \ldots, 47$, hospital h is represented equally often in the treated and control groups. In the second set of matched pairs, hospital h is represented in the treated group if it assigns most patients to treatment, and it is represented in the control group if it assigns most patients to control, but no one hospital h is represented in both groups. The first set of matched pairs, the 'finely balanced set,' is not much affected by

unmeasured systematic differences between the types of hospitals that typically use some regional anesthesia and those that rarely use any regional anesthesia; after all, these two types of hospitals are represented equally often in the treated and control groups. The second set of matched pairs, the 'usual practice set,' is less affected by decision making about individual patients within hospitals; after all, the main reason that a patient received regional or general anesthesia in the second type of matched pair is that the patient received the care typically provided in that hospital. The description just given is qualitative, but there is a quantitative dimension also. Hospitals that assign almost all patients to treatment or almost all patients to control dominate the second type of match, the 'usual practice match,' while hospitals that divide their patients fairly evenly dominate the 'finely balanced match,' even though most hospitals contribute at least a few patients to both matched sets.

The case-study is presented first in §2, whereas the technical details of the new matching algorithm are described later in §4. In §2.1, the basic structure of the study is described. Then, §2.2 asks whether the matching has been effective in balancing 44 observed covariates. Of course, the key source of uncertainty in an observational study stems not from imbalances in observed covariates, which can be removed, but rather from possible imbalances in unobserved covariates. Section 2.3 looks at two outcomes and their combination, deep vein thrombosis, a serious complication of knee surgery, and readmission to an acute care hospital within 30 days. Section 2.5 discusses sensitivity to unmeasured biases. In §3, the assumptions underlying various analyses are contrasted, in particular, the assumptions that would lead both types of match to produce consistent estimates, violations that would lead the two matches to disagree with each other, and the relationship with instrumental variable analysis.

1.3 Brief review of the logic of evidence factors

The proposed matched design is intended to produce two independent matched comparisons, each of which is not extremely susceptible to one type of unmeasured bias while being very susceptible to another. Differences between hospitals are largely controlled in the finely balanced match but they are intensified in the usual practice match. Selection biases within hospitals have reduced effect on the usual practice matches — each hospital is doing what it typically does — but they are intensified in the finely balanced match. Why is this structure desirable? Why is it advantageous to enlarge one bias while shrinking another when both biases may be present? The logic of evidence factors is related to the goals of replication in observational studies, which will now be discussed.

Because treatments are not assigned at random, observational studies may vary in their susceptibility to particular unmeasured biases but it is difficult if not impossible to design an observational study that is absolutely immune to all possible unmeasured biases. When a new investigator sets out to replicate a previous observational study, the new investigator has the choice of imitating the first study exactly as reported or, alternatively, of varying the design in specific ways to reduce its susceptibility to one worrisome bias while perhaps producing a study more susceptible to some other type of bias. The attempt to replicate exactly increases the sample size, but many observational studies are large and a limited sample size is not the principal source of uncertainty (Cochran 1965). An exact replicate may also serve as a check on the candor and competence of the first study, which is, alas, sometimes a source of genuine concern. On the other hand, the attempt to remove one source of potential bias in the original study comes closer to addressing the central concern in observational studies, namely bias from nonrandom treatment assignment. As expressed by Mervyn Susser, the investigator seeks "consistency of results in a variety of repeated tests" (Susser 1987, page 88) and "diverse approaches [which] produce similar results"

(Susser 1973, page 148). Replicates that vary key design elements yield statistically independent tests of no treatment effect that are unequally susceptible to specific forms of bias, so consistent results from varied designs can gradually reduce, albeit not eliminate, uncertainty about unmeasured biases (Rosenbaum 2001).

Evidence factors attempt to produce independent replicates with varied design elements within a single study. Evidence factors are independent tests of the null hypothesis of no treatment effect that are especially susceptible to different type of unmeasured biases (see Rosenbaum 2010a, 2011a; Zhang et al. 2011). Unlike previous work, the current manuscript is concerned with the design of observational studies to yield two evidence factors, specifically with a matching algorithm that produces such a design.

2 Example: Knee Surgery with Regional or General Anesthesia

2.1 'Finely balanced' and 'usual practice' matches

Although the Obesity and Surgical Outcomes Study (OBSOS; Silber et al. 2011a,b) abstracted charts for five categories of surgery, the discussion here will focus on the one largest category, namely knee replacements. Table 1 describes the distribution of 4596 Medicare patients undergoing knee surgery in 2298 matched pairs. Each pair contains one patient who had only general anesthesia and one patient who had some regional anesthesia. The pairs were matched for the 44 covariates in Table 2. The pairs are of two types, 1 and 2. The 1354 pairs of type 1 finely balanced the 47 hospitals, which is to say that the number of regional patients from each hospital equals the number of general patients from that hospital; however, individual pairs are closely matched for the covariates in Table 2 and rarely matched for the hospital. In some hospitals, most patients receive regional anesthesia, while in other hospitals most patients receive general anesthesia, but this pattern has been removed from the type 1 pairs. The 944 type 2 pairs, or 'usual practice pairs,'

are built from the remnant of the type 1 match. A hospital which gives most patients regional anesthesia contributes only regional patients to the type 2 match, while a hospital that gives most patients general anesthesia contributes only general patients to the type 2 match.

Consider, for example, hospital 8. Most patients in this hospital receive regional anesthesia during knee surgery. That hospital contributed 18 regional and 18 general patients to the type 1 match and 66 regional patients to the type 2 match. In contrast, in hospital 27 most patients receive general anesthesia. Hospital 27 contributed 35 regional and 35 general patients to the type 1 match and 168 general patients to the type 2 match. Because hospital 3 gives regional anesthesia to about half its patients, it contributes many patients, 202 patients, to the type 1 match and only 4 to the type 2 match. This situation is reversed in hospital 6.

Hospitals that prefer regional anesthesia may differ from those that prefer general anesthesia, but to a large extent unmeasured bias from this source is controlled in the type 1 match. After all, every hospital is represented in the regional and general groups with the same frequency in the type 1 match. However, we do not know how patients were divided between general and regional anesthesia in hospital 3. As will be seen in Table 2, we do know that the regional and general patients were similar in terms of 44 measured covariates from chart abstraction and Medicare claims, and we do know this was true in type 1 matches and in type 2 matches. Nonetheless, it is easy to imagine that in the type 1 matches in hospital 3, the division into regional and general groups involved consideration of some pretreatment risk factors that are not adequately represented by the 44 measured covariates, so the type 1 matches may be biased.

To a considerable extent, the type 2 matches are a comparison of hospitals such as 2 and 6, where most patients receive regional anesthesia, and hospitals 27 and 37, where

most patients receive general anesthesia. In hospitals 6 and 27, the regional-versus-general decision reflects the usual practice in that hospital — the choice to have surgery in hospital 6 or hospital 27 — not a patient-by-patient decision.

In brief, the type 1 matches suffer from one defect — anesthesiologists divided patients in the same hospital into regional and general groups for reasons we do not fully understand — while the type 2 matches suffer from a very different defect — hospitals that typically use regional anesthesia may differ from those that typically use general anesthesia. Although we can and do adjust for measured covariates, we cannot be certain that these adjustments have created a situation free of both defects; however, by contrasting results in type 1 and type 2 pairs, to a large extent we can view the situation without the first defect or without the second. This is less than we might like, but it is progress nonetheless. Possible biases simultaneously affecting both types of pairs are discussed in §2.5.

Treated and control subjects are paired using treatment, covariates and hospitals but not using outcomes, so this is "matching on the basis of covariates and hospitals alone" in the sense of Rosenbaum and Rubin (1985, §1.5), and therefore the matching does not alter conditional distributions of outcomes given treatments, covariates and hospitals.

2.2 Covariate balance attained by matching for 44 covariates

The matching algorithm sought to create matched pairs who were similar in terms of 44 observed covariates. The algorithm did this in two steps: it first created the type 1 pairs finely balanced for the 47 hospitals, then took the unmatched remnants from the first step and constructed the type 2 pairs that reflect the usual practice in the hospitals, either regional or general anesthesia. How successful was the algorithm at balancing the 44 observed covariates? Figures 1 and 2 and Table 2 display covariate balance for 44 covariates describing a patient upon admission or the patient's history. Some of the

information, such as the ASA score and blood pressure, is from chart abstraction, while some information is from the patient's history of Medicare claims.

It is, perhaps, easiest to begin with the four continuous covariates in Figure 1. Figure 1 displays boxplots, four for each covariate, describing the distribution of the covariate in type 1 and type 2 pairs, and for the regional and general patients in those pairs. One hopes to see that the first two boxplots, Gen-1 and Reg-1, are similar, and that the second two boxplots, Gen-2 and Reg-2, are similar; if one saw this, then the marginal distribution of the covariate would be approximately balanced. In fact, more than this has happened: all four boxplots for each covariate are similar, even though this is not the algorithm's goal, nor is it likely to occur in all examples, nor for all covariates in any one example. In Figure 1, age is greater than or equal to 65 by the requirement of eligibility for Medicare and is at most 80 by the design requirements for the OBSOS study. Obesity places a strain on the knees and knee surgery is more common among the obese, so the median body mass index of 30.5 and the typical systolic blood pressure of 142 are high but not surprising.

Table 2 describes all 44 observed covariates. Many of these are binary indicators of other diseases or comorbid conditions that a patient undergoing knee surgery may have, such diabetes or congestive heart failure. The propensity score is the estimated probability of regional anesthesia given the covariates fitted using a logit model; it is a standard tool in multivariate matching (Rosenbaum and Rubin 1983, 1985). Recall that hospitals charts were abstracted for 47 hospitals in Illinois, New York and Texas, and only Medicare data are available for the remaining patients from these states. The risk score is an estimated probability of death within 30 days of surgery based on Medicare data from a logit model fitted to all patients undergoing knee surgery in Illinois, New York and Texas except the patients in the 47 hospitals in Table 1; it is an independent estimate of Hansen's (2008) prognostic score which makes no use of outcomes at the 47 hospitals. Because knee surgery

is typically elective surgery aimed at improving quality of life, it is not surprising that the estimated mortality risk is very low. (The OBSOS study also looked at colectomies and thoracotomies often used in the treatment of cancer, and here the mortality within 30 days is higher.) The ASA score and systolic blood pressure were from chart abstraction and were missing for some patients, and the matching attempts to balance the observed values of these covariates and the binary indicators of missing values, both of which appear in Table 2; see Rosenbaum (2010b, §9.4).

The covariates in Table 2 are not perfectly balanced, but the balance on observed covariates is considerably greater than would have been expected had an equivalent number of patients been assigned completely at random to regional or general anesthesia. This is exhibited in Figure 2. If a two-sample randomization test, such as Wilcoxon's rank sum test or Fisher's exact test for a 2×2 table, is applied to a covariate in a large completely randomized experiment then the distribution of the resulting P-value is approximately uniform on the interval $[0, 1]$; it would be exactly uniform but for the discreteness of randomization distributions. Figure 2 depicts $132 = 3 \times 44$ such P-values, three for each covariate in Table 2, using Wilcoxon's rank sum test for continuous and scored covariates and Fisher's exact test for binary covariates. The open circles in Figure 2 compare general and regional groups in the type 1 matches, while the x's refer to the type 2 matches. The solid circles combine the type 1 and type 2 matches into a single group. None of the 132 P-values in Figure 2 is below 0.05, whereas $6.6 = 132 \times 0.05$ P-values below 0.05 would be expected by chance under complete randomization. Moreover, the entire distribution of P-values is stochastically larger than the uniform distribution, so there is greater balance for observed covariates in Table 2 than expected from complete randomization. Importantly, randomization tends to balance covariates that were not measured, but matching for observed covariates cannot be expected to do this.

An obvious concern about Figures 1 and 2 and Table 2 is that balance is appraised one covariate at a time. It is possible that the marginal distribution of each of the 44 covariates is balanced, yet the joint distribution is not. After all, the match in Table 2 was produced by an algorithm that was aiming for covariate balance, and it is conceivable that an algorithm has done something odd in a high-dimensional sense. We examine multivariate balance using the crossmatch test (Rosenbaum 2005) as suggested by Heller et al. (2011). The crossmatch test momentarily forgets who is treated, who is control and who is matched to whom, and then uses optimal nonbipartite matching (Lu et al. 2011, Derigs 1988, Papadimitriou and Steiglitz 1982, §11.3) to repair the momentarily unpaired groups using the covariates alone. Having done this, the crossmatch test counts the number of times a treated subject was paired with a control, rejecting the hypothesis of covariate balance if that count is small. The test is an exact, distribution-free randomization test of covariate balance. The idea is that if treated subjects are rarely paired with controls using covariates alone, the covariate distributions must differ. See Heller et al. (2011) for discussion of the relationship between the crossmatch test and the propensity score. When applied to the finely balanced pairs, the usual practice pairs and all the pairs, the P-values are 1.00, 0.98 and 0.99, because their are more, rather than fewer, crossmatches than expected under covariate balance. That is, in a multivariate sense also, the observed covariates exhibit greater balance than would be expected from assignment to treatment completely at random.

In brief, the matching algorithm does appear to have balanced the 44 observed covariates, perfectly balanced the 47 hospitals in the type 1 match, and perfectly unbalanced the hospitals in the type 2 match.

2.3 Naïve Analysis: What would one conclude if matching had removed all bias?

In an observational study, a naïve analysis is one that assumes adjustments for observed covariates suffice to remove all bias; see §3. In §2.3, a naïve analysis is presented and then §2.5 examines the degree of sensitivity to unmeasured bias of one of the results in §2.3.

Mortality within 30 days of admission was low, 13 deaths among 4596 patients, with 6 among general anesthesia patients and 7 among regional anesthesia patients. Of these 13 deaths, 9 occurred prior to discharge from the hospital and four occurred shortly after discharge. Two further patients died prior to discharge from the hospital at 51 and 72 days after admission; both received general anesthesia. Albers (1988) proposed a rank test for censored matched pairs. Using this test three times to compare mortality over 180 days (i.e. six months) after admission in type 1, type 2 and all pairs yields three two-sided P-values, the smallest of which is 0.22. In brief, among the few deaths, there is no indication of a difference in survival associated with regional-versus-general anesthesia. An exponential distribution with a hazard rate of $13/4596 = 0.00283$ per month would have an expectation of more than 29 years, far greater than the life expectancy in the Medicare population. Taken together, this suggests events following surgery may or may not have killed a few patients, but despite this their overall 30-day mortality was low compared with the Medicare population, perhaps because elective knee surgery is undertaken by a comparatively healthy subset of the Medicare population.

Although elective knee surgery often proceeds uneventfully, two of the more common but serious events associated with knee surgery are deep vein thrombosis and readmission to an acute-care hospital within 30 days. Table 3 examines deep vein thrombosis, readmission within 30 days, and their combination. Table 3 refers to "alive without readmission" or "alive without deep vein thrombosis," so the small number of deaths are always counted in the unfavorable category. The McNemar-Mantel-Haenszel test for paired binary data

is the large sample approximation to the uniformly most powerful unbiased test against a constant odds ratio not equal to one — see Birch (1964) and Cox (1966) — and this test focuses on discordant pairs, that is, the subset of matched pairs in which exactly one of the two people exhibited the binary response in question. For this reason, Table 3 counts discordant pairs. The null hypothesis that the odds ratio is the same in finely balanced (type 1) pairs and usual practice (type 2) pairs was tested using Gart's (1969) test for interaction for paired binary responses; it is essentially Fisher's exact test for a 2×2 table comparing type 1 and type 2 discordant pairs.

In Table 3, the type 1 and type 2 pairs yield different impressions. In the finely balanced or type 1 pairs, the regional and general groups have the same number of patients drawn from hospital h, for $h = 1, \ldots, 47$, whereas the usual practice or type 2 pairs, one is contrasting patients at hospitals that typically use regional or typically use general anesthesia. In discordant finely balanced pairs, it is more common for the patient receiving regional anesthesia to be alive without deep vein thrombosis or readmission. In the usual practice pairs, it is quite plausible that there is no difference, and the point estimates of odds ratios are closer to the null value of 1. However, Gart's test does not reject the null hypothesis that the finely balanced and ususal practice pairs have the same odds ratio.

Not shown in Table 3 is an analysis of an ordinal 3-category outcome with score 0 for patients who had neither deep vein thrombosis, 1 for patients who either but not both, and 2 for patients who had both or died. Although there were 52 patients with score 2, the analysis is barely distinguishable from the combined analysis in Table 3.

Figure 3 depicts time-to-readmission for up to 30 days after discharge, or more accurately, it depicts "alive without readmission at d days after discharge" for $d = 0, \ldots, 30$. The curves do not begin at 1 on $d = 0$ in part because of deaths prior to discharge, which again are *included* in Figure 3 as *not* "alive without readmission at d days after discharge."

The two-sided P-values in Figure 3 are again from Albers (1988) test, with censoring imposed at 30 days to reflect the notion that a hospital admission in the Medicare population several months after knee surgery may not be connected with the surgery. All censoring occurred at exactly 30 days because all patients were observed for at least 30 days. Generally, the impression from Figure 3 is consistent with the impression from Table 3.

2.4 Testing multiple hypotheses with multiple control groups

When several comparisons are made, as in Table 3, each test performed at level α, there is an increased chance, typically above α, of at least one false rejection simply by chance when all null hypotheses are true. Two simple testing plans control the chance of false rejection. In the first plan, the overall test — "All" in Table 3 — is performed first, and only if this test rejects at level α are the two types of pairs tested separately, each at level α. This plan performs either one test or three tests depending upon the outcome of the first test. In the second plan, priority is given to one type of pair, say the balanced type 1 pairs, which are tested at level α, and only if this P-value is less than or equal to α are the type 2 pairs also tested at level α. The second plan performs either one test or two tests. If each test separately has level α, then by the argument in Rosenbaum (2008; 2010b, §19.3), the chance that either testing plan tests and rejects at least one true hypothesis is at most α.

If the first testing plan were applied to deep vein thrombosis in Table 3, then it would reject no effect for all pairs at the conventional level $\alpha = 0.05$, would therefore test both types of pairs at level 0.05, and would also reject no effect for the finely balanced pairs, but not for the usual practice pairs. For readmission, the first testing plan would not reject for all pairs and would stop, testing no further hypotheses.

2.5 Sensitivity analysis: What magnitude of unmeasured bias would need to be present to alter the conclusions of the naïve analysis?

The naïve analysis in §2.3 acted as if the matching had recreated a randomized pair experiment in which randomization tests, such as McNemar's test and Albers' test could safely be used. The current section performs in Table 4 a sensitivity analysis which asks about the magnitude of departure from random assignment that would need to be present to alter these conclusions. Because McNemar's and Albers' statistics equal a constant plus a so-called "sign-score statistic," the sensitivity analysis of I pairs is a special case of an available technique (Rosenbaum 1987; 2010b, §3.4), so the method will be described very briefly. As discussed in Rosenbaum (2010a, 2011a) and Zhang et al. (2011), the two sensitivity analyses for the two types of pairs may be combined using Fisher's method for combining independent P-values, and this is done in Table 4. For a few alternative methods of sensitivity analysis, see Cornfield et al. (1959), Yanagawa (1984), Marcus (1997), Lin, Psaty and Kronmal (1998), Robins, Rotnitzky and Scharfstein (1999), Imbens (2003), Yu and Gastwirth (2005) and Small (2007).

The sensitivity analysis is indexed by a parameter $\Gamma \geq 1$. The sensitivity parameter Γ says that two individuals with the same observed covariates may differ in their odds of treatment — here, regional anesthesia — by at most a factor of Γ. If $\Gamma = 1$, then two individuals with the same observed covariates have the same odds of treatment, and this results in the randomization distribution or the naïve analysis in §2.3. For each $\Gamma > 0$, the distribution of treatment assignments is unknown but to a bounded degree, so it is possible to determine upper and lower bounds on inference quantities such as P-values or point estimates or endpoints of confidence intervals. Table 4 presents upper bounds on one-sided P-values. The test statistics are of the form $\sum q_i$, where q_i is a score for pair i, so that, under Fisher's sharp null hypothesis of no treatment effect, pair i contributes

$\pm q_i$ and in a randomized paired experiment $\pm q_i$ occur with equal probabilities 1/2. Under Fisher's sharp null hypothesis, McNemar's test has q_i equal to 1, 0 or -1 with $q_i = 0$ for a concordant pair, while for Albers' (1988) test for censored paired times, q_i is a score computed from the time and censoring distributions. For either test, allowing for a bias of $\Gamma \geq 1$ in treatment assignment, the approximate upper bound on the one-sided P-value is:

$$1 - \Phi\left[\frac{\sum q_i - \left\{(\Gamma-1)/(\Gamma+1)\right\}\sum |q_i|}{\sqrt{\left\{4\Gamma/(\Gamma+1)^2\right\}\sum q_i^2}}\right], \tag{1}$$

where $\Phi(\cdot)$ is the Normal cumulative distribution; see Rosenbaum (1987; 2010b, §3.4).

The value of Γ indexes the magnitude of departure from random assignment. It is sometimes useful to reinterpret a magnitude of Γ in terms of failure to adjust for a single unobserved covariate associated with both treatment assignment and matched pair outcome difference. Under mild conditions (Rosenbaum and Silber 2009, Proposition 1), a single value of Γ corresponds with a curve of values of two parameters, namely $\Gamma = (\Delta\Lambda + 1)/(\Delta + \Lambda)$ where Λ controls the odds of assignment to treatment and Δ controls the odds of a positive response difference, and $\Lambda \to \Gamma$ as $\Delta \to \infty$. For instance, the curve for $\Gamma = 1.25$ includes $(\Delta = 2, \Lambda = 2)$ for an unobserved covariate that doubles the odds of treatment and doubles the odds of a positive response difference. In parallel, the curve for $\Gamma = 1.1$ includes $(\Delta = 2, \Lambda = 4/3)$, $(\Delta = 4/3, \Lambda = 2)$, $(\Delta = 1.56, \Lambda = 1.56)$, and infinitely many other values of (Δ, Λ) satisfying $1.1 = (\Delta\Lambda + 1)/(\Delta + \Lambda)$. See Rosenbaum and Silber (2009) for detailed discussion.

Table 4 displays the upper bounds (1) on the one-sided P-values for a binary outcome using McNemar's test and for a paired censored time using Albers' test. Success for the combined binary outcome is to be alive without deep vein thrombosis and without readmission at 30 days. Success for the temporal outcome is to be alive without readmission

at time t days for $t \in [0, 30]$. In Table 4, Γ_1 refers to biases affecting the type 1 pairs which balanced the 47 hospitals, and Γ_2 refers to biases affecting the type 2 pairs which contrast outcomes in hospitals with differing usual practices. For $\Gamma_1 = \Gamma_2$, the sensitivity analysis using all $I = 2298$ pairs is given in the last column. Because type 1 and type 2 pairs do not overlap, the tests based on them are independent, and the upper bounds on their P-values may be combined using Fisher's method which adds -2 times the logs of the two P-values and compares the result with the chi-square distribution on 4 degrees of freedom; see Rosenbaum (2010a, 2011a) for discussion of using Fisher's method in sensitivity analyses. The table considers only $\Gamma_2 = 1$ because no effect is plausible in the type 2 pairs even in a randomization test that assumes there is no unmeasured bias.

The combined binary outcome would be significant for type 1 pairs and for the combined analysis in the absence of biased treatment assignment, $\Gamma_1 = \Gamma_2 = 1$, and those analyses resist a small unmeasured bias of $\Gamma_1 = 1.1$, $\Gamma_2 = 1$, but even slightly larger biases could explain the somewhat better results for patients receiving regional anesthesia. In a two-sided test, the time-to-readmission outcome is only barely significant only for type 1 pairs even in the absence of bias $\Gamma_1 = \Gamma_2 = 1$. The multiple testing procedure of §2.4 may be combined with the sensitivity analysis, and at level $\alpha = 0.05$ it would terminate testing in Table 4 at $\Gamma_1 = 1.1$, $\Gamma_2 = 1$; see Rosenbaum (2010b, §19).

2.6 Is logit regression an alternative analytic strategy?

Our sense is that the analyses just described cannot easily be done using a model such as logit regression. A key element in the analysis just described is that there are two independent analyses, one largely unaffected by differences between hospitals, the other comparing different hospitals with different typical patterns of anesthesia practice. A logit regression would include the patient covariates in Table 2 and might or might not include

the 47 hospital indicators in Table 1. Those two logit analyses, with and without hospital indicators, would use the same patients twice, so they would be far from independent replicates. Moreover, both analyses would derive some, perhaps most, of their information about the association of an outcome with regional-versus-general anesthesia from patients in the same hospital, where selection biases may assign one patient to regional, another to general. For instance, in Table 1, hospital 3 uses both types of anesthesia for knee surgery with about equal frequencies while hospital 6 typically uses regional anesthesia. Logit regression would not clarify the distinct types of information provided by hospitals 3 and 6. The comparison within hospital 3 is quite a different thing from the comparison of hospital 6 with hospital 27, yet logit regression would not aid in distinguishing them.

At a purely practical level, there are 44 covariates and 47 hospitals, or $44 + 47 = 91$ predictors of just 172 cases of deep vein thrombosis, so even a logit model without interactions would be thinly supported by the data. Indeed, in the usual practice pairs, there are only 62 cases of deep-vein thrombosis but there are still 44 covariates. In contrast, matching balanced the 91 observed covariates to a greater degree than randomization is expected to balance observed covariates, and it sought individually similar patients using a Mahalanobis distance, so it is attempting to control interactions among predictors.

2.7 Discussion of the example

Selection bias refers to the thoughtful, considered selection of some patients for treatment, others for control; it is one mechanism that can lead to biased comparisons in an observational study. A worry focused on selection biases within hospitals might lead to a comparison of different hospitals that typically prefer either regional or general anesthesia. Alas, two hospitals that differ in terms of typical anesthesia practice may differ also in other ways, such as other aspects of the care they provide and the populations they serve.

In §2.3, two sets of matched pairs were examined, one balanced for the 47 hospitals and hence largely controlling differences among hospitals, the other emphasizing comparisons of different hospitals with different usual anesthesia practices. Both sets of matched pairs controlled biases in measured covariates using matching; see Table 2 and Figures 1 and 2. In each set of matched pairs, one form of unmeasured bias is reduced while the other is enlarged. If one were viewing without bias an effect actually caused by regional anesthesia, then one would hope to see evidence of this in both matched sets.

How strong is the evidence in Tables 3 and 4 and Figure 3 that regional anesthesia reduces the risk of readmission or deep vein thrombosis? Clearly, the proposed design might have produced evidence that is considerably stronger or considerably weaker than the evidence it did produce. We might have seen, but we did not see, an association between regional anesthesia and the outcomes in both the balanced and usual practice pairs. There was, however, no compelling indication of an association in the usual practice pairs, albeit with no compelling indication that the two type of pairs were actually different; see the test for interaction in Table 3. In other words, looking inside hospitals that use both forms of anesthesia, we see somewhat better results with regional anesthesia, but comparing hospitals that mostly do one or mostly do the other, we see little evidence of better results at hospitals that mostly use regional anesthesia. Moreover, we might have seen, but we did not see, an association that was insensitive to moderately large unmeasured biases, whereas the combined analysis with 2298 pairs at the right in Table 4 is sensitive to biases $\Gamma > 1.1$, that is, to biases that might be produced by an unobserved covariate that increases the odds of both regional anesthesia and a favorable outcome by 50%, $(\Delta = 1.56, \Lambda = 1.56)$. In brief, stronger evidence in both senses was possible with this design. At the same time, the analysis on the left for the finely balanced pairs in Table 3 and Figure 3 would meet the usual standard in the typical empirical journal, with conventional tests that yield

significant P-values and point estimates of odds ratios that are far from inconsequential in size. This evidence meets the standards usually set for studies that guide the treatment of patients, but the evidence is not overwhelmingly strong; that is, in our judgement, the evidence is interesting but not decisive. In helping us reach this cautious judgment, our sense is that the proposed design improved upon the usual design and analysis that omits evidence factors and sensitivity analyses.

3 Assumptions for matching within and between institutions and for instrumental variables

The current section considers the relationship between the analysis in §2 and two other analyses, namely an analysis that matched for hospital H and one that used hospital H as an instrument for anesthesia type Z. Following Neyman (1923), Welch (1937), Rubin (1974) and Reiter (2000), write r_T for the outcome a patient would exhibit if given the treatment, here regional anesthesia, and r_C for the response of this same patient if given the control, here general anesthesia, and write $Z = 1$ if the patient does receive the treatment or $Z = 0$ if the patient does receive the control, so the response actually exhibited by the patient is $R = Z\, r_T + (1 - Z)\, r_C$, and the causal effect of the treatment on this patient, namely $r_T - r_C$, is not seen for any patient. Fisher's sharp null hypothesis of no effect asserts $r_T = r_C$ for all patients, so that, for instance, changing anesthesia type does not change whether a patient has deep vein thrombosis. Patients exhibit an observed covariate $\mathbf{x}$ but may also differ in terms of a relevant unobserved covariate u. Each patient is treated in some hospital H. Finally, following Dawid (1979), write $A \perp\!\!\!\perp B \,\big|\, C$ for A is conditionally independent of B given C.

The condition

$$(r_T,\, r_C) \;\underline{||}\; (Z,\, H) \;\Big|\; \mathbf{x} \;\text{ with }\; \Pr(\, Z = z,\, H = h \mid \mathbf{x}) > 0 \text{ for all } z,\, h, \tag{2}$$

says that the effects of regional-versus-general anesthesia on health outcomes $(r_T,\, r_C)$ are biased neither by selection into particular hospitals H nor by selection for particular forms of anesthesia Z providing adjustments are made for measured covariates $\mathbf{x}$, a condition similar to strong ignorability as defined in Rosenbaum and Rubin (1983). If (2) were true, then one could estimate an average treatment effect, such as $\mathrm{E}\,(r_T - r_C|\,\mathbf{x})$ or $\mathrm{E}\,(r_T - r_C)$, by comparing the observed responses R of treated and control subjects at any hospitals h, h' adjusting for $\mathbf{x}$, because (2) implies $\mathrm{E}\,(r_T - r_C|\,\mathbf{x}) = \mathrm{E}\,(R|\,Z = 1, H = h, \mathbf{x}) - \mathrm{E}\,(R|\,Z = 0, H = h', \mathbf{x})$ for all h, h'. That is, if (2) were true, both the type 1 balanced matches and the type 2 usual practice matches would provide consistent estimates of average treatment effects $\mathrm{E}\,(r_T - r_C|\,\mathbf{x})$ at $\mathbf{x}$. If the two types of matches produce very different estimates of $\mathrm{E}\,(r_T - r_C|\,\mathbf{x})$, this suggests that condition (2) is false.

In parallel, if Fisher's sharp null hypothesis of no effect were true, so $R = r_T = r_C$, then (2) would imply $R \;\underline{||}\; (Z,\, H) \;\Big|\; \mathbf{x}$. Therefore, if (2) were true, Fisher's null hypothesis could be tested by testing the hypothesis of conditional independence of observed response R and treatment Z given either $\mathbf{x}$ or $(H, \mathbf{x})$. In particular, this test could be performed with pairs matched using $(Z, H, \mathbf{x})$.

Condition (2) is related to two other conditions which underly certain analyses. By Dawid's (1979) Lemma 4, condition (2) is equivalent to the conjunction of conditions (3) and (4):

$$(r_T,\, r_C) \;\underline{||}\; Z \;\Big|\; (H,\, \mathbf{x}) \quad \text{with }\; \Pr(\, Z = z \mid\, H = h,\, \mathbf{x}) > 0 \text{ for all } z,\, h, \tag{3}$$

$$(r_T,\, r_C) \;\underline{||}\; H \;\Big|\; \mathbf{x} \;\text{ with }\; \Pr(\, H = h \mid\, \mathbf{x}) > 0 \text{ for all } h. \tag{4}$$

Here, after adjustment for $\mathbf{x}$, condition (3) says there is no selection bias in treatment assignment Z within each hospital H, while condition (4) says that there is no selection bias in assigning patients to hospitals H. If (3) were true, an analysis that matched patients for $\mathbf{x}$ within the same hospital could estimate $\mathrm{E}(r_T - r_C|\, H = h,\, \mathbf{x}) = \mathrm{E}(R|\, Z = 1, H = h, \mathbf{x}) - \mathrm{E}(R|\, Z = 0, H = h, \mathbf{x})$ even when (2) is false, and could then estimate $\mathrm{E}(r_T - r_C|\, \mathbf{x})$ by direct adjustment as $\mathrm{E}(r_T - r_C|\, \mathbf{x}) = \sum_h \mathrm{E}(r_T - r_C|\, H = h,\, \mathbf{x})\, \Pr(H = h|\, \mathbf{x})$ where $\Pr(H = h|\, \mathbf{x})$ is directly estimable. Moreover, pairs matched for $\mathbf{x}$ within hospitals could be used to test the null hypothesis of no treatment effect if (3) is true even if (2) is false. Aside from some relatively minor detail, the type 1 pairs could be used if (3) were true, whether or not (2) is true. (The minor detail has to do with balancing H and then ignoring it, rather than matching exactly for H, but this makes the comparison slightly conservative; see Hollander et al. (1974) with their $\lambda_1 = \lambda_2 = 0$.)

In contrast, at each $\mathbf{x}$, condition (4) is one of the several assumptions underlying an instrumental variable analysis that views the hospital H as an instrument for the type of anesthesia received; see Angrist, Imbens and Rubin (henceforth AIR, 1996, Assumption 2). With additional assumptions (AIR 1996, Assumption 1-5), an instrumental variable analysis can estimate a treatment effect parameter when (4) is true but there is selection bias within hospitals, so (3) is false. Moreover, an attempt to strengthen the instrumental variable (Baiocchi et al. 2010) would focus attention on hospitals with $\Pr(Z = 1|\, H = h,\, \mathbf{x})$ near 0 or 1, such as $H = 2$, 6, 27, and 37 in Table 1, which are the hospitals that dominate the type 2, usual practice matches. In the analysis in §2.3, the type 2 matches are not entirely free of selection biases within hospitals — that is, violations of (3) — but they would be if one could use only hospitals h with $\Pr(Z = 1|\, H = h,\, \mathbf{x}) = 0$ or $\Pr(Z = 1|\, H = h,\, \mathbf{x}) = 1$. The sensitivity analysis in §2.5 assumes (2), (3), and (4) are all false with the magnitude of the violation of these conditions controlled by Γ.

4 Matching algorithms

4.1 Review of the optimal assignment algorithm

The remainder of the paper concerns the new algorithm used to create the matched comparison, an algorithm that extends two existing techniques, namely optimal matching with fine balance (Rosenbaum, Ross and Silber 2007) and optimal subset matching (Rosenbaum 2011b). There is a finite set of T treated subjects, $\mathcal{T}$, and a finite set of C potential controls, $\mathcal{C}$, with $\mathcal{T} \cap \mathcal{C} = \emptyset$, and based on covariates there is a distance $0 \leq \delta_{\tau,\gamma} < \infty$ between each $\tau \in \mathcal{T}$ and $\gamma \in \mathcal{C}$. For a finite set, $\mathcal{S}$, the number of elements in $\mathcal{S}$ is $|\mathcal{S}|$, so $|\mathcal{T}| = T$. Write $\mathbf{\Delta}$ for the $T \times C$ matrix of $\delta_{\tau,\gamma}$ whose rows are indexed by $\tau \in \mathcal{T}$ and whose columns are indexed by $\gamma \in \mathcal{C}$. If $C \geq T$, then an assignment α pairs each τ with a different γ; that is, $\alpha : \mathcal{T} \rightarrow \mathcal{C}$ with $\alpha(\tau) \neq \alpha(\tau')$ if $\tau \neq \tau'$. An optimal assignment is an assignment α that minimizes the total distance within pairs, $\sum_{\tau \in \mathcal{T}} \delta_{\tau,\alpha(\tau)}$. Kuhn (1955) and Bertsekas (1981) proposed two solutions of the optimal assignment problem, and Bertsekas' solution is available in `R` as the `pairmatch` function in Hansen's (2007) `optmatch` package. Some solutions to the optimal assignment problem can be produced in $O\left(C^3\right)$ arithmetic operations which is the same order as multiplying two $C \times C$ matrices in the conventional way. For a textbook discussion of optimal assignment algorithms, see Papadimitriou and Steiglitz (1982, §11) and Bertsimas and Tsitsiklis (1997, §7.8). Use of optimal assignment for matching in observational studies is discussed in Rosenbaum (1989).

4.2 Notation and definitions: fine balance, matching a subset

There are I institutions, $i = 1, \ldots, I$, with T_i treated subjects, $\mathcal{T}_i = \{\tau_{i1}, \ldots, \tau_{iT_i}\}$, and C_i potential controls, $\mathcal{C}_i = \left\{\gamma_{i1}, \ldots, \gamma_{iC_i}\right\}$, from institution i. Institutions and treated and control groups are disjoint; that is, $\mathcal{T}_i \cap \mathcal{T}_j = \emptyset$ and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ for $i \neq j$ and $\mathcal{T}_i \cap \mathcal{C}_j = \emptyset$ for all i, j. Write $\mathcal{T} = \mathcal{T}_1 \cup \cdots \cup \mathcal{T}_I$ and $\mathcal{C} = \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_I$, and $T = \sum T_i = |\mathcal{T}|$, $C = \sum C_i = |\mathcal{C}|$.

Also, write $\iota(E)$ for the indicator of event E, so $\iota(E) = 1$ if E occurs and $\iota(E) = 0$ otherwise, and define $\infty \times \iota(E) = \infty$ if E occurs and $\infty \times \iota(E) = 0$ if E does not occur.

A match indicates which treated subjects are matched and to which controls they are matched. That is, a match is a subset $\mathcal{T}_r \subseteq \mathcal{T}$ and a function $\mu : \mathcal{T}_r \rightarrow \mathcal{C}$ such that $\mu(\tau) \neq \mu(\tau')$ if $\tau \neq \tau'$, so the match is $(\mathcal{T}_r, \mu)$. The total distance $\zeta(\mathcal{T}_r, \mu)$ associated with a match $(\mathcal{T}_r, \mu)$ is $\zeta(\mathcal{T}_r, \mu) = \sum_{\tau \in \mathcal{T}_r} \delta_{\tau,\mu(\tau)}$. The size of a match is $|\mathcal{T}_r|$.

A match $(\mathcal{T}_r, \mu)$ is finely balanced if the number of treated subjects from institution i equals the number of controls from institution i, even if treated subjects from institution i are not matched to controls from institution i, that is, if

$$\sum_{\tau \in \mathcal{T}_r} \iota(\tau \in \mathcal{T}_i) = \sum_{\tau \in \mathcal{T}_r} \iota\{\mu(\tau) \in \mathcal{C}_i\} \text{ for } i = 1, \ldots, I. \tag{5}$$

Equation (5) extends the definition in Rosenbaum, Ross and Silber (2007) which required all treated subjects be matched, $\mathcal{T}_r = \mathcal{T}$, and this in turn required $C_i \geq T_i$, $i = 1, \ldots, I$, so fine balance in this earlier sense may not be possible. In the current study, in some hospitals most patients receive regional anesthesia, $T_i > C_i$, whereas in others most patients do not, $C_i > T_i$, so $\mathcal{T}_r = \mathcal{T}$ is not possible with pair matching.

In optimal subset matching (Rosenbaum 2011b), a particular distance $\widetilde{\delta}$ is selected with the view that it would be preferable to not match some treated subjects than to match them at a distance greater than $\widetilde{\delta}$. A slight subtlety must be addressed because the matching decisions are interdependent. If two finely balanced matches, $(\mathcal{T}_r, \mu)$ and $(\mathcal{T}_r^*, \mu^*)$, have the same size, $|\mathcal{T}_r| = |\mathcal{T}_r^*|$, then $(\mathcal{T}_r, \mu)$ is better if it has smaller distance, $\sum_{\tau \in \mathcal{T}_r} \delta_{\tau,\mu(\tau)} < \sum_{\tau \in \mathcal{T}_r^*} \delta_{\tau,\mu^*(\tau)}$. Fix a number $\widetilde{\delta} \geq \min\left(\delta_{\tau_{it},\gamma_{i'c}}\right)$. As in Rosenbaum (2011b), if $|\mathcal{T}_r| < |\mathcal{T}_r^*|$, we prefer $(\mathcal{T}_r, \mu)$ to $(\mathcal{T}_r^*, \mu^*)$, written $(\mathcal{T}_r, \mu) \succ (\mathcal{T}_r^*, \mu^*)$ if

$$\frac{\sum_{\tau \in \mathcal{T}_r^*} \delta_{\tau,\mu^*(\tau)} - \sum_{\tau \in \mathcal{T}_r} \delta_{\tau,\mu(\tau)}}{|\mathcal{T}_r^*| - |\mathcal{T}_r|} > \widetilde{\delta}, \tag{6}$$

we prefer $(\mathcal{T}_r^*, \mu^*)$ to $(\mathcal{T}_r, \mu)$, written $(\mathcal{T}_r^*, \mu^*) \succ (\mathcal{T}_r, \mu)$, if the inequality in (6) is reversed, and we are indifferent, written $(\mathcal{T}_r^*, \mu^*) \sim (\mathcal{T}_r, \mu)$, if there is equality in (6), so we prefer to use more treated subjects rather than fewer treated subjects provided they can be included at an average change in total distance of less than $\widetilde{\delta}$. Also write $(\mathcal{T}_r, \mu) \succsim (\mathcal{T}_r^*, \mu^*)$ if either $(\mathcal{T}_r, \mu) \succ (\mathcal{T}_r^*, \mu^*)$ or $(\mathcal{T}_r^*, \mu^*) \sim (\mathcal{T}_r, \mu)$.

4.3 A minimum distance finely balanced optimal subset match

Write $m_i = \min(T_i, C_i)$, $\overline{m}_i = \max(T_i, C_i)$, $M = \sum m_i$, and $\overline{M} = \sum \overline{m}_i$. Fine balance requires that at least $T_i - m_i$ treated subjects and at least $C_i - m_i$ controls be removed, so the maximum size $|\mathcal{T}_r|$ of a finely balanced match is M. Augment $\mathbf{\Delta}$ to form the square $\overline{M} \times \overline{M}$ matrix $\mathbf{\Lambda}$ with entries $\lambda_{k\ell}$ which has $\sum_{i=1}^{I}(C_i - m_i)$ additional rows labeled ε_{ik}, $k = 1, \ldots, C_i - m_i$, $i = 1, \ldots, I$, and $\sum_{i=1}^{I}(T_i - m_i)$ additional columns labeled $\epsilon_{i\ell}$, $\ell = 1, \ldots, T_i - m_i$, $i = 1, \ldots, I$ defined in the following way:

$$
\begin{aligned}
\lambda_{\tau\gamma} &= \delta_{\tau\gamma} \text{ for } \tau \in \mathcal{T} \text{ and } \gamma \in \mathcal{C}, \\
\lambda_{\tau,\epsilon_{i\ell}} &= \infty \times \iota(\tau \notin \mathcal{T}_i),\ i = 1, \ldots, I,\ \ell = 1, \ldots, T_i - m_i, \\
\lambda_{\varepsilon_{ik},\gamma} &= \infty \times \iota(\gamma \notin \mathcal{C}_i),\ i = 1, \ldots, I,\ k = 1, \ldots, C_i - m_i, \\
\lambda_{\varepsilon_{ik},\epsilon_{i'\ell}} &= \infty,\ i, i' = 1, \ldots, I,\ k = 1, \ldots, C_i - m_i,\ \ell = 1, \ldots, T_{i'} - m_{i'}.
\end{aligned}
$$

Table 5 is a small illustration. There are $I = 3$ institutions, with $\mathcal{T}_1 = \{\tau_{11}\}$, $\mathcal{T}_2 = \{\tau_{21}, \tau_{22}\}$, $\mathcal{T}_3 = \{\tau_{31}, \tau_{32}\}$, $\mathcal{C}_1 = \{\gamma_{11}, \gamma_{12}\}$, $\mathcal{C}_2 = \{\gamma_{21}\}$, $\mathcal{C}_3 = \{\gamma_{31}, \gamma_{32}\}$, and $\mathbf{\Delta}$ is 5×5. Therefore, $m_1 = 1$, $m_2 = 1$, $m_3 = 2$, $M = 1 + 1 + 2 = 4$ and $\overline{M} = 2 + 2 + 2 = 6$, so one row, namely ε_{11} is added, and one column, namely ϵ_{21}, is added. The patterns of 0's and ∞'s in ε_{11} will force either γ_{11} or γ_{12} to be paired with ε_{11}, while either τ_{21} or τ_{22} is paired with

ϵ_{21}. As shown in Proposition 1, an optimal assignment in $\mathbf{\Lambda}$ yields a minimum distance finely balanced match of maximum size M once pairs involving ε's or ϵ's are removed.

Optimal subset matching combined with fine balance entails possibly using fewer than M pairs, guided by the preference (6), in such a way that fine balance is preserved. For each i, fix a number $\widetilde{m}_i$ with $0 \leq \widetilde{m}_i \leq m_i$, where the algorithm will require at least $\widetilde{m}_i$ treated subjects and $\widetilde{m}_i$ controls from institution i; however, based on (6), the algorithm may use more than $\widetilde{m}_i$ treated subjects and $\widetilde{m}_i$ controls from institution i. Augment $\mathbf{\Lambda}$ to form $\mathbf{\Upsilon}$ by appending $m_i - \widetilde{m}_i$ rows labeled $\varkappa_{ik}$ and columns labeled $\kappa_{i\ell}$, so that $\mathbf{\Upsilon}$ is square with $\overline{M} + \sum (m_i - \widetilde{m}_i)$ rows and columns, where column γ of $\varkappa_{ik}$ has entry $\infty \times \iota(\gamma \notin \mathcal{C}_i) + \widetilde{\delta}\,\iota(\gamma \in \mathcal{C}_i)/2$ and row τ of $\kappa_{i\ell}$ has entry $\infty \times \iota(\tau \notin \mathcal{T}_i) + \widetilde{\delta}\,\iota(\tau \in \mathcal{T}_i)/2$, $i = 1, \ldots, I$, $\ell = 1, \ldots, m_i - \widetilde{m}_i$, the one diagonal entry linking $\varkappa_{ik}$ and κ_{ik} is zero, and all other entries in both $\varkappa_{ik}$ and $\kappa_{i\ell}$ are ∞. If $\widetilde{m}_i = m_i$ for all i, there is no augmentation and $\mathbf{\Upsilon} = \mathbf{\Lambda}$. Table 5 is an example with $\widetilde{m}_1 = m_1$, $\widetilde{m}_2 = m_2$ and $\widetilde{m}_3 = 0$ so that $m_3 - \widetilde{m}_3 = 2$ rows and columns are added to $\mathbf{\Lambda}$ to form $\mathbf{\Upsilon}$.

Proposition 1 *Let α be a minimum distance assignment in $\mathbf{\Upsilon}$. Define $\mathcal{T}_r = \{\tau \in \mathcal{T} : \alpha(\tau) \in \mathcal{C}\}$ and let μ be the restriction of α to the domain $\mathcal{T}_r$. Then the match $(\mathcal{T}_r, \mu)$ is finely balanced with at least $\widetilde{m}_i$ treated subjects and $\widetilde{m}_i$ controls from institution i,*

$$\sum_{\tau \in \mathcal{T}_r} \iota(\tau \in \mathcal{T}_i) = \sum_{\tau \in \mathcal{T}_r} \iota\{\mu(\tau) \in \mathcal{C}_i\} \geq \widetilde{m}_i \text{ for } i = 1, \ldots, I. \tag{7}$$

Moreover, if $(\mathcal{T}_r^*, \mu^*)$ is any other finely balanced match with at least $\widetilde{m}_i$ treated subjects and $\widetilde{m}_i$ controls from institution i, then

$$\text{if } |\mathcal{T}_r| = |\mathcal{T}_r^*| \text{ then } \sum_{\tau \in \mathcal{T}_r^*} \delta_{\tau,\mu^*(\tau)} \geq \sum_{\tau \in \mathcal{T}_r} \delta_{\tau,\mu(\tau)} \tag{8}$$

and

$$\textit{if } |\mathcal{T}_r| \neq |\mathcal{T}_r^*| \textit{ then } (\mathcal{T}_r, \mu) \gtrsim (\mathcal{T}_r^*, \mu^*)\,. \tag{9}$$

Proof. First, we show that there exists an assignment α' in $\boldsymbol{\Upsilon}$ with finite total distance. To construct such an α', assign $\varkappa_{ik}$ to κ_{ik} for all ik with a distance of 0, assign each ε_{ik} to a different γ_{ij} from the same institution i with a distance of 0, and assign each ϵ_{ik} to a different τ_{ij} from the same institution i with a distance of 0 — this is always possible by the definition of m_i — and complete the construction of α' by arbitrarily pairing the unpaired τ's and γ's, yielding an assignment with finite total distance. Because there exists an assignment α' in $\boldsymbol{\Upsilon}$ with finite total distance, an optimal assignment α must also have finite total distance. Let α^* be *any* assignment in $\boldsymbol{\Upsilon}$ with finite total distance, let $\mathcal{T}_r^* = \{\tau \in \mathcal{T} : \alpha^*(\tau) \in \mathcal{C}\}$ and let μ^* be the restriction of α^* to $\mathcal{T}_r^*$. Now α^* has finite total distance if and only if it avoids all of the ∞'s in $\boldsymbol{\Upsilon}$. In particular, α^* must pair each ε_{ik} to a different γ_{ij} at a distance of 0, thereby removing $C_i - m_i$ of the γ_{ij} from institution i. In parallel, α^* must pair each ϵ_{ik} to a different τ_{ij} at a distance of 0 thereby removing $T_i - m_i$ of the τ_{ij} from institution i. Also, for each ik, if α^* does not pair $\varkappa_{ik}$ to κ_{ik} with a distance of 0, then $\varkappa_{ik}$ must be paired to some γ_{ij} from institution i and κ_{ik} must be paired to some τ_{ij} from institution i with a total distance for these two pairs of $\widetilde{\delta}/2 + \widetilde{\delta}/2 = \widetilde{\delta}$. If follows that α^* has

$$m_i \geq \sum_{\tau \in \mathcal{T}_r^*} \iota\left(\tau \in \mathcal{T}_i\right) = \sum_{\tau \in \mathcal{T}_r^*} \iota\left\{\mu^*(\tau) \in \mathcal{C}_i\right\} \geq \widetilde{m}_i \text{ for } i = 1, \ldots, I, \tag{10}$$

so it is finely balanced, and α^* has a total distance of $\mathcal{D}(\alpha^*) = \sum_{\tau \in \mathcal{T}_r^*} \delta_{\tau,\alpha^*(\tau)} + (M - |\mathcal{T}_r^*|)\,\widetilde{\delta}$. So every assignment α^* with finite total distance is finely balanced (10); moreover, any match μ^* that satisfies (10) may be extended to an assignment α^* with finite total distance. If α^* has finite total distance and α is a minimum distance assignment, then

$\mathcal{D}(\alpha) \leq \mathcal{D}(\alpha^*) < \infty$, so if $|\mathcal{T}_r| = |\mathcal{T}_r^*|$ then (8) holds, whereas otherwise rearranging $\mathcal{D}(\alpha^*) - \mathcal{D}(\alpha) \geq 0$ yields (9). ■

4.4 Construction of two sets of matched pairs

The match in §2 first applied Proposition 1 to a distance matrix formed using a robust Mahalanobis distance with calipers on the propensity score; see Rosenbaum (2010b, §8) for discussion of these standard devices. Some patients were not matched in this first step. In the second step, hospitals were classified as either "predominantly general" or "predominantly regional," and unmatched patients from their predominant group were candidates for matching in the second step. The second step used optimal subset matching (without fine balance), as discussed in Rosenbaum (2011b). Several values of $\widetilde{\delta}$ were tried until an acceptable match was obtained. A good starting value for $\widetilde{\delta}$ is the lower 5% or 10% quantile of all of the distances in $\mathbf{\Delta}$. Matching makes no use of outcomes, so it is part of the design of the study.

5 Discussion

The design in §2 may be used in other contexts. The institutions need not be hospitals; instead, they might be schools, states, courts, judges, and so on. Indeed, the groups need not be defined by institutions that provide treatment. For instance, sometimes a new treatment replaces an older treatment gradually over a period of years. The process that selects individuals for the new treatment may be different in the early years, when the new treatment is something novel, compared to late years, when the older treatment has become a rarity. In this case, year of treatment may take the place of the institutions. One match balances the year, while the other compares early years to late years. Unlike the comparison that balances year of treatment, the comparison of early years to late years

is confounded with time but compares usual practice in early and late years.

The new matching algorithm defined by Proposition 1 may be used on its own to construct a single set of matched pairs similar to the type 1 matched pairs. For instance, institutions might be replaced by all combinations or interactions of several nominal variables, and the algorithm would then create one set of matched pairs to balance the combinations of these nominal variables. As originally developed, fine balance required all treated subjects to be matched, $\mathcal{T}_r = \mathcal{T}$ in (5), so it was a constraint on an optimization problem, but the constraint was not feasible if some $C_i < T_i$. The algorithm in Proposition 1 combines matching with fine balance and optimal subset matching so that fine balance (5) is always feasible, and it does this by permitting the matched treated subjects to be a proper subset of all treated subjects, $\mathcal{T}_r \subset \mathcal{T}$. An alternative approach to maintaining feasibility is to require all treated subjects to be matched, $\mathcal{T}_r = \mathcal{T}$, but to permit slight deviations from fine balance (5); see Yang et al. (2012) and Yang's `finebalance` package in `R`.

In selecting multiple control groups, a general principle is Bitterman's "control by systematic variation," that is, select control groups likely to be differently influenced by unmeasured biases; see Campbell (1969) and Rosenbaum (2002, §8; 2010b, §5.2.2, §11.3.1). Typically, such control groups exist as groups prior to matching. In contrast, the current study created two types of controls, balanced or usual practice, from a single population in which these two types were not previously distinguished. The example in Heller et al. (2010) illustrates a different method for building two types of controls differently influenced by unmeasured biases; specifically, that method uses tapered matching applied a seemingly innocuous covariate.

References

Albers, W. (1988), "Combined rank tests for randomly censored paired data," *Journal of the American Statistical Association*, 83, 1159-1162.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996), "Identification of causal effects using instrumental variables (with Discussion)," *Journal of the American Statistical Association*, 91, 444-455.

Baiocchi, M., Small, D. S., Lorch, S. and Rosenbaum, P. R. (2010), "Building a stronger instrument in an observational study of perinatal care for premature infants," *Journal of the American Statistical Association*, 105, 1285-1296.

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

Bertsimas, D. and Tsitsiklis, J. N. (1997), *Introduction to Linear Optimization*, Nashua, NH: Athena.

Birch, M. W. (1964), "The detection of partial association, I: the 2×2 case," *Journal of the Royal Statistical Society* B, 26, 313-324.

Block, B. M., Liu, S. S., Rowlingson, et al. (2003), "Efficacy of postoperative epidural analgesia," *Journal of the American Medical Association,* 290, 2455-2463.

Campbell, D. T. (1969), "Prospective: artifact and control," in R. Rosenthal and R. Rosnow, eds., *Artifact in Behavioral Research*, New York: Academic Press, pp. 351-382.

Cochran, W.G. (1965), "The planning of observational studies of human populations (with Discussion)," *Journal of the Royal Statistical Society* A, 128, 234-265.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., Wynder, E. (1959), "Smoking and lung cancer," *Journal of the National Cancer Institute*, 22, 173-203.

Cox, D. R. (1966), "A comparison involving quantal data," *Biometrika*, 53, 215-220.

Dawid, A. P. (1979, "Conditional independence in statistical theory," *Journal of the Royal*

Statistical Society, B 41, 1-31.

Derigs, U. (1988), "Solving nonbipartite matching problems by shortest path techniques," *Annals of Operations Reseach*, 13, 225-261.

Fisher, R.A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.

Gart, J. J. (1969), "An exact test for comparing matched proportions in crossover designs," *Biometrika*, 56, 75-80.

Hansen, B.B. (2007), "Optmatch," *R News*, 7, 18-24.

Hansen, B.B. (2008), "The prognostic analogue of the propensity score," *Biometrika*, 95, 481-488.

Heller, R., Small, D., Rosenbaum, P. R. (2010), "Using the cross-match test to appraise covariate balance in matched pairs," *American Statistician*, 64, 299-309.

Hollander, M., Pledger, G. and Line, P. (1974), "Robustness of the Wilcoxon test to a certain dependency between samples," *Annals of Statistics*, 2, 177-181.

Imbens, G. W. (2003), "Sensitivity to exogeneity assumptions in program evaluation," *American Economic Review*, 93, 126-132.

Knaus, W. A., Wagner, D. P., Draper, E. A., et al. (1991), "The APACHE III prognostic system," *Chest*, 100, 1619-1636.

Kuhn, H.W. (1955), "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, 2, 83-97.

Lu, B., Greevy, R., Xu, X., Beck, C. (2011), "Optimal nonbipartite matching and its statistical applications," *American Statistician*, 65, 21-30. (Also nbpmatching in R)

Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998), "Assessing the sensitivity of regression results to unmeasured confounders," *Biometrics*, 54, 948-963.

Marcus, S. M. (1997), "Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect," *Journal of Educational Statistics*, 22, 193-201.

Neyman, J. (1923), "On the application of probability theory to agricultural experiments: Essay on principles, Section 9," reprinted in *Statistical Science*, 5, 463-480.

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

R Development Core Team. (2007), *R*, Vienna: R Foundation, http://www.R-project.org.

Reiter, J. (2000), "Using statistics to determine causal relationships," *American Mathematical Monthly*, 107, 24-32.

Robins, J. M., Rotnitzky, A. and Scharfstein, D. (1999), "Sensitivity analysis for selection bias and unmeasured confounding in causal inference,"In *Statistical Models in Epidemiology*, Ed. E. Halloran and D. Berry, pp. 1-94," NY: Springer.

Rodgers, A., Walker, N., Schug, S., et al. (2000), "Reduction of postoperative mortality and morbidity with epidural or spinal anaesthesia," *British Medical Journal*, 321, 1493-1499.

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

Rosenbaum, P.R. and Rubin, D.B. (1985), "The bias due to incomplete matching," *Biometrics*, 41, 103-116.

Rosenbaum, P. R. (1987), "Sensitivity analysis for certain permutation inferences in matched observational studies," *Biometrika*, 74, 13–26.

Rosenbaum, P.R. (1989), "Optimal matching in observational studies," *Journal of the American Statistical Association*, 84, 1024-1032.

Rosenbaum, P.R. (2001), "Replicating effects and biases," *American Statistician*, 55, 223-227.

Rosenbaum, P. R. (2002), *Observational Studies*, New York: Springer.

Rosenbaum, P.R. (2005), "Exact distribution free test comparing two multivariate distributions based on adjacency," *Journal of the Royal Statistical Society*, B67, 515-30.

Rosenbaum, P. (2008), "Testing hypotheses in order," *Biometrika*, 95, 248-252.

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

Rosenbaum, P. R. and Silber, J. H. (2009), "Amplification of sensitivity analysis in observational studies," *Journal of the American Statistical Association*, 102, 75-83.

Rosenbaum, P. R. (2010a), "Evidence factors in observational studies," *Biometrika*, 97, 333-345.

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

Rosenbaum, P. R. (2011a), "Some approximate evidence factors in observational studies," *Journal of the American Statistical Association*, 106, 285-295.

Rosenbaum, P. R. (2011b), "Optimal matching of an optimally chosen subset in observational studies," *Journal of Computational and Graphical Statistics*, to appear. (On-line early: DOI: 10.1198/jcgs.2011.09219.)

Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66, 688-701.

Silber, J. H., Rosenbaum, P. R., Even-Shoshan, O., Mi, L. Y., Kyle, F. A., Teng, Y., Bratzler, D. W., and Fleisher, L. A. (2011a), "Estimating anesthesia time using the Medicare Claim: A validation study," *Anesthesiology*, 115, 322-333.

Silber, J. H., Rosenbaum, P. R., Ross, R. N., Even-Shoshan, O., Kelz, R. R., Neuman, M. D., Reinke, C. E., David, G., Saynisch, P. A., Kyle, F., Bratzler, D. W., Fleisher, L. A. (2011b), "Medical and financial risks associated with surgery in the elderly obese," *Annals of Surgery*, to appear.

Small, D. S. (2007), "Sensitivity analysis for instrumental variables regression with overidentifying restrictions," *Journal of the American Statistical Association*, 102, 1049-1058.

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

Susser, M. (1973), *Causal Thinking in the Health Sciences*, New York: Oxford.

Susser, M. (1987), "Falsification, verification and causal inference in epidemiology," in Susser, M., ed., *Epidemiology, Health and Society*, pp. 82-93, New York: Oxford.

Welch, B. L. (1937), "On the z-test in randomized blocks," *Biometrika*, 29, 21-52.

Wiklund, R. A., Rosenbaum, S. H. (1997), "Anesthesiology," *New England Journal of Medicine*, 337, 1132-1141.

Yanagawa, T. (1984), "Case-control studies: assessing the effect of a confounding factor," *Biometrika*, 71, 191-194.

Yang, D., Small, D., Silber, J. H., and Rosenbaum, P. R. (2012), "Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes," *Biometrics*, to appear. (On-line early: DOI: 10.1111/j.1541-0420.2011.01691).

Yu, B. B., Gastwirth, J. L. (2005), "Sensitivity analysis for trend tests: application to the risk of radiation exposure," *Biostatistics*, 6, 201-209.

Zhang, K., Small, D. S., Lorch, S., Srinivas, S. and Rosenbaum, P. R. (2011), "Using split samples and evidence factors in an observational study of neonatal outcomes," *Journal of the American Statistical Association*, 106, 511-524.

Table 1: Counts of patients by hospital, general or regional anesthesia, and pair type. Type 1 pairs finely balanced hospitals, while type 2 pairs contrast hospitals with different usual practices.

	Finely Balanced (type 1)		Usual Practice (type 2)	
Hospital	General	Regional	General	Regional
1	77	77	0	11
2	8	8	0	102
3	101	101	0	4
4	10	10	0	52
5	30	30	0	2
6	3	3	0	163
7	42	42	3	0
8	18	18	0	66
9	48	48	0	10
10	14	14	0	54
11	34	34	0	22
12	25	25	20	0
13	27	27	48	0
14	30	30	0	17
15	16	16	0	44
16	19	19	0	32
17	10	10	0	16
18	21	21	45	0
19	32	32	13	0
20	19	19	0	38
21	5	5	0	87
22	17	17	0	16
23	1	1	0	0
24	29	29	0	14
25	14	14	0	41
26	17	17	71	0
27	35	35	168	0
28	92	92	0	0
29	34	34	11	0
30	11	11	19	0
31	16	16	0	60
32	10	10	42	0
33	39	39	11	0
34	0	0	63	0
35	33	33	0	8
36	27	27	25	0
37	4	4	135	0
38	123	123	11	0
39	13	13	23	0
40	44	44	0	5
41	18	18	0	68
42	9	9	27	0
43	47	47	21	0
44	73	73	0	4
45	22	22	97	0
46	33	33	91	0
47	4	4	0	8

Table 2: Balance on 44 covariates in matched pairs, regional-vs-general anesthesia. Type 1 pairs finely balance the 47 hospitals, while type 2 pairs make contrasts between hospitals with different usual practices.

	Finely Balanced Pairs (type 1)		Usual Practice Pairs (type 2)	
Covariate	Regional	General	Regional	General
Propensity score (mean)	52.3	52.3	52.7	52.6
Risk score (mean)	0.003	0.002	0.002	0.002
Knee 8154 %	97.5	97.0	98.9	99.0
Knee 8155 %	2.5	3.0	1.1	1.0
Age (mean)	72.4	72.5	72.6	72.4
Sex % Female	64.0	65.5	64.4	63.1
White %	92.8	92.8	95.9	95.8
Black %	2.8	3.4	2.1	2.0
Emergency Room Admission %	0.4	0.3	0.0	0.0
Body Mass Index, BMI (mean)	31.2	31.4	31.4	31.2
BMI<18 %	0.3	0.3	0.0	0.0
BMI ge 30 %	52.4	52.4	54.1	54.1
Height (mean)	65.7	65.6	65.9	66.1
ASA Score, 1-5 (mean)	2.5	2.5	2.5	2.5
ASA Score missing %	1.3	1.6	0.8	0.8
Systolic blood pressure (mean)	143.0	143.0	142.5	142.9
Blood pressure missing %	1.4	1.0	0.2	0.2
APACHE Score (mean)	22.4	22.6	22.3	22.4
	Comorbid Conditions			
Diabetic on medication %	15.7	15.2	13.9	15.4
Diabetic score, 1-3 (mean)	0.4	0.4	0.3	0.4
Number of cardiac comorbidities (mean)	0.4	0.4	0.4	0.4
Congestive heart failuare %	6.5	5.8	4.1	5.2
Past MI %	4.3	3.9	3.9	4.1
Past arrhythmia %	14.7	15.1	16.0	16.5
Unstable angina %	0.5	0.9	1.0	0.4
Angina %	2.0	2.5	2.3	1.5
Hypertension %	81.3	81.2	78.8	79.6
Coagulopathy %	0.2	0.1	0.0	0.0
Stroke %	3.8	4.6	1.1	1.5
Dementia %	2.1	1.5	0.6	1.5
Electrolyte abnormality %	2.3	2.7	1.9	2.4
Valvulardis %	10.4	10.1	8.2	8.2
Chronic pulmonary disease %	11.0	11.4	8.8	10.6
Asthma %	5.9	6.2	5.9	7.1
Liver disease %	1.2	1.3	0.1	0.3
Renal dysfunction %	2.0	2.3	0.5	0.6
Renal failure %	1.5	1.5	0.3	0.3
Paraplegia %	0.2	0.4	0.1	0.1
Smoking history %	5.3	6.3	2.6	3.0
Pulmonary fibrosis %	1.8	2.7	1.1	1.1
Cancer %	13.3	13.9	12.1	12.8
Abdominal cancer %	0.1	0.0	0.0	0.0
Weight loss %	0.2	0.2	0.1	0.1
Sleep apnea %	1.0	0.7	0.1	0.2

Table 3: Counts of discordant pairs for two outcomes and their combination, with McNemar P-values, and Gart's test for interaction with type. There are 2298 pairs, 1354 pairs that finely balance the 47 hospitals (type 1) and 944 pairs that contrast hospitals with different usual practices (type 2). Only discordant pairs are recorded here. For readmission, there were 13 deaths within 30 days, 6 with general anesthesia, 7 with regional anesthesia, and these are included in the unfavorable outcome category in each comparison.

	Alive without deep vein thrombosis		
	All	Finely Balanced Pairs (type 1)	Usual Practice Pairs (type 2)
General	72	42	30
Regional	100	66	34
P-value	0.040	0.027	0.708
Odds Ratio (OR)	1.39	1.57	1.13
95% CI for OR	[1.03, 1.88]	[1.07, 2.31]	[0.69, 1.85]
P-value for interaction		0.34	

	Alive without readmission at 30 days		
	All	Finely Balanced Pairs (type 1)	Usual Practice Pairs (type 2)
General	106	61	45
Regional	135	84	51
P-value	0.071	0.068	0.610
Odds Ratio (OR)	1.27	1.38	1.13
95% CI for OR	[0.99, 1.64]	[0.99, 1.91]	[0.75, 1.69]
P-value for interaction		0.51	

	Alive without both deep vein thrombosis and readmission at 30 days		
	All	Finely Balanced Pairs (type 1)	Usual Practice Pairs (type 2)
General	152	84	68
Regional	200	128	72
P-value	0.012	0.003	0.800
Odds Ratio (OR)	1.32	1.52	1.06
95% CI for OR	[1.07, 1.62]	[1.16, 2.01]	[0.76, 1.47]
P-value for interaction		0.10	

Table 4: Sensitivity analysis for two outcomes. The table reports the upper bound on a one-sided P-value, which may be doubled for a two-sided P-value. The sensitivity analysis for all 2298 pairs takes the common value of Γ equal to the tabled value for Γ_1.

Alive without deep vein thrombosis or readmission at 30 days (McNemar's test)					
Γ_1	Γ_2	Finely balanced (type 1) (1354 type 1 pairs)	Usual practice (type 2) (944 type 2 pairs)	Fisher combination of type 1 and type 2	All pairs, one Γ 2298 pairs
1.000	1.000	0.001	0.368	0.004	0.005
1.100	1.000	0.010	0.368	0.024	0.048
1.200	1.000	0.044	0.368	0.083	0.196
1.250	1.000	0.079	0.368	0.132	0.317
Time to death or readmission within 30 days (Albers' test)					
Γ_1	Γ_2	Finely balanced (type 1) (1354 type 1 pairs)	Usual practice (type 2) (944 type 2 pairs)	Fisher combination of type 1 and type 2	All pairs, one Γ 2298 pairs
1.000	1.000	0.025	0.312	0.045	0.033
1.100	1.000	0.081	0.312	0.119	0.135

Table 5: The distance matrix Δ, augmented first to Λ, then to Υ. Here, Δ is the first five rows and five columns, Λ is the first six rows and six columns, and Υ is the full eight rows and eight columns.

	γ_{11}	γ_{12}	γ_{21}	γ_{31}	γ_{32}	ϵ_{21}	κ_{31}	κ_{32}
τ_{11}	$\delta_{\tau_{11},\gamma_{11}}$	$\delta_{\tau_{11},\gamma_{12}}$	$\delta_{\tau_{11},\gamma_{21}}$	$\delta_{\tau_{11},\gamma_{31}}$	$\delta_{\tau_{11},\gamma_{32}}$	∞	∞	∞
τ_{21}	$\delta_{\tau_{21},\gamma_{11}}$	$\delta_{\tau_{21},\gamma_{12}}$	$\delta_{\tau_{21},\gamma_{21}}$	$\delta_{\tau_{21},\gamma_{31}}$	$\delta_{\tau_{21},\gamma_{32}}$	0	∞	∞
τ_{22}	$\delta_{\tau_{22},\gamma_{11}}$	$\delta_{\tau_{22},\gamma_{12}}$	$\delta_{\tau_{22},\gamma_{21}}$	$\delta_{\tau_{22},\gamma_{31}}$	$\delta_{\tau_{22},\gamma_{32}}$	0	∞	∞
τ_{31}	$\delta_{\tau_{31},\gamma_{11}}$	$\delta_{\tau_{31},\gamma_{12}}$	$\delta_{\tau_{31},\gamma_{21}}$	$\delta_{\tau_{31},\gamma_{31}}$	$\delta_{\tau_{31},\gamma_{32}}$	∞	$\widetilde{\delta}/2$	$\widetilde{\delta}/2$
τ_{32}	$\delta_{\tau_{32},\gamma_{11}}$	$\delta_{\tau_{32},\gamma_{12}}$	$\delta_{\tau_{32},\gamma_{21}}$	$\delta_{\tau_{32},\gamma_{31}}$	$\delta_{\tau_{32},\gamma_{31}}$	∞	$\widetilde{\delta}/2$	$\widetilde{\delta}/2$
ε_{11}	0	0	∞	∞	∞	∞	∞	∞
$\varkappa_{31}$	∞	∞	∞	$\widetilde{\delta}/2$	$\widetilde{\delta}/2$	∞	0	∞
$\varkappa_{32}$	∞	∞	∞	$\widetilde{\delta}/2$	$\widetilde{\delta}/2$	∞	∞	0

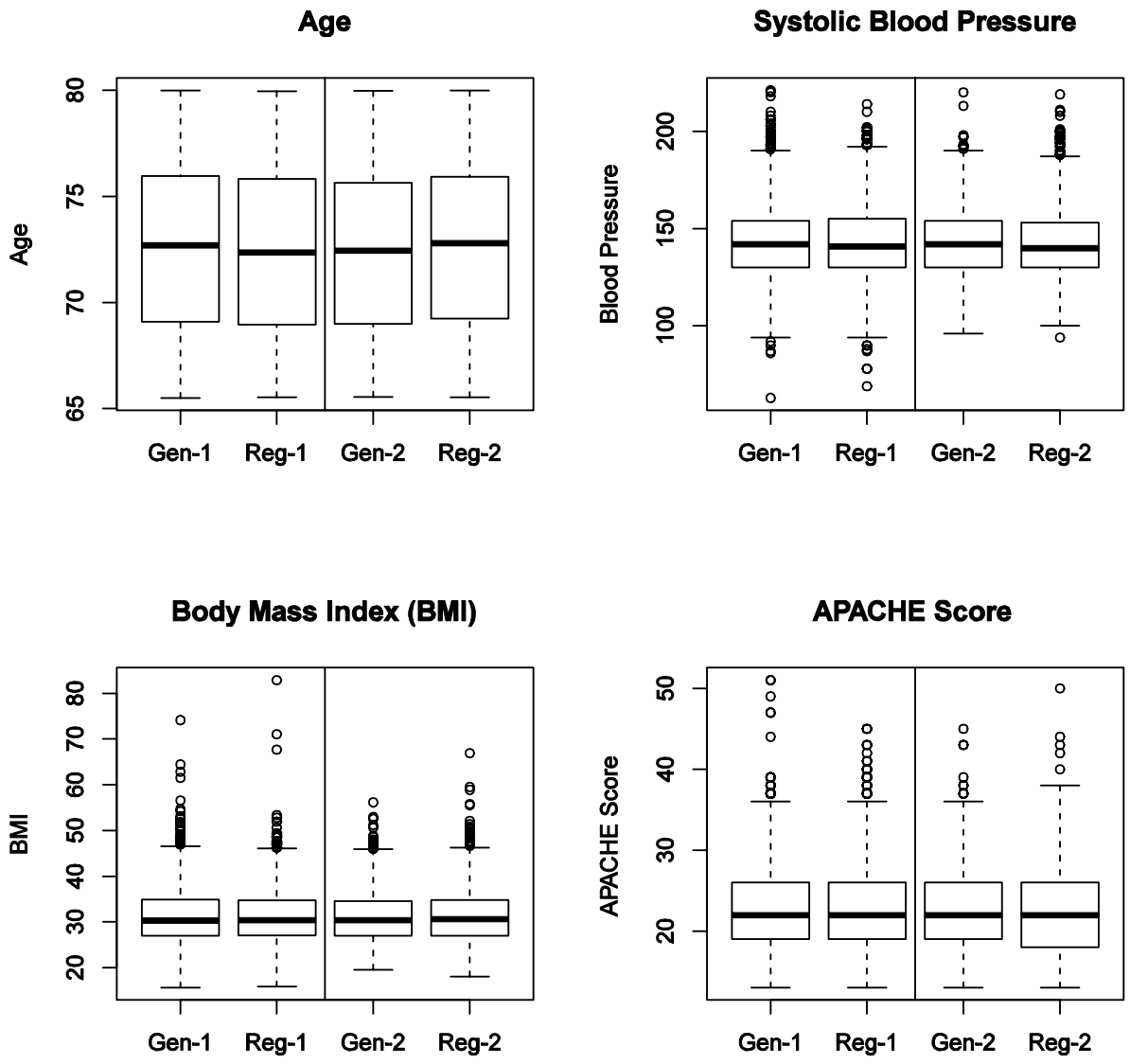

Figure 1: Covariate balance after matching for four continuous covariates. Gen = only general anesthesia, Reg = some regional anesthesia. The 1354 matched pairs of type 1 are finely balanced for the 47 hospitals, while the 944 pairs of type 2 contrast hospitals that typically use general anesthesia to those that typically use regional anesthesia.

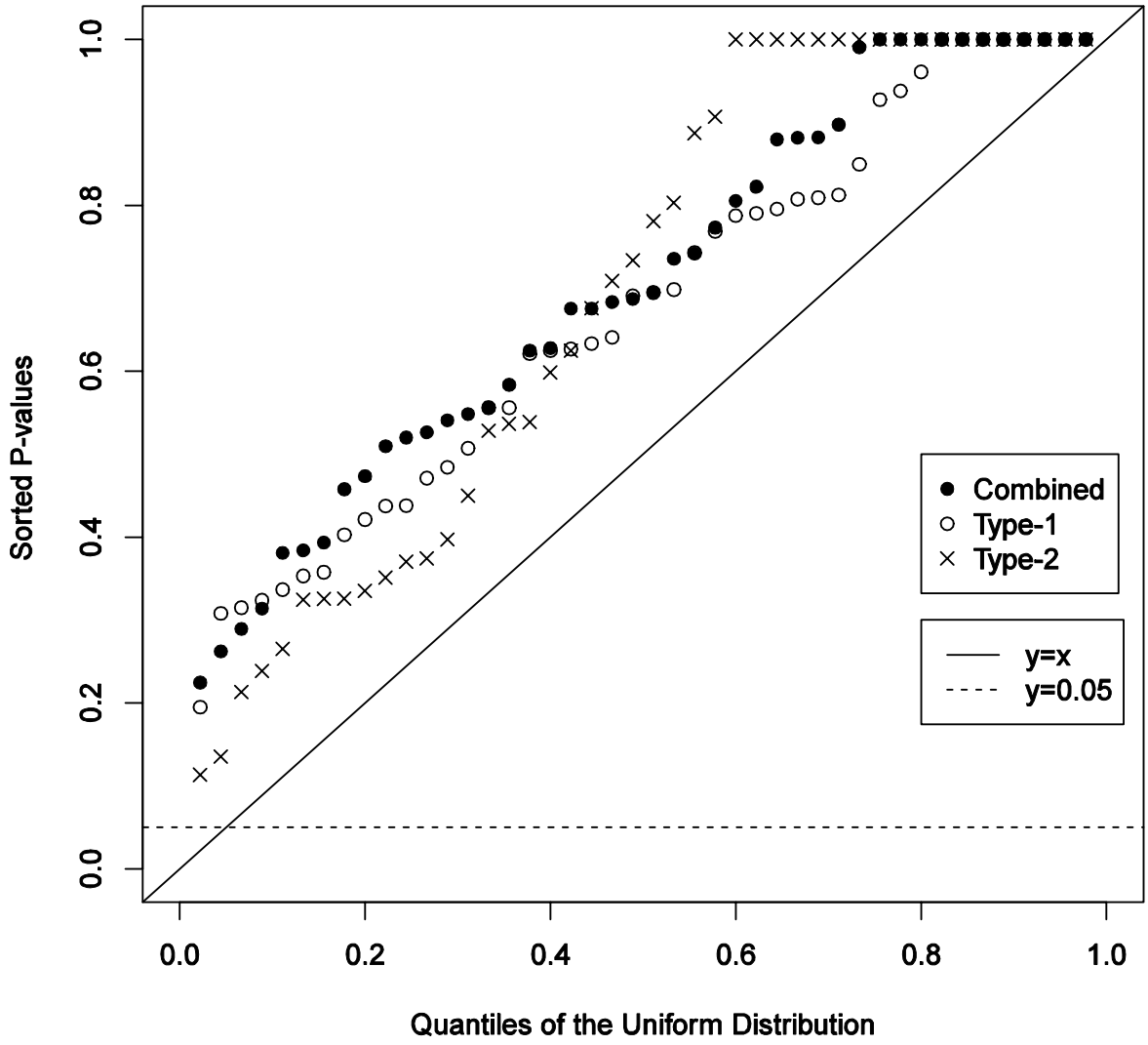

Figure 2: Balance on 44 covariates compared to balance expected from complete randomization. The figure contains three quantile-quantile plots of 44 P-values against the uniform distribution. The P-values contrast the marginal distributions of the 44 covariates in the regional and general anesthesia groups, for 2708 patients in type 1 matches, 1888 patients in type 2 matches and 4596 = 2708 + 1888 patients in both types combined. For continuous covariates, the P-values are from Wilcoxon's rank sum test, and for binary covariates they are from Fisher's exact test for a 2x2 table. All 132 = 3 x 44 P-values are greater than 0.05, and they are larger than expected from a uniform distribution.

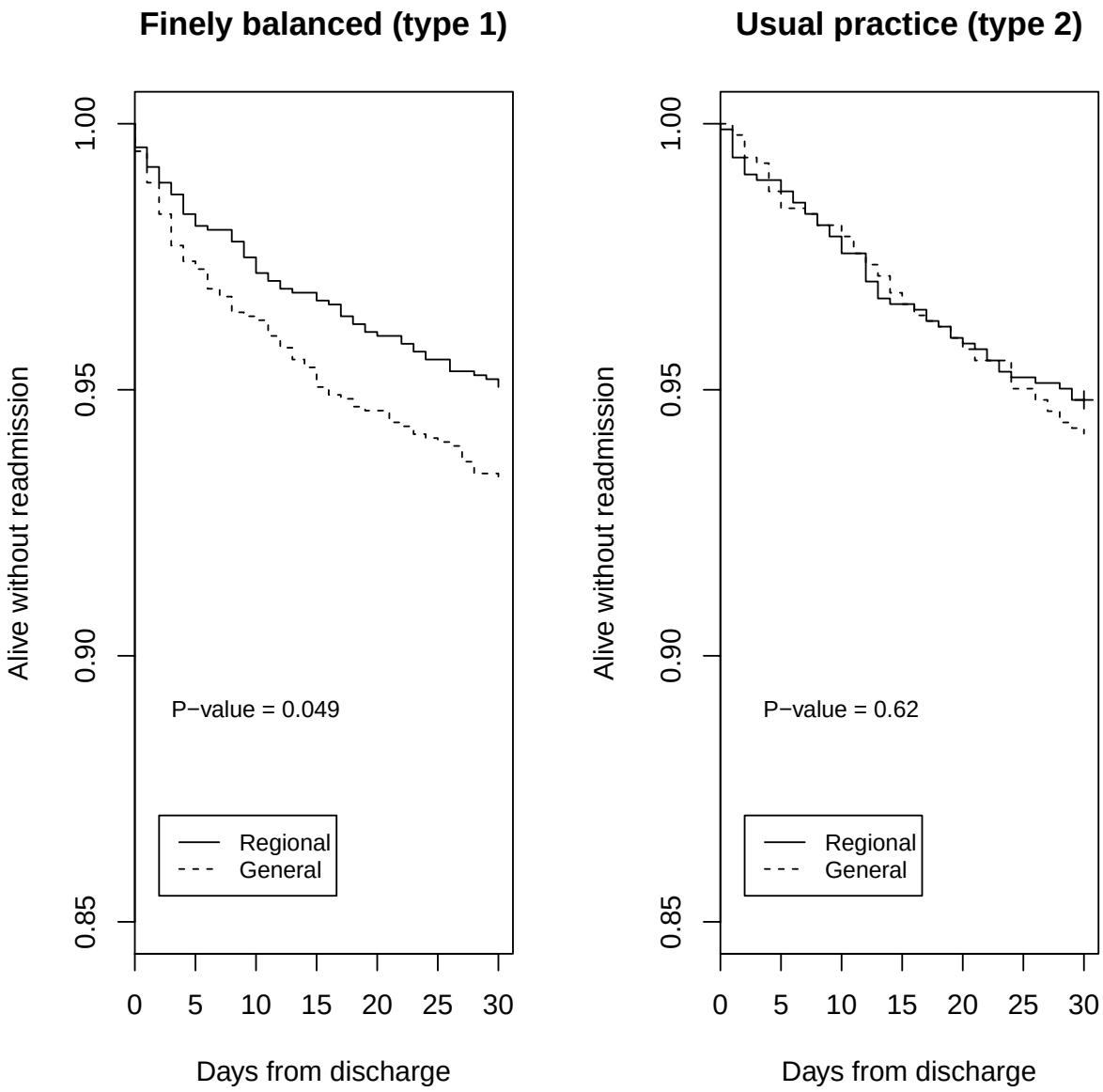

Figure 3: Kaplan-Meier curves for readmission within 30 days of discharge from the hospital. The figure plots the estimate of the probability of being alive without readmission at various days after discharge. The few patients who died in the hospital or who died within 30 days of discharge are counted at the appropriate time as not "alive without readmission". The P-values are from Albers (1988) test for paired censored survival times with Wilcoxon scores.