Balanced Risk Set Matching

Yunfei Paul Li, Kathleen J. Propert, and Paul R. Rosenbaum

A new form of matching—optimal balanced risk set matching—is applied in an observational study of a treatment, cystoscopy and hydrodistention, given in response to the symptoms of the chronic, nonlethal disease interstitial cystitis. When a patient receives the treatment at time t, that patient is matched to another patient with a similar history of symptoms up to time t who has not received the treatment up to time t; this is risk set matching. By using a penalty function in integer programming in a new way, we force the marginal distributions of symptoms to be balanced in the matched treated and control groups. Among all balanced matchings, we pick the one that is optimal in the sense of minimizing the multivariate pretreatment covariate distance within matched pairs. Under a simple model for the treatment assignment mechanism, we study the sensitivity of the findings to hidden biases. In particular, we show that a simple, conventional sensitivity analysis is appropriate with risk set matching when the time to treatment follows a proportional hazards model with a time-dependent unobserved covariate.

KEY WORDS: Integer programming; Matched sampling; Network flow; Observational study; Sensitivity analysis.

1. MATCHING TO ACHIEVE COMPARABLE HISTORIES

1.1 Observational Study of Interstitial Cystitis

With chronic, symptomatic diseases, medical intervention often is given to patients in response to severe, perhaps recently intensified, symptoms. Ideally, the effects of such an intervention would be studied in a controlled trial in which subjects were randomly assigned to treatment or control, so that treated and control subjects would be comparable before treatment. Without random assignment, treated patients may be more severely ill than untreated patients, or they may have received the treatment in response to a transient but acute bout of symptoms. However, randomized trials in such settings are not always possible for ethical or practical reasons.

An example of this occurs in interstitial cystitis (IC), whose symptoms are bladder pain and irritative voiding, which resemble the symptoms of a urinary tract infection, but there is no evidence of infection. Although IC was described more than 80 years ago (Hunner 1918), knowledge of its causes and natural history are limited (Curhan et al. 1999).

We examine the effects of a surgical intervention, cystoscopy and hydrodistention, on the symptoms of IC by using data from the Interstitial Cystitis Data Base (ICDB), a multicenter cohort study sponsored by the National Institute of Diabetes, Digestive, and Kidney Diseases (Simon et al. 1997; Propert et al. 2000). Patients began enrolling in the database in 1993. To be eligible for the database, a patient must have exhibited, for at least 6 months before entry, symptoms of urinary urgency, urinary frequency, or pelvic pain and so would have been considered to have IC (Hanno et al. 1999). Patients were evaluated at entry into the database and at intervals of approximately every 3 months thereafter for up to 4 years. Three quantities were measured repeatedly over time: pain, urgency, and nocturnal frequency of voiding. Pain and urgency are subjective appraisals on a scale from 0 to 9, with higher numbers signifying greater intensity. Patients were treated by usual clinical practice with no specific treatments required by the study protocol. At some point after enrollment, some patients were treated by cystoscopy and hydrodistention, perhaps in response to acute symptoms.

1.2 Risk Set Matching

If patient m received the treatment at time T_m after entry into the study, one would like to compare the response of this patient to a patient who did not receive the treatment up to time T_m but who otherwise appeared similar during that pretreatment interval, that is, who had a similar history of symptoms. Or, at least, one would like treated and control groups whose aggregate distributions of symptoms were similar. Our new matching algorithm, described in Section 2 and applied in Section 3, works at both goals, that is, it aims for similar patients in each pair and balanced groups in aggregate. In the end, we will have S matched pairs, $s = 1, \ldots, S$, containing $2S$ distinct patients, such that the treated patient in pair s received the treatment at time T_s, and the control in pair s either did not receive the treatment at all or received it strictly after T_s.

The term "risk set matching" refers to the risk set that arises in the partial likelihood associated with Cox's (1972, 1975) proportional hazards model. Recall that this partial likelihood compares an individual who experiences an event at time t to all other individuals at risk of the event at time t, thereby eliminating a nuisance parameter of infinite dimension describing variations in risk over time. The matching method used here pairs a patient treated at time T_s with a similar patient untreated at time T_s but at risk of treatment at time T_s, that is, to a similar patient in the risk set. Sampling or matching from a risk set was discussed by Prentice and Breslow (1978) as a model for case-control studies, by Prentice (1986) as a model for case-cohort studies, and by Oakes (1981) as a computational simplification; see Langholz and Goldstein (1996) for a survey. Unlike these authors, we build a model for the time to treatment, not the time to an outcome event. Like Prentice

Yunfei Paul Li is Senior Consultant, Economic Consulting Services, KPMG LLP, Washington, DC 20036 (E-mail: *yunfeili@kpmg.com*). Kathleen J. Propert is Associate Professor, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6021. Paul R. Rosenbaum is Professor, Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6302. The data were provided courtesy of the Interstitial Cystitis Data Base Study Group. Propert was partially supported by a grant from the National Institute of Diabetes, Digestive, and Kidney Diseases. Rosenbaum was supported by a grant from the National Science Foundation and by the Center for Advanced Study in the Behavioral Sciences. The authors thank Abba Krieger and Monique Guignard-Spielberg for advice on integer programming and Colleen Brensinger for data analysis.

© 2001 American Statistical Association

and Breslow (1978, sec. 2) but unlike most applications of risk sets, because our algorithm yields nonoverlapping samples from the risk set, we obtain a conditional distribution rather than a partial likelihood; see Section 4. Also, the sensitivity analysis model in Section 4 involves an unobserved time-dependent covariate that could not be controlled by matching.

Notice that we match only on past data, never on future data. This is why a patient treated at time T_s is matched to a patient not yet treated at time T_s rather than to a patient who was never treated. To make this clear, consider an extreme hypothetical but straightforward illustration. Imagine a strict rule that assigned patients to treatment whenever their symptoms became acute. In this hypothetical case, to know that a patient never received treatment is to know the patient's symptoms never became acute, that is, to know that the patient had a relatively favorable outcome. If the control group consisted of all patients who never received treatment, then it would contain only patients with favorable outcomes, because any patient whose symptoms later became acute received the treatment. In sharp contrast to this case, our algorithm would immediately reveal that there is no matching that balances covariates in the treated and control groups at the time of treatment, thereby clearly warning of the extreme biases that are present. We must compare a patient treated at time T_s with a similar patient not yet treated at time T_s, but we must not use future data on either patient in deciding whether this is a good match. In Section 4, we develop formally the issues in this sort of risk set matching, and we consider the possibility of bias due to an unobserved time-varying covariate that was not controlled by the matching.

Risk set matching differs from matching on baseline variables in two ways. First, when a potential control is considered as a possible match to patient m, who was treated at time T_m, the relevant covariates for matching are from baseline to time T_m, but when this same control is considered as a possible match to patient m', who was treated at a different time $T_{m'}$, the relevant covariates are from baseline to time $T_{m'}$. Second, a patient m treated at time T_m can enter the study in exactly one of two ways—as a treated patient at time T_m or as a not-yet-treated control for a patient treated strictly before T_m.

2. OPTIMAL BALANCED MATCHING

2.1 Matching by Minimum Cost Flow in a Network

Stated abstractly, the optimal pair matching problem involves a finite set $\mathcal{A} = \{\alpha_1, \ldots, \alpha_M\}$ called units, a subset $\mathcal{T} \subseteq \mathcal{A}$ called treated units, and a subset of their direct product $\mathcal{E} \subseteq \mathcal{T} \times \mathcal{A}$ called edges. If the pair $e = (\alpha_p, \alpha_q)$ is an edge $e \in \mathcal{E}$, then it is permitted to match α_p to α_q, but if $e \notin \mathcal{E}$, then this match is forbidden.

In our study, $\mathcal{A}$ consists of 400 patients randomly sampled from the IC database, and $\mathcal{T} \subseteq \mathcal{A}$ consists of all patients in the sample who eventually received the treatment. The pair $e = (\alpha_p, \alpha_q)$ is an edge $e \in \mathcal{E}$ if α_p received the treatment, say, at time T_p, and α_q either never received the treatment or received it strictly after time T_p. In principle, the set $\mathcal{E}$ may exclude certain pairs for additional reasons, such as being too far apart on an important covariate, but that was not done in the current study. Notice also that the abstract statement may be applied without a risk set in a study in which everyone receives either treatment or control immediately at baseline; then, $\mathcal{T}$ contains the treated subjects, $\mathcal{A} - \mathcal{T}$ contains the controls, and $\mathcal{E} \subseteq \mathcal{T} \times (\mathcal{A} - \mathcal{T})$ requires treated subjects to be matched to untreated controls.

For each $e \in \mathcal{E}$, there is a nonnegative distance $\delta_e \geq 0$. A commonly used distance in matching is the Mahalanobis distance; see Rubin (1980). Suppose $(\alpha_p, \alpha_q) = e \in \mathcal{E}$, and α_p received the treatment, say at time T_p, and α_q received the treatment later or not at all. Then, in our study, the distance δ_e is the Mahalanobis distance between subject α_p and control α_q on a six-dimensional covariate describing the three symptoms at baseline and at time T_p when α_p received treatment.

A pair matching of size S is a subset $M \subseteq \mathcal{E}$ with $|M| = S$ edges such that each unit $\alpha_q \in \mathcal{A}$ appears in at most one matched pair, possibly as $(\alpha_p, \alpha_q) \in M$ or as $(\alpha_q, \alpha_p) \in M$ but not as both. A pair matching is optimal of size S if it minimizes the total distance within pairs, $\sum_{e \in M} \delta_e$ over all pair matchings M of size S obtainable with the given structure $\mathcal{A}$, $\mathcal{T}$, $\mathcal{E}$. In our study, we picked $|M| = S = 100$ matched pairs of a treated patient and a not-yet-treated control.

In the fields of operations research and computer algorithms, there is a large literature on optimal matching by minimum cost flow in a network, and fast algorithms are available. Textbook discussions were given by Papadimitriou and Steiglitz (1982, sec. 11.2) and Bertsekas (1991, sect. 1.1). Optimal pair matching in observational studies was discussed by Rosenbaum (1989) and Gu and Rosenbaum (1993), and an implementation in the computer package SAS was discussed by Bergstralh, Kosanke, and Jacobsen (1996). Optimal matching with multiple controls was illustrated by Ming and Rosenbaum (2000) in a study of mortality following surgery.

In a clinical trial, Pocock and Simon (1975) used a multivariate sequential procedure to approximate covariate balance. Their goal was similar in some respects to our goal, but their method was quite different, because patients entered the trial gradually and treatment assignment was under experimental control.

Because of the structure of the IC database, time is measured in discrete 3-month intervals, so if $e = (\alpha_p, \alpha_q)$ is an edge in $\mathcal{E}$ connecting α_p treated at time T_p to α_q not yet treated at T_p, then α_q was untreated until at least $T_p + 3$. Hence, a comparison of two paired subjects 3 months after treatment is always a comparison of a treated subject and an untreated subject. In contrast, 6 months after treatment, a few not-yet-treated controls will have received treatment after measurements are taken at 3 months. In general, at all time points, the effect under study is the effect of treating now versus not treating now but possibly treating later, that is, the effect of delaying treatment, and of course that is the treatment choice that patients and surgeons keep facing.

Some implementation decisions merit brief mention. Although for balancing we divide covariates at quantiles, when using the Mahalanobis distance δ_e, we use the covariates themselves without division. A single covariance matrix for the Mahalanobis distance was computed from all baseline measurements and all later measurements for all patients in $\mathcal{A}$. More precisely, the six-dimensional variable containing the baseline and current pain, urgency, and frequency measurements is found for every patient at every time point, and

the six-by-six covariance matrix is computed from all these measurements, so most patients count several times when the covariance matrix is computed. Other definitions of the covariance matrix with time varying covariates are possible, and, at this time, we lack a firm basis for advocating any one definition. We excluded from $\mathcal{E}$ any edge $e = (\alpha_p, \alpha_q)$, which would pair individuals who were missing needed data at relevant time points; therefore, all our final pairs have complete data. However, an assessment of a patient is always recorded before the patient is treated, so pretreatment measurements are always available for treated patients.

2.2 Balanced Pair Matching

Associated with each treated unit $\alpha_p \in \mathcal{T}$ are K binary variables $B_{pk} = 1$ or $B_{pk} = 0$ for $k = 1, \ldots, K$. For instance, B_{pk} might describe the gender of the pth treated subject. Alternatively, B_{pk} might describe the status of a time-varying attribute at the time α_p receives the treatment or at some other time before treatment. In addition, associated with each potential pairing $(\alpha_p, \alpha_q) = e \in \mathcal{E}$, there are K binary variables, $B_{ek} = 1$ or $B_{ek} = 0$ for $k = 1, \ldots, K$. In the simplest case, with $(\alpha_p, \alpha_q) = e$, the binary variable B_{ek} might describe a baseline measure of potential control α_q, such as gender; however, with a time-varying attribute, B_{ek} might describe the status of potential control α_q at the moment that treated subject α_p received treatment.

This notation is necessary but unusual, and it is helpful to emphasize the sense in which it is unusual. For $(\alpha_p, \alpha_q) = e \in \mathcal{E}$, the variable B_{pk} describes just the treated patient α_p a moment before she received treatment. In contrast, B_{ek} is not just a description of the not-yet-treated control patient α_q. Rather, B_{ek} describes the potential pairing, (α_p, α_q), and depends on information for both patients; specifically, it describes α_q a moment before α_p received the treatment. In a different potential pairing, say $e' = (\alpha_{p'}, \alpha_q)$, this same control α_q often would have a different value for $B_{e'k}$, because α_q is now being described at the moment $\alpha_{p'}$ received treatment, not the moment at which α_p received treatment.

A pair matching M is balanced with respect to these K variables if

$$\sum_{(\alpha_p, \alpha_q) \in M} B_{pk} = \sum_{e \in M} B_{ek} \quad \text{for } k = 1, \ldots, K.$$

Notice that both sums refer to the same matched pairs in M; however, the first sum describes treated patients, and the second sum describes their matched controls at times determined by attributes of the treated subject. For instance, if the first binary variable indicates gender, then in a balanced matching, the total number of males in the treated group equals the total number of males in the matched control group, although individual pairs may not be matched for gender. For a continuous covariate such as age, one can define, say, three additional binary variables indicating whether a subject is older than each of the quartiles in the treated group. Then a balanced matching for these four binary variables would produce matched controls with the same number of males and the same age quartiles as the treated group. For a time-dependent binary variable, say, fever above 101°F, a balanced matching might insist that the number of patients with fevers above 101°F at the time of treatment equals the number of controls with fevers above 101°F at the times treated subjects were treated.

An optimal balanced matching is a balanced pair matching M that minimizes the total distance $\sum_{e \in M} \delta_e$ over all balanced pair matchings. A simple form of optimal balanced matching, in which the binary variables code the categories of a single nominal variable, remains a network flow problem (Rosenbaum 1989); however, in general, optimal balanced matchings must be found by integer programming methods, which are described in the Appendix.

We wanted to balance the three symptom variables at two times: pain, urgency, and frequency at baseline and at the time at which a treated subject received treatment. That is, we wanted to balance a $6 = 2 \times 3$ dimensional variable. We divided each of these 6 variables into three groups of equal sizes, that is, at the one-third and two-thirds percentiles, and we introduced 2 binary variables indicating the group, making 12 binary variables in total. As a result, our matching perfectly balances the one-third and two-thirds quantiles of the six symptom covariates. In addition, we used these same covariates, without the division into coarse groups, in defining the Mahalanobis distance δ_e. As a result, among all perfectly balanced matches, ours minimizes the total distance within matched pairs.

A few of our implementation decisions merit brief discussion. We very much wanted to produce a simple comparison that would be perceived as compelling by urologists treating interstitial cystitis. Cystoscopy and hydrodistention is not a new and experimental treatment; rather, it is the most standard surgical intervention for interstitial cystitis. A negative evaluation of this treatment, should it occur, must be compelling to its audience if it is to have any chance to change current practice. A paired comparison of 200 similar patients at similar moments, half treated, half not yet treated, seemed to us to be the simplest reasonable comparison. This yields treated and control groups that are easily checked for comparability and permits a simple, conventional analysis. We could, as an alternative, have used all 400 patients by matching with a variable number of controls (Ming and Rosenbaum 2000) or by full matching (Rosenbaum 1991; Gu and Rosenbaum 1993). Matching with sets of unequal size is not difficult in a technical sense, but it requires a weighted definition of comparability or covariate balance, so a nontechnical audience may not easily be persuaded that comparable patients are being compared. In a different medical context, we might prefer the larger study, despite its greater complexity. For the same reason, we were content to partition risk sets into comparable pairs, rather than to compare every treated measurement to every untreated measurement. See Rosenbaum (1995, sec. 10) for a discussion of trade-offs of this kind. Unlike a well-conducted clinical trial, our observational study may be biased by a failure to control for important unobserved covariates; however, our study's sampling variability should be comparable to that of a clinical trial with 200 patients in 100 pairs, and there is no study of cystoscopy and hydrodistention of comparable size. Notice also that we did not decide who would be treated and who would be not yet treated; rather, the algorithm considered all possible balanced pairings of 100 pairs and picked the closest one.

In the ICDB, time is measured from entry into the cohort study. In chronic, nonprogressive diseases such as IC, although symptoms may wax and wane over time, there are few differences in overall symptom severity between newly diagnosed patients and those who have had the disorder for many years (Propert et al. 2000; Rovner et al. 2000). In contrast, in progressive diseases such as cancer, time typically would be measured from the date of diagnosis.

3. EFFECTS OF CYSTOSCOPY AND HYDRODISTENTION

3.1 Quality of the Matching

Recall that we balanced the one-third and two-thirds quantiles of pain, urgency, and frequency at baseline and at the time of treatment, and then we minimized the Mahalanobis distance among matched samples that balanced these quantiles. These variables are discrete, so the thirds are not exactly 33%; however, they are exactly matched. For instance, for pain at baseline, 37% of the 100 matched treated patients had pain scores of 3 or lower, and 37% of the 100 matched controls had pain scores of 3 or lower. Similarly, in both groups, 44% had baseline pain scores strictly above 3 and no more than 6, and in both groups, 19% had baseline pain scores strictly above 6. The other five variables are also perfectly balanced at their thirds.

Among all such balanced matchings, the algorithm found a match that minimized the Mahalanobis distance within matched pairs. Before matching, among all potential pairings, the median Mahalanobis distance was 8.8; after matching, it was less than .5. Before matching, the upper quartile of the distances was 14.4; after matching, it was .75. Before matching, the maximum distance was 81.8; after matching, 2.7. As a standard for comparison, two subjects one standard deviation apart on each of six uncorrelated variables would have a Mahalanobis distance of 6. In a single homogeneous six-dimensional multivariate Normal population, Mahalanobis distances between two independent people have a distribution that is twice a chi-square on 6 degrees of freedom, so the expectation is 12. The individual pairs appear to be quite close before treatment.

3.2 Graphical Comparisons

The boxplots in Figures 1–3 give an informal description of the results for $S = 100$ matched pairs. There are three variables—pain score, urgency score, and nocturnal frequency. For each variable, there are six pairs of boxplots, two pairs describing covariates before treatment and four pairs describing outcomes after treatment. Each pair of boxplots compares the treated patients to their matched not-yet-treated controls, labeled "Never/Later Treated" in the plots.

Consider, first, the comparability of the groups before treatment. For each variable, the first pair of boxplots describes patients at entry into the study. The second pair describes patients at the moment the treated patient received treatment. The matching tried to make the treated patients and their not-yet-treated controls comparable before the moment at which the treated patient received treatment. The boxplots indicate that the distributions of the six pretreatment variables were closely balanced. For these six observed covariates, treated and not-yet-treated controls look comparable.

The remaining four pairs of boxplots for each variable describe patient outcomes after treatment. The plots describe the outcome 3 and 6 months after treatment, and the change in the outcome from treatment to 3 or 6 months after treatment. The plots hint at small benefits from treatment for nocturnal frequency and possibly for urgency at 3 months, but there is no visible benefit for pain or urgency at 6 months.

3.3 Inference in the Absence of Hidden Bias

In this section, differences in patient outcomes are estimated by using simple methods that would be appropriate in a paired randomized experiment. In contrast, in Section 3.4, the sensitivities of these findings to departures from randomization are examined. Formal conditions under which these analyses are appropriate are developed in Section 4.

Table 1 compares the 100 matched pairs for the measures frequency, pain, and urgency. The baseline measure describes patients on entry into the study, and the treatment measure describes them a moment before the treated patient in the pair received treatment; these are pretreatment measures controlled by matching. The 3-month measure is 3 months after the time the treated patient received the treatment but before the matched not-yet-treated control received treatment. The boxplots indicate some extreme observations, so a robust estimator, the trimean, is used as a measure of location in Table 1. Recall that the trimean is twice the median plus the quartiles divided by four; see Andrews et al. (1972, p. 8) for discussion. Notice that, as groups, the treated patients and the not-yet-treated controls look quite comparable on all three measures at baseline and before treatment. As the boxplots suggest, 3 months after treatment, the treated responses appear slightly lower (better) than the control responses.

Each pair yields one value of the variable contrast, which is formed as an interaction contrast of six measurements in that pair, three from the treated patient and three from the not-yet-treated control. Specifically, the average of the two pretreatment measures is subtracted from the 3-month measure for the treated and the control patient in each pair, and then the treated-minus-control difference of these quantities is the contrast; i.e., the contrast is

$$\left(\text{Treated}_3 - \frac{\text{Treated}_{\text{base}} + \text{Treated}_0}{2}\right) - \left(\text{Control}_3 - \frac{\text{Control}_{\text{base}} + \text{Control}_0}{2}\right)$$

where base refers to baseline, 0 refers to the time immediately before the treated patient received treatment, and 3 refers to 3 months after that. There is one contrast for each pair, and Table 1 reports the trimean of these 100 contrasts. The value of -0.50 for the trimean of the 100 contrasts for frequency suggests that treated patients improved slightly more than controls did, reducing their frequency of nocturnal voiding by about half a trip a night. The significance levels are based on the signed rank statistic applied to the contrasts.

For each measure, the signed rank test was applied to the contrast to test the hypothesis of no treatment effect. There is

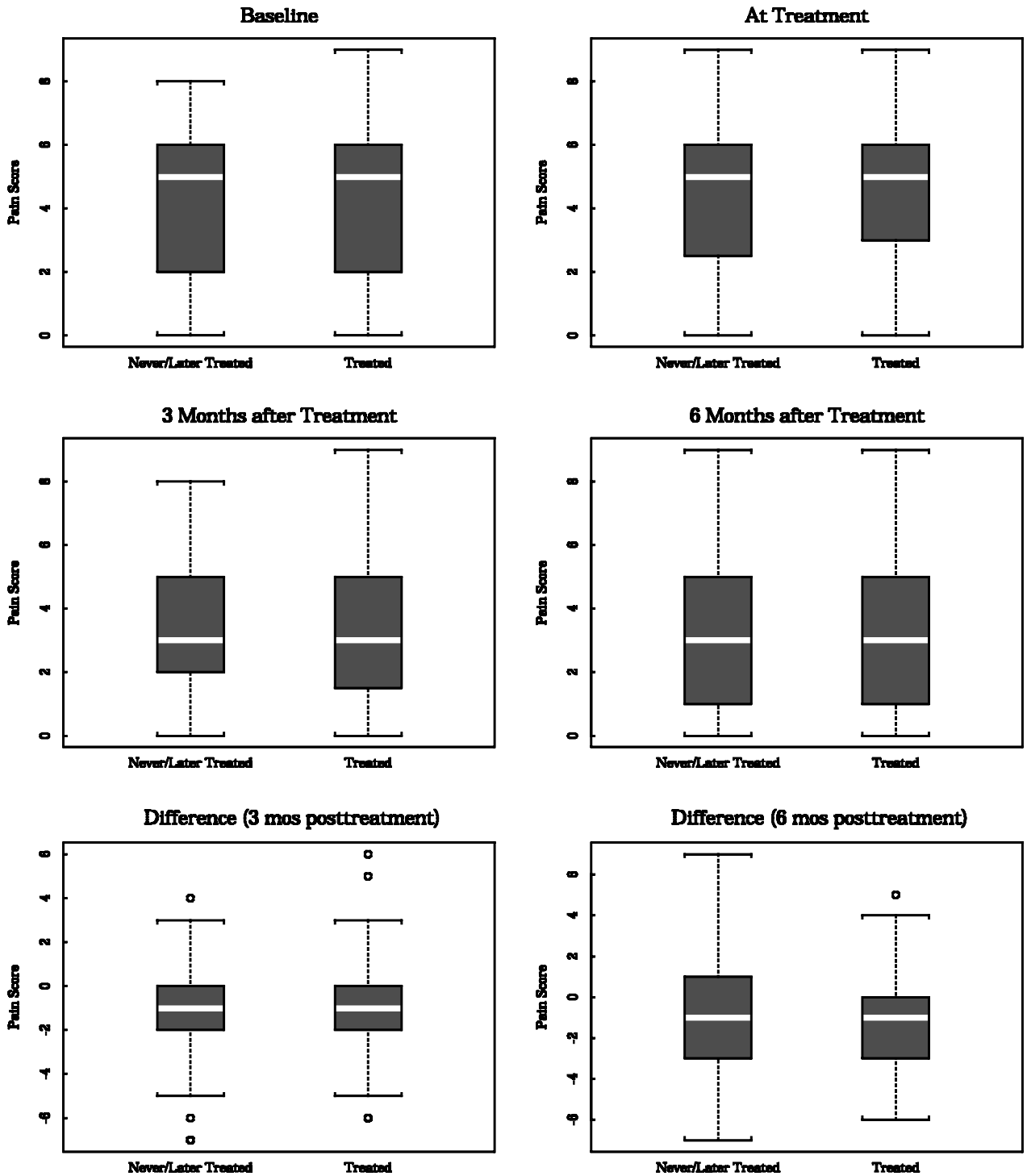

Figure 1. Pain Score.

a significant difference at the .05 level for frequency but not for pain or urgency. On the basis of the Bonferroni inequality, the difference in frequency would remain significant if allowance were made for the testing of three hypotheses. Keep in mind that the point estimate of the magnitude of the gain is about half a visit to the washroom per night for patients whose trimean was two trips before surgery.

Because we hope for improvement in all three outcomes, another approach is to perform a single multivariate test formed by adding the three separate signed rank tests, as discussed in Rosenbaum (1997), where technical details of this simple test may be found. When this is done, the standardized deviate is -1.62, just missing significance in a large-sample, one-sided .05 level test. A different multivariate nonparametric test was proposed by Wei and Lachin (1984).

Although we have not found strong evidence of dramatic effects of delay of this surgery, one might reasonably ask whether we have found strong evidence against dramatic effects. To answer this, we performed a type of equivalence test. In such a test, the null hypothesis asserts that the

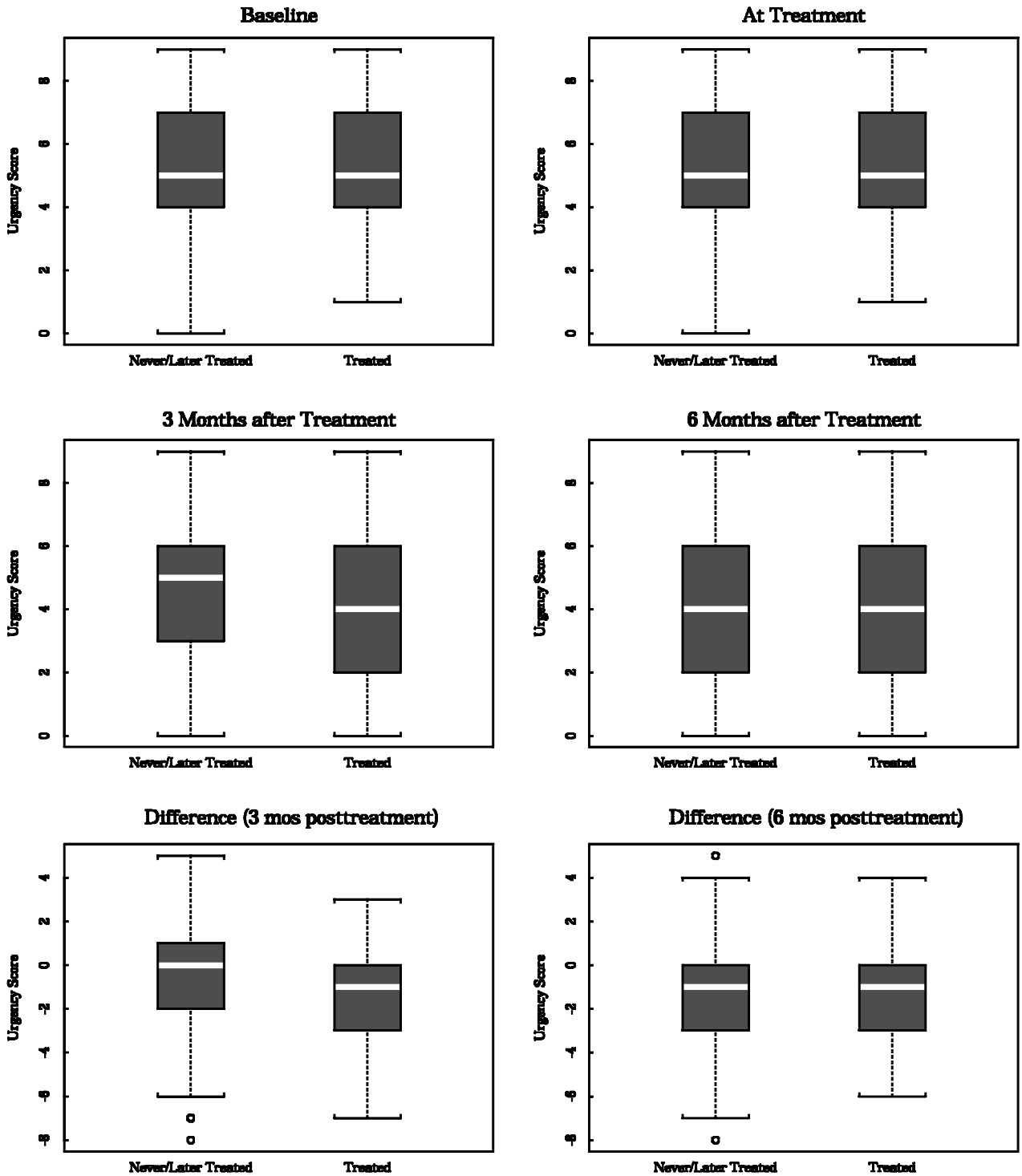

Figure 2. Urgency Score.

treatment has been quite effective, and the alternative says the effects are small or nonexistent, so rejecting the null is strong evidence of the absence of large effects; see, for instance, Hsu et al. (1994). We defined two null hypotheses, asserting that the treatment has an additive effect and is either moderately effective or quite effective. Additive effects are the simplest and most common models of effect, and they are familiar from experimental design (e.g., Cox 1958, sec. 2) and nonparametric shift models (e.g., Hollander and Wolfe 1999, sect. 3); see Rosenbaum (1999a, 2000) for alternative models. For each of the three outcomes, the smallest improvement for one patient that can be recorded on the scales used is a one-unit improvement. This is our first null hypothesis of effectiveness, which states that every treated patient improves on each outcome by one unit. A patient who experienced such an effect would have experienced the smallest simultaneous improvement in all three outcomes that can be recorded for one patient on this measurement scale. (Of course, smaller typical improvements are possible if the improvements affect some patients but not others or some outcomes but not others.) We again used the

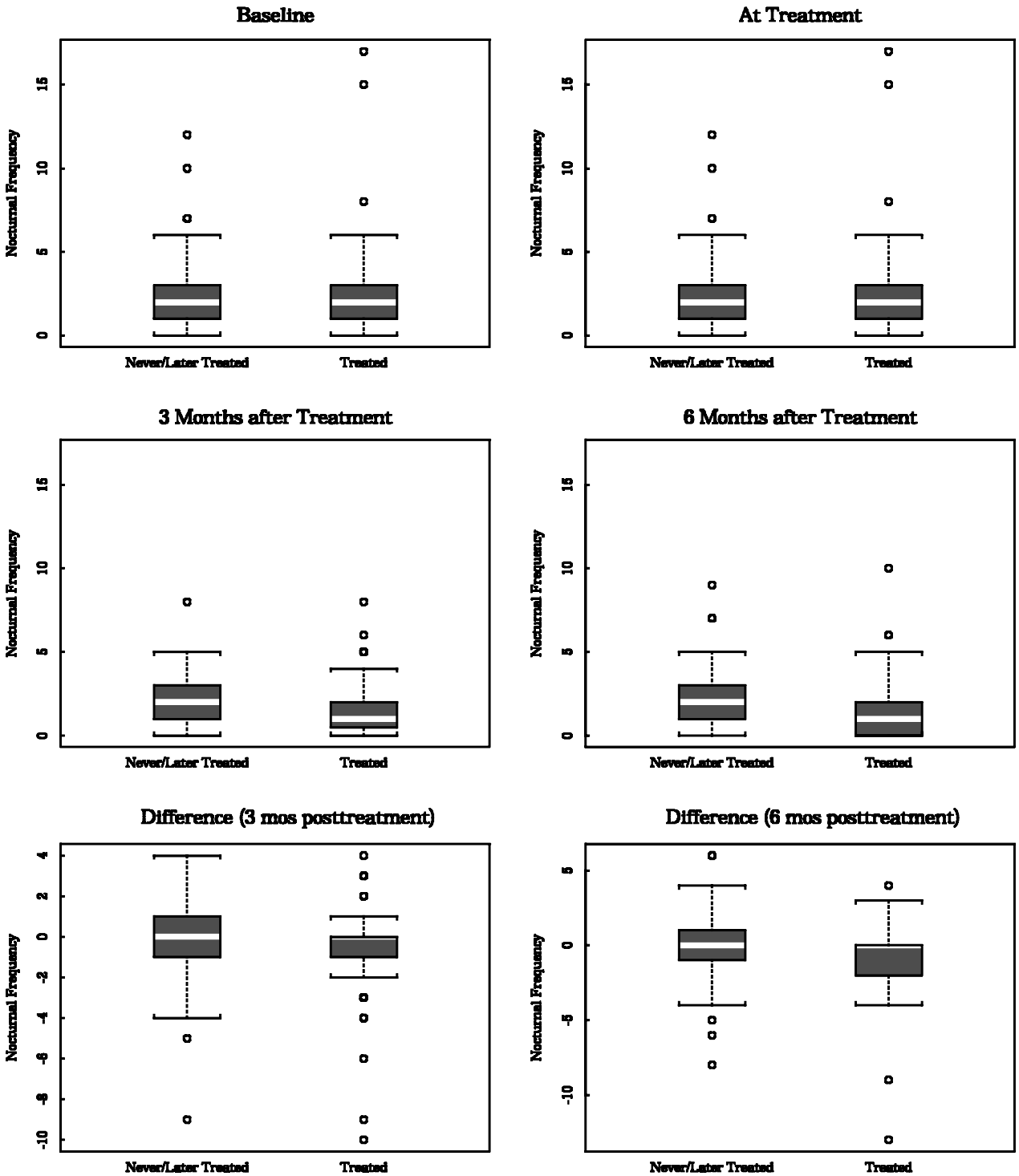

Figure 3. Nocturnal Frequency.

sum of the three signed rank statistics to test this hypothesis, yielding a standardized deviate of 3.57 and a significance level of .00018. This test assumed that there is no hidden bias, so methods for a randomized experiment may be used. If this assumption is correct, then it would not be plausible that the treatment produced a one-unit improvement in all three outcomes.

Our second hypothesis of effectiveness, called half iqr improvement, states that each patient improved by half of the interquartile range at baseline for each of the three measures. In a standard Normal distribution, half of the interquartile range is .674, or about two-thirds a standard deviation, whereas about 95% of the data fall in a range that is four standard deviations in length, so this is a substantial improvement. For the Normal, the interval that includes 95% of the data can be transversed in about six steps, each of which has length equal to half the interquartile range. Visually, in a pair of boxplots, a half iqr improvement would place the endpoint of the box in one group at the center of the box in the other, thereby aligning the median in one group with a quartile in the

Table 1. Trimeans of Pretreatment and Posttreatment Measures

Measure	*Group*	*Baseline*	*Treatment*	*3 Months*	*Contrast*	*P-value*
Frequency	Treated	2.00	2.00	1.19	−0.50	.0032
	Not yet	2.00	2.00	2.00		
Pain	Treated	4.50	4.75	3.19	.12	.78
	Not yet	4.50	4.69	3.25		
Urgency	Treated	5.25	5.25	4.00	−.22	.26
	Not yet	5.25	5.25	4.75		

NOTE: Baseline is upon entry into the study, Treatment is the pretreatment measure at the time at which the treated patient in a pair received treatment, and 3 Months is the outcome 3 months after treatment. For each pair, the interaction Contrast of the six measures in the pair is the treated-minus-control difference between 3 months posttreatment and the average of the two pretreatment measures. A trimean is twice the median plus the quartiles divided by four. The P-value is the two-sided significance level from Wilcoxonfs signed rank statistic applied to the contrasts.

other. At baseline, that is, at entry into the study, in both the treated and matched control groups, the interquartile ranges for pain, urgency, and frequency were 4, 3, and 2, respectively. We considered the hypothesis that the treatment has the effect of reducing each of the three outcomes by half of the interquartile range at baseline, that is, by 2 for pain, by 1.5 for urgency, and by 1 for frequency. An improvement of this magnitude would be clinically important. The standardized deviate for the sum of signed rank statistics is 5.18 with significance level <.00001, so that, again, in the absence of hidden bias, there would be strong evidence against such a large treatment effect.

We now consider how these conclusions might be altered by hidden biases of various magnitudes.

3.4 Sensitivity to Hidden Bias

A sensitivity analysis asks how hidden biases of various magnitudes might alter the conclusion of an observational study (Rosenbaum 1995, sect. 4). In a paired, randomized experiment, a coin flip decides the treatment assignment independently in each pair, and the two patients in a pair have the same chance, one-half, of receiving the treatment. In an observational study, even after matching on the observed covariates, one patient in a matched pair may be more likely than the other to receive the treatment because that patient differed before treatment in some important but unobserved way.

The sensitivity analysis involves a parameter, Γ, that describes the magnitude of the departure from a randomized experiment; specifically, two matched subjects may differ in their chances of receiving the treatment by at most a factor of $\Gamma \geq 1$. For $\Gamma = 1$, there is no hidden bias, the treatment assignment probabilities equal one-half, and the test is the conventional randomization test, reported in Section 3.3. For $\Gamma = 2$, two matched subjects may differ in their chances of receiving the treatment by a factor of 2, so one might be twice as likely to receive the treatment as the other; that is, within a pair, the treatment assignment probabilities might range from one-third to two-thirds, and there is a corresponding range of plausible inferences. If small departures from $\Gamma = 1$ alter the qualitative conclusions of the study, then the study is sensitive to hidden bias, but if only large values of Γ can alter the conclusions, then the study is insensitive. For example, Hammond's (1964) study of heavy smoking and lung cancer became sensitive at $\Gamma = 6$, and Herbst, Ulfelder, and Poskanzer's (1971) study of DES and vaginal cancer became sensitive at $\Gamma = 7$, so these studies are quite insensitive to hidden bias. In contrast, the study by Jick et al. (1973) of coffee and myocardial infarction is sensitive at $\Gamma = 1.3$. See Rosenbaum (1995, sec. 4) for details. Sensitivity to small hidden biases does not imply that such biases are present but implies only that small biases, if present, could materially alter the conclusions.

Table 2 is a sensitivity analysis for the combined analysis in Section 3.3 using the sum of the three signed rank statistics. The easy computational details of the procedure are described in Rosenbaum (1997). The significance level tabulated in Table 2 is the largest possible one-sided significance level for treatment assignment probabilities compatible with the given value of Γ. If this largest significance level is less than or equal to .05, then every possible set of treatment assignment probabilities compatible with this Γ yields a one-sided significance level less than .05. The bounds in Table 1 are sharp—they are attained for particular treatment assignment probabilities compatible with the stated value of Γ. In particular, they are attained for an unobserved u_{si} strongly associated with improved outcomes. For example, u_{si} might represent a patient attribute of prognostic value, visible to the attending physician but not recorded in the medical record.

In the combined, three-variable test, the null hypotheses of no treatment effect is just barely plausible, even in the absence of hidden bias, $\Gamma = 1$, where the one-sided significance level is .052, as in Section 3.3. The hypothesis of no treatment effect is entirely plausible in the presence of a small hidden bias, $\Gamma = 1.1$, where the significance level is .11, which is not recorded in Table 2. The hypothesis of a half-quartile effect, defined in Section 3.3, is insensitive to a bias of $\Gamma = 2.5$, where significance level is .046, but it is sensitive to $\Gamma = 3$. The hypothesis of a one-unit effect is insensitive to $\Gamma = 1.5$ but sensitive to $\Gamma = 2$. In short, the null hypothesis of no treatment effect is sensitive to small biases, $\Gamma = 1.1$, whereas the null hypothesis of a fairly large half-quartile effect is rejected even in the presence of moderately large biases, $\Gamma = 2.5$. Small biases, $\Gamma < 1.5$, could create the false impression that an ineffective treatment is slightly effective, but biases of that small size could not, in this study, create the false impression that a highly effective treatment was ineffective.

Consider the one variable, nocturnal frequency, for which a significant improvement was found in Section 3.3 by using

Table 2. Sensitivity Analysis for the Combined Test

	Null hypothesis		
Γ	*No effect*	*One-unit effect*	*Half iqr effect*
1.0	.052	.00018	<.00001
1.5	.5	.029	.00020
2.0		.218	.0066
2.5			.046
3.0			.15

NOTE: The tabulated values are sharp upper bounds on one-sided P-values for testing the three null hypotheses. The null hypothesis of no effect is tested against the alternative that treatment is beneficial. The hypotheses of a one-unit effect and a half iqr effect are tested against the alternative that the treatment has either no effect or a smaller effect than stated by the null hypothesis. The blank entries are not significant and are larger than the largest displayed P-value in the column.

the randomization distribution of the signed rank statistic. For nocturnal frequency, in the absence of hidden bias, $\Gamma = 1$, the one-sided significance level is .0016; however, the bound on the significance level is .059 for a bias of $\Gamma = 1.4$, so the ostensible improvement in nocturnal frequency is sensitive to a bias of modest size. In the absence of hidden bias, $\Gamma = 1$, a one-unit effect on nocturnal frequency is rejected as too large, with significance level of .0011; however, the bound on the significance level is .078 for $\Gamma = 1.5$. For nocturnal frequency, both tests—the test to detect an effect and the test to detect near equivalence—are quite sensitive to hidden bias. If a modest hidden bias of $\Gamma = 1.5$ is plausible, then the observed data for nocturnal frequency are consistent with either no treatment effect or a substantial one-unit effect, although neither hypothesis is plausible in the absence of hidden bias, $\Gamma = 1$.

Although this analysis suggests that the effects of treatment on nocturnal frequency are sensitive to moderate biases due to an unobserved covariate u_{si}, participants in the ICDB believe that the most important determinants of treatment and the best predictors of later symptoms are symptoms before treatment. In particular, there is tangible evidence that neither prior treatment nor bladder biopsy results are of incremental value (Propert et al. 2000, Rovner et al. 2000). Sensitivity to hidden bias is not evidence that bias is actually present but rather a measure of the magnitude of unobserved bias that would have to be present to alter conclusions.

The analyses in this section applied conventional methods for matched pairs to matches formed from risk sets. The formal discussion in Section 4 shows that these simple analyses are appropriate although the matching is based on time-dependent covariates.

4. INFERENCE IN RISK SET MATCHING

4.1 Risk Set Matching and Permutation Inference

In risk set matching, a treated patient is compared to an as-yet-untreated control who appeared similar in terms of observed covariates up to the moment before the treated patient received the treatment. Earlier sections depended on the informal, intuitive sense that such a comparison is reasonable. In this section, we show formally that risk set matching justifies simple, conventional permutation inferences under a general model for the decision to apply the treatment at a particular moment in response to time-varying symptoms. In other words, this section shows that the informal, intuitive sense that risk set matching produces reasonable comparisons is formally justified in terms of a model and specific methods of inference. To borrow a felicitous phrase from Susser (1973, sec. 7), optimal balanced risk set matching "simplifies the conditions of observation," comparing ostensibly comparable individuals at comparable moments, permitting simple comparisons and appropriate analyses by elementary methods.

The model says that the chance that patient m will receive the treatment at time t if the treatment has not been given up to time t—that is, the hazard of treatment—combines, in a proportional hazards model, an arbitrary function of the patient's observed symptom history with a multiple of an unobserved time-varying covariate describing this patient. The unobserved covariate expresses the possibility of hidden bias because the covariates we recorded are an incomplete record of the symptoms that determine treatment assignment.

Under this model, we obtain two conclusions. First, if hidden biases are absent so the unobserved covariate is irrelevant, then risk set matching produces matched pairs in which treatment assignments follow a randomization distribution. In this case, conventional methods, such as the signed rank test, produce appropriate inferences. Second, if hidden biases are present as expressed by the unobserved covariate, then the distribution of treatment assignments in matched pairs follows a familiar model for sensitivity analysis in observational studies, and this model may be applied directly. In other words, although the treatment was given in response to time-varying symptoms, risk set matching has simplified and restructured the problem in such a way that simple, conventional methods for matched pairs may be used. Because these conventional methods are standard and are described in existing journals and texts, it suffices here to show that risk set matching reproduces the distribution of treatment assignments which justifies their use. See Lehmann (1998) for discussion of nonparametric methods in randomized experiments, and see Rosenbaum (1995, sec. 4) for discussion of sensitivity analyses in observational studies.

This section is organized as follows. The model for treatment assignment in the unmatched population is defined in Section 4.2. The matched sampling of this population is defined in Section 4.3. The key result is Proposition 1 in Section 4.4. It says the distribution of treatment assignments in matched sets has a simple form, justifying simple, conventional permutation inferences. Although our study used matched pairs, the description permits matching with multiple controls. The ideas in this section benefit from and are related in spirit to those of Robins et al. (1992), Joffe et al. (1998), Robins (1999) and Keiding et al. (1999). However, the ideas developed here differ in several technical specifics, and because of the simplifications resulting from risk set matching, the methods are simpler and more conventional.

4.2 Notation: Effects of Treatment Delay

The population contains M patients, $m = 1, \ldots, M$, where patient m entered d_m months ago and may be treated at any one time $T_m \in [0, d_m]$ or not at all, signified by $T_m = c$, where c is censored. Following Neyman (1923) and Rubin (1974, 1977), each patient m has a potential response $\mathbf{r}_{tvm}$ that would be observed from patient m at time t if the patient received the treatment at time $v \in [0, d_m] \cup \{c\}$. In Section 3, $\mathbf{r}_{tvm}$ was three-dimensional and described pain, urgency, and frequency. To say that delaying treatment for patient m until $v = 6$ months would cause a one-unit increase in the response at $t = 9$ months in each of three coordinates of a trivariate response compared to starting treatment immediately with $v = 0$ is to say that $\mathbf{r}_{96m} - \mathbf{r}_{90m} = (1, 1, 1)^T$. Because for each t, only one of the potential responses $\mathbf{r}_{tvm}$ is observed for patient m, namely, $\mathbf{R}_{tm} = \mathbf{r}_{t, T_m, m}$, causal statements depend on inference about responses that would be observed under treatments not received. The null hypothesis of no treatment effect asserts that the response patient m exhibits at time t is the same as the control response, no matter when v the patient receives the

treatment; that is, it asserts

$$H_o : \mathbf{r}_{tvm} = \mathbf{r}_{tcm} \quad \text{for all} \quad t \in [0, d_m], \quad v \in [0, d_m] \cup \{c\}, \quad m = 1, \ldots, M.$$

As emphasized by Robins et al. (1992), a treatment applied at time v may have an effect at time v or after, but it has no effect at times t before v, so that $\mathbf{r}_{tvm} = \mathbf{r}_{tcm}$ for $v > t$. For this reason, time plays a unique role in structuring a study of this sort, a role quite different from that of covariates.

In randomization inference (Fisher 1935), patient m's potential responses, namely, $\langle \mathbf{r}_{tvm}, t \in [0, d_m], v \in [0, d_m] \cup \{c\} \rangle$, are fixed, but the treatment assignment T_m is a random variable, so patient m's observed response at time t, namely, $\mathbf{R}_{tm} = \mathbf{r}_{t, T_m, m}$, is also a random variable. To motivate the later discussion of observational studies of treatment delay, consider two possible randomized experiments. The simplest design for a study of treatment delay is to randomly assign half the patients to treatment immediately and half to a fixed delay, say, a delay of 6 months. This simplest design is a conventional clinical trial for which conventional randomization inference might be used. A less conventional randomized trial might randomly and independently pick treatment times for different patients from a single continuous distribution of treatment times, say, an exponential distribution with hazard λ. In both these randomized experiments, the fixed responses $\langle \mathbf{r}_{tvm}, t \in [0, d_m], v \in [0, d_m] \cup \{c\} \rangle$ that subject m might exhibit under different treatments do not alter the chances of receiving the treatments—that is the essence of random assignment.

The hazard of receiving treatment is modeled in terms of a proportional hazards model with time-varying covariates (Cox 1972), where some covariates are observed and recorded and one covariate is unobserved. For each $t \in [0, d_m]$, write $\mathbf{A}_{tm}$ for all the accumulated observed information about patient m until the instant before time t, and write $\mathbf{a}_{tm}$ for all the information that would have been observed about patient m before t if patient m had been assigned to control throughout the interval $t \in [0, d_m]$, so $\mathbf{A}_{tm} = \mathbf{a}_{tm}$ whenever $T_m \geq t$ or $T_m = c$. In addition to this observed information, there is a single unobserved variable U_{tm} whose value under control would have been u_{tm}, so also $U_{tm} = u_{tm}$ whenever $T_m \geq t$ or $T_m = c$.

Treatment decisions for distinct patients are assumed to be mutually independent, and are orderly in the sense that no two patients start treatment at exactly the same instant. Write h_{tm} for patient m's hazard of receiving treatment at time t given that patient m has not received treatment just before t, that is, $h_{tm} = \lim_{\delta \to 0} \text{prob}(t + \delta \geq T_m \geq t | T_m \geq t))\delta$, provided this limit exists. The following model is assumed:

$$h_{tm} = \exp\{\xi_t(\mathbf{a}_{tm}) + \gamma u_{tm}\}, \tag{1}$$

where $\xi_t(\cdot)$ is an unknown function for each t and γ is an unknown scalar parameter, called the sensitivity parameter. Notice that the hazard at time t describes someone with $T_m \geq t$, and for such a person $\mathbf{a}_{tm} = \mathbf{A}_{tm}$ is observed. Because the function $\xi_t(\cdot)$ can be any function at all, when $\gamma = 0$, the hazard of treatment at time t is any unknown function of observed data up to time t, that is, of $\mathbf{a}_{tm}$. Hence, when $\gamma = 0$, model (1) is the same as the model of no unmeasured confounders in the sense of Robins et al. (1992), which is the time-dependent version of Rubin's (1977) "randomization on the basis of a covariate" and of "strong ignorability" as discussed by Rosenbaum and Rubin (1983). Because γ need not equal 0, model (1) is more general than the model of no unmeasured confounders. When $\gamma \neq 0$, the unobserved time-dependent covariate u_{tm} may introduce an unobserved or hidden bias, as discussed by Rosenbaum (1987, 1995).

If the value of γ is to have meaning, the scale of the unobserved u_{tm} must be specified in some way. We assume $1 \geq u_{tm} \geq 0$, so (1) asserts that two subjects with identical observed covariate histories $\mathbf{a}_{tm}$ up to time t may differ in their hazards of treatment at time t by at most a multiplicative factor of $\Gamma = \exp(\gamma)$. For instance, u_{tm} might be an unobserved, time-varying, binary attribute. Other scale restrictions are discussed by Rosenbaum (1987), and they produce only small changes in the sensitivity analysis.

4.3 Matching on Observed Histories

At first, all patients are unmatched. At time t, let $L_t(\mathbf{a})$ be the set of unmatched patients who have a history of observed information equal to $\mathbf{a}$ up to time t and who did not receive the treatment an instant before t; that is, $L_t(\mathbf{a})$ is the subset of $\{1, 2, \ldots, M\}$ such that $m \in L_t(\mathbf{a})$ implies $T_m \geq t$ and $\mathbf{a}_{tm} = \mathbf{A}_{tm} = \mathbf{a}$, and m is unmatched at time t. Under model (1) for the hazard, if someone in $L_t(\mathbf{a})$ receives the treatment at time t, then the chance it is patient $m \in L_t(\mathbf{a})$ is $\exp(\gamma u_{tm})) \sum_{j \in L_t(\mathbf{a})} \exp(\gamma u_{tj})$, in parallel with Cox (1972). Matching will entail a partitioning of the risk set $L_t(\mathbf{a})$.

If a patient in $L_t(\mathbf{a})$ receives the treatment at time t, a matched set is formed, say, set number s with treatment time T_s, containing this newly treated patient and n_s as-yet-untreated patients also in $L_t(\mathbf{a})$. In our paired study, $n_s = 1$ for $s = 1, \ldots, 100$. Because all patients in $L_t(\mathbf{a})$ have identical observed covariate histories $\mathbf{a}_{tm} = \mathbf{A}_{tm} = \mathbf{a}$ a moment before t, the n_s matched controls are selected at random from $L_t(\mathbf{a})$. Patients in $L_t(\mathbf{a})$ have identical observed covariate histories but may differ in terms of the unobserved covariate u_{tm} which cannot be controlled by matching.

The $n_s + 1$ patients in matched set s are randomly assigned a second subscript i from $i = 1, \ldots, n_s + 1$, so i carries no information. Write $Z_{si} = 1$ if the ith patient in matched set s is the treated patient, and write $Z_{si} = 0$ otherwise, so $1 = \sum_{i=1}^{n_s+1} Z_{si}$ for each s. Write $\mathbf{a}_s$ for the value of $\mathbf{a}_{tm}$ common to all patients in matched set s, and write u_{si} for value of the unobserved covariate at the time T_s for the ith patient in matched set s, where the subscript t is dropped, because T_s is fixed within set s. Notice that when matched set s is formed at time t, as described in the previous paragraph, the $n_s + 1$ patients who form this matched set are taken out of $L_t(\mathbf{a})$; that is, for all $\varepsilon > 0$, the set $L_{t+\varepsilon}(\mathbf{a})$ does not include the $n_s + 1$ patients who were just matched. In this sense, $L_t(\mathbf{a})$ behaves differently from the risk set in Cox's proportional hazards model, which would exclude the newly treated patient at time $t + \varepsilon$ but not the n_s matched controls.

Write $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{1, n_1+1}, Z_{21}, \ldots, Z_{S, n_S+1})^T$. Then let $\mathcal{Z}$ be the set containing the $|\mathcal{Z}| = \prod_{s=1}^{S}(1 + n_s)$ possible values of $\mathbf{Z}$, that is, $\mathbf{z} \in \mathcal{Z}$ if and only if $\mathbf{z} = (z_{11}, \ldots, z_{S, n_S+1})^T$ with $z_{si} = 0$ or $z_{si} = 1$, $1 = \sum_{i=1}^{n_s+1} z_{si}$ for each s.

4.4 Treatment Assignment in Matched Sets

The following proposition says two things based on model (1) and the matching procedure described earlier. First, if there is no hidden bias in the assignment of treatments, in the sense that $\gamma = 0$, then the conditional distribution of treatment assignments within matched sets is a permutation or randomization distribution, and conventional methods of analysis, such as the signed-rank test or McNemar's test, may be used. Second, if there is hidden bias, then (1) leads directly to the sensitivity analysis model in Rosenbaum (1988, sec. 3; 1995, sec. 4), given by (2) to follow, which may again be used with the signed-rank or McNemar statistics. In other words, Proposition 1 asserts that, under model (1), conventional methods of analysis, such as those used in Section 3, may be used with pairs formed from risk set matching.

The form of expression (2) is commonly associated with the partial likelihood for Cox's proportional hazards model. However, here, because of the removal from the risk set of matched individuals, expression (2) is actually a conditional distribution as distinct from a partial likelihood; see Prentice and Breslow (1978, sec. 2) for a related conditional distribution. Write $\mathbf{L}$ for the set-valued stochastic function $L_t(\mathbf{a})$ of two arguments, t and $\mathbf{a}$.

Proposition 1. Under model (1), for each $\mathbf{z} \in \mathcal{Z}$,

$$\text{prob}(\mathbf{Z} = \mathbf{z}|\mathbf{L}) = \prod_{s=1}^{S} \frac{\exp(\gamma \sum_{i=1}^{n_s+1} z_{si} u_{si})}{\sum_{i=1}^{n_s+1} \exp(\gamma u_{si})}. \tag{2}$$

In particular, when there is no hidden bias from the unobserved covariate u_{si}, that is when $\gamma = 0$, for each $\mathbf{z} \in \mathcal{Z}$,

$$\text{prob}(\mathbf{Z} = \mathbf{z}|\mathbf{L}) = \frac{1}{|\mathcal{Z}|}. \tag{3}$$

Proof. The information in $\mathbf{L}$ provides exactly two types of information about the $n_s + 1$ patients in matched set s. First, it provides complete information about these $n_s + 1$ patients up to the instant before one of them received the treatment. Second, for all $\varepsilon > 0$, these $n_s + 1$ patients are absent from $L_{t+\epsilon}(\mathbf{a})$ for all $\mathbf{a}$. So $\mathbf{L}$ indicates that these $n_s + 1$ patients were matched at time T_s and that exactly one of these $n_s + 1$ patients received the treatment at time T_s, but $\mathbf{L}$ provides no other information about what happened to these $n_s + 1$ patients at time T_s and no information at all about these $n_s + 1$ patients after time T_s. Moreover, the treatment assignment at time T_s for these $n_s + 1$ patients is governed by (1). Matched set s contains one patient who received the treatment at time T_s and n_s other patients randomly sampled from $L_t(\mathbf{a})$, so the conditional probability that the ith patient in set s received the treatment given that one did is $\exp(\gamma u_{si})) \sum_{j=1}^{n_s+1} \exp(\gamma u_{sj})$. Moreover, given $\mathbf{L}$, the treatment assignment vectors $\mathbf{Z}_s$ for different matched sets s are independent.

The matching discussed in the proposition differs from the matching actually performed in Section 2.2, and the relationship between these two matchings merits brief discussion. Consider, for simplicity, the case of matched pairs, as in Section 2.2. The matching in the proposition would pair two individuals with identical covariate histories—one just treated, the other not yet treated. This type of matching is not practical with covariates of high dimension, but it permits a theoretical analysis. The matching in Section 2.2 was the closest matching that balanced marginal distributions of the covariates. Notice that exact matching on covariate histories, if feasible, would both balance the marginal distributions of covariates and be the closest matching in terms of these covariates. Speaking very informally, the matching in Section 2.2 is as close as we can get to the idealized exact matching discussed in Section 4.

Inspection of the proof of Proposition 1 shows that exact matching on the entire covariate history is not needed to obtain the result, for reasons that closely parallel permutation distributions obtained by using propensity scores (Rosenbaum 1984). Specifically, suppose patients were matched who did not have identical untreated multidimensional covariate histories $\mathbf{a}_{tm}$ up to time t, but instead had identical, unidimensional hazard components $\xi_t(\mathbf{a}_{tm})$ from these covariates; then the same distributions (1) and (2) are obtained by the same argument. In the proof, identical covariate histories served only to ensure identical hazard components $\xi_t(\mathbf{a}_{tm})$. For instance, in the absence of hidden bias, with $\gamma = 0$, it suffices to match patients with the same chance or hazard of receiving the treatment, $\exp\{\xi_t(\mathbf{a}_{tm})\}$, even if that same hazard reflects different covariate histories $\mathbf{a}_{tm}$.

4.5 Models for Effects

The discussion so far in this section emphasized testing the hypothesis of no treatment effect stated in Section 4.2. In the conventional way, as illustrated in Section 3, one can test hypotheses about additive effects by subtracting the hypothesized effects from treated subjects and testing that no effect remains. We did this in our equivalence tests, but the same approach yields confidence intervals by inverting the test; see Rosenbaum (1999b) for an example of such confidence intervals in sensitivity analysis. Additive models are reasonable in our application at 3 months because the not-yet-treated controls are still untreated at 3 months.

In some other study with a different structure, if many not-yet-treated controls had switched to treatment, then one might want to incorporate this information in modeling the treatment effect—that is, the constant effect model might be less reasonable than a model that used subsequent information about treatment. In this case, one might use the treatment decision at the time of matching as an instrument for the actual treatment received and would perform an instrumental variable analysis. In an instrumental variable analysis, the assigned treatment and the received treatment are not the same, and both variables play distinct roles in the inference. See Sheiner and Rubin (1995) and Angrist, Imbens, and Rubin (1996) for a conceptual discussion, and see Rosenbaum (1996, 1999b) for discussion of exact permutation inference and sensitivity analysis with an instrumental variable.

APPENDIX: OPTIMAL MATCHING AS INTEGER PROGRAMMING

This appendix writes the optimal balanced matching problem as a particular integer programming problem, that is, as a problem of

minimizing a linear function of integer variables, actually binary variables, subject to linear equality and inequality constraints. Unlike linear and network programming, where very fast algorithms exist, the speed at which a large integer programming problem is solved can depend, in a delicate way, on the way the problem is formulated. The formulation we describe has performed well for us in a substantial number of simulated and actual examples. The balance conditions are treated not as linear constraints but rather as penalty in the objective function. Expressed in this way, the solution to the integer programming problem is either an optimal balanced matching or a demonstration that no optimal balanced matching exits. With it expressed in this way, we used the package GAMS to obtain a solution. Specifics follow.

For each edge, $(\alpha_p, \alpha_q) = e \in \mathcal{E}$, introduce a binary flow variable $f_e = 1$ if α_p is matched to α_q and introduce $f_e = 0$ otherwise, and write $\mathbf{f}$ for the vector of $|\mathcal{E}|$ flows f_e, $e \in \mathcal{E}$. For each balance condition, introduce two gap variables, the positive gap g_{k+} and the negative gap g_{k-}, and write $\mathbf{g} = (g_{1+}, g_{1-}, g_{2+}, \ldots, g_{K+}, g_{K-})$. The gap variables will soon measure the degree of positive or negative departure from balance for the kth binary variable, and when perfect balance is obtained, $g_{k+} = g_{k-} = 0$. Let $\lambda_k > 0$ be a penalty, typically a large number, that will be paid when the kth binary variable is out of balance by 1. Specifically, take $\lambda_k > \sum_{e\in\mathcal{E}} \delta_e$, so the penalty for each binary variable is larger than the total of all the distances within pairs. Consider the following integer programming problem, called IP, and notice that motivation for this problem follows immediately after its statement:

$$\min_{\mathbf{f},\mathbf{g}} \sum_{e\in\mathcal{E}} f_e \cdot \delta_e + \sum_{k=1}^{K} \lambda_k (g_{k+} + g_{k-}) \tag{A.1}$$

subject to

$$S = \sum_{\{e:\, \in\mathcal{E}\}} f_e, \tag{A.2}$$

$$1 \geq \sum_{\{e\in\mathcal{E}:\, e=(\alpha_p,\alpha_q) \text{ or } e=(\alpha_q,\alpha_p)\}} f_e \quad \text{for} \quad \alpha_q \in \mathcal{A}, \tag{A.3}$$

$$g_{k+} \geq \sum_{e=(\alpha_p,\alpha_q)\in\mathcal{E}} f_e \cdot B_{pk} - \sum_{e=(\alpha_p,\alpha_q)\in\mathcal{E}} f_e \cdot B_{ek} \quad \text{for} \quad k = 1, \ldots, K, \tag{A.4}$$

$$g_{k-} \geq \sum_{e=(\alpha_p,\alpha_q)\in\mathcal{E}} f_e \cdot B_{ek} - \sum_{e=(\alpha_p,\alpha_q)\in\mathcal{E}} f_e \cdot B_{pk} \quad \text{for} \quad k = 1, \ldots, K, \tag{A.5}$$

$$g_{k+} \geq 0, \quad g_{k-} \geq 0 \quad \text{for} \quad k = 1, \ldots, K, \quad f_e \in \{0,1\} \quad \text{for} \quad e \in \mathcal{E}. \tag{A.6}$$

Notice the following. The objective function (A.1) is linear in the variables $\mathbf{f}$, $\mathbf{g}$, and (A.6) requires each flow f_e to be 0 or 1. Condition (A.2) says that there are S matched sets. Condition (A.3) says that for each fixed unit $\alpha_q \in \mathcal{A}$, at most one edge $e = (\alpha_p, \alpha_q) \in \mathcal{E}$ or $e = (\alpha_q, \alpha_p) \in \mathcal{E}$ has $f_e = 1$; i.e., each unit is either unmatched or in a single matched pair. Inequalities (A.4) and (A.5) imply that if $g_{k+} = g_{k-} = 0$, then the kth binary variable is perfectly balanced. Because (A.6) requires $g_{k+} \geq 0$ and $g_{k-} \geq 0$, if $g_{k+} + g_{k-}$ is made smaller, the imbalance in binary variable k is made smaller. Because we picked the penalties λ_k so that $\lambda_k > \sum_{e\in\mathcal{E}} \delta_e$ for $k = 1, \ldots, K$, it follows that even a single imbalance, $g_{k+} + g_{k-} = 1$, has a greater impact on the objective function (A.1) than does the distance within matched pairs.

Proposition A.1. Any solution $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$, if one exists, to *IP* is an optimal balanced matching if $\tilde{g}_{k+} = \tilde{g}_{k-} = 0$ for $k = 1, \ldots, K$, and otherwise no optimal balanced matching exists.

Proof. It is clear that any flow $\mathbf{f}$ describes a pair matching of size S if and only if conditions (A.2) and (A.3) are satisfied, so $\tilde{\mathbf{f}}$ is indeed a pair matching. It is also clear that any flow $\mathbf{f}$ that describes a pair matching will be a balanced pair matching if and only if it is possible to satisfy (A.4), (A.5), and (A.6) by taking $g_{k+} = g_{k-} = 0$ for $k = 1, \ldots, K$. So any balanced pair matching $\mathbf{f}$ can be written as a satisfying the constraints (A.2)–(A.6) with $(\mathbf{f}, \mathbf{g}) = (\mathbf{f}, \mathbf{0})$, in which case the objective function (A.1) is the total distance within pairs, $\sum_{e\in\mathcal{E}} f_e \cdot \delta_e$. It follows that a solution $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ to IP with $\tilde{g}_{k+} = \tilde{g}_{k-} = 0$ is a balanced pair matching that minimizes the total distance $\sum_{e\in\mathcal{E}} f_e \cdot \delta_e$ within pairs. If there is no solution to IP, then there is no pair matching, because every pair matching, $\mathbf{f}$, no matter how imbalanced, will satisfy the constraints (A.2)–(A.6) for some choice of $\mathbf{g}$. If there is a solution $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ to IP but it has $\tilde{g}_{k+} > 0$ or $\tilde{g}_{k-} > 0$ for some k, then it is imbalanced, and, moreover, any balanced pair matching would have a smaller value of the objective function (A.1), proving that no such balanced matching exists.

[Received July 1999. Revised October 2000.]

REFERENCES

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972), *Robust Estimates of Location*, Princeton, NJ: Princeton University Press.

Angrist, J., Imbens, G., and Rubin, D. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association,* 91, 444–469.

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

Bertsekas, D. P. (1991), *Linear Network Optimization,* Cambridge, MA: MIT Press.

Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley.

——— (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 74, 187–200.

——— (1975), "Partial Likelihood," *Biometrika*, 62, 269–276.

Curhan, G. C., Speizer, F. E., Hunter, D. J., Curhan, S. G., and Stampfer, M. J. (1999), "Epidemiology of Interstitial Cystitis: A Population Based Study," *Journal of Urology*, 161, 549.

Fisher, R. A. (1935), Design of Experiments. Edinburgh: Oliver and Bogel.

Gu, X. S., and Rosenbaum, P. R. (1993), "Comparison of Multivariate Matching Methods: Structures, Distances and Algorithms," *Journal of Computational and Graphical Statistics*, 2, 405–420.

Hammond, E. C. (1964), "Smoking in Relation to Mortality and Morbidity," *Journal of the National Cancer Institute*, 32, 1161–1188.

Hanno, P. M., Landis, J. R., Matthews-Cook, Y., Kusek, J., Nyberg, L., and Interstitial Cystitis Database Study Group (1999), "The Diagnosis of Interstitial Cystitis Revisited," *Journal of Urology*, 161, 553.

Herbst, A., Ulfelder, H., and Poskanzer, D. (1971), "Adenocarcinoma of the Vagina: Association of Maternal Stibestrol Therapy with Tumor Appearance in Young Women," *New England Journal of Medicine*, 284, 878–881.

Hollander, M., and Wolfe, D. A. (1999), *Nonparametric Statistical Methods*, New York: Wiley.

Hsu, J. C., Hwang, J. T. G., Liu, H. K., and Ruberg, S. J. (1994), "Confidence Intervals Associated With Tests for Bioequivalence," *Biometrika*, 81, 103–114.

Hunner, G. L. (1918), "A Rare Type of Bladder Ulcer," *Journal of the American Medical Association*, 70, 203.

Jick, H., Miettinen, O., Neff, R., Shapiro, S., Heinonen, O. P., and Sloan, D. (1973), "Coffee and Myocardial Infarction," *New England Journal of Medicine*, 289, 63–77.

Joffe, M. M., Hoover, D. R., Jacobson, L. P., Kingsley, L., Chmiel, J. S., Visscher, B. R., and Robins, J. M. (1998), "Estimating the Effect of Zidovudine on Kaposi's Sarcoma from Observational Data Using a Rank Preserving Structural Failure-Time Model," *Statistics in Medicine*, 17, 1073–1102.

Keiding, N., Filiberti, M., Esbjerg, S., Robins, J. M., and Jacobsen, N. (1999), "The Graft Versus Leukemia Effect After Bone Marrow Transplantation: A Case-Study Using Structural Nested Failure Time Models," *Biometrics*, 55, 23–28.

Langholz, B. and Goldstein, L. (1996), "Risk Set Sampling in Epidemiologic Cohort Studies" (with discussion), *Statistical Science*, 11, 35–53.

Lehmann, E. L. (1998), *Nonparametrics*, Upper Saddle River, NJ: Prentice-Hall.

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

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Reprinted in English in *Statistical Science*, 1990, 5, 463–480 (with discussion by T. Speed and D. Rubin).

Oakes, D. (1981), "Survival Times: Aspects of Partial Likelihood" (with discussion), *International Statistical Review,* 49, 235–264.

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

Pocock, S. J., and Simon, R. (1975), "Sequential Treatment Assignment With Balancing for Prognostic Factors in the Controlled Clinical Trial," *Biometrics*, 31, 103–116.

Prentice, R. L. (1986), "A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials," *Biometrika*, 73, 1–11.

Prentice, R. L., and Breslow, N. E. (1978), "Retrospective Studies and Failure Time Models," *Biometrika*, 65, 153–158.

Propert, K. J., Schaeffer, A., Brensinger, C., Kusek, J. W., Nyberg, L. M., Landis, J. R., and ICDB Study Group (2000), "A Prospective Study of Interstitial Cystitis: Results of Longitudinal Follow-up of the Interstitial Cystitis Database Cohort," *Journal of Urology*, 163, 1434–1439.

Robins, J. M. (1999), "Association, Causation, and Marginal Structural Models," *Synthese*, 121, 151–179.

Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992), "G-Estimation of the Effect of Prophylaxis Therapy for Pneumocystis Carinii Pneumonia on the Survival of AIDS Patients," *Epidemiology,* 3, 319–336.

Rosenbaum, P. R. (1984), "Conditional Permutation Tests and the Propensity Score in Observational Studies," *Journal of the American Statistical Association*, 79, 565–574.

——— (1987), "Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies," *Biometrika*, 74, 13–26.

——— (1988), "Sensitivity Analysis for Matching With Multiple Controls," *Biometrika*, 75, 577–581.

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

——— (1991), "A Characterization of Optimal Designs for Observational Studies," *Journal of the Royal Statistical Society*, Ser. B, 53, 597–610.

——— (1995), *Observational Studies*, New York: Springer.

——— (1996), Comment on "Indentification of Causal Effects Using Instrumental Variables by J. Angrist, G. Imbens and D. Rubin," *Journal of the American Statistical Association*, 91, 465–468.

——— (1997), "Signed Rank Statistics for Coherent Predictions," *Biometrics*, 53, 556–566.

——— (1999a), "Reduced Sensitivity to Hidden Bias at Upper Quantiles in Observational Studies With Dilated Treatment Effects," *Biometrics*, 55, 560–564.

——— (1999b), "Using Combined Quantile Averages in Matched Observational Studies," *Applied Statistics*, 48, 63–78.

——— (2001), "Effects Attributable to Treatment: Inference in Experiments and Observational Studies With a Discrete Pivot," *Biometrika*, 88, 219–231.

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

Rovner, E., Propert, K. J., Brensinger, C., Wein, A. J., Foy, M., Kirkemo, A., Landis, J. R., Kusek, J. W., Nyberg, L. M., and ICDB Study Group (in press), "Treatments Used in Women With Interstitial Cystitis: The Interstitial Cystitis Data Base (ICDB) Study Experience," *Urology*.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (1977), "Randomization on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1–26.

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

Sheiner, L. B., and Rubin, D. B. (1995), "Intention-to-Treat Analysis and the Goals of Clinical Trials," *Clinical Pharmacology and Therapeutics*, 57, 6–15.

Simon, L. J., Landis, J. R., Erickson, D. R., Nyberg, L. M., and ICDB Study Group (1997), "The Interstitial Cystitis Data Base Study: Concepts and Preliminary Baseline Descriptive Statistics," *Urology*, 49, 64–75.

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

Wei, L. J., and Lachin, J. M. (1984), "Two Sample Asymptotically Distribution Free Test for Incomplete Multivariate Observations," *Journal of the American Statistical Association*, 79, 653–661.