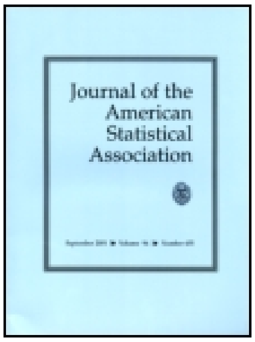

Journal of the
American
Statistical
Association

Journal of the American Statistical Association

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

Sensitivity Analysis for Equivalence and Difference in an Observational Study of Neonatal Intensive Care Units

Paul R. Rosenbaum & Jeffrey H. Silber

To cite this article: Paul R. Rosenbaum & Jeffrey H. Silber (2009) Sensitivity Analysis for Equivalence and Difference in an Observational Study of Neonatal Intensive Care Units, Journal of the American Statistical Association, 104:486, 501-511, DOI: 10.1198/jasa.2009.0016

To link to this article: http://dx.doi.org/10.1198/jasa.2009.0016

Published online: 01 Jan 2012.

Submit your article to this journal

Article views: 141

Citing articles: 4 View citing articles

Full Terms & Conditions of access and use can be found at
http://amstat.tandfonline.com/action/journalInformation?journalCode=uasa20

Download by: [Mount Sinai Health System Libraries] **Date:** 09 May 2017, At: 11:10

Sensitivity Analysis for Equivalence and Difference in an Observational Study of Neonatal Intensive Care Units

Paul R. ROSENBAUM and Jeffrey H. SILBER

In randomized experiments, it is sometimes important to demonstrate that two treatments do not differ greatly in their effects, or to demonstrate that the treatments have very different effects on one outcome but similar effects on another outcome. These ''demonstrations'' may take the form of rejecting a null hypothesis of inequivalence in an equivalence test, or rejecting a null hypothesis of equal and inequivalent effects on two outcomes. The procedures often express a complex hypothesis in terms of component hypotheses, and combine the component significance levels to test the complex hypothesis. If used in a randomized trial, randomization provides valid significance levels for each component test, and hence also for the combined procedure. In an observational study—that is, in a study of treatments that were not randomly assigned—there is typically concern that significance levels for testing hypotheses about treatment effects may be distorted by failure to control for some unobserved pretreatment covariate. This concern is raised in the evaluation of virtually all observational studies. The possible impact of such an unobserved covariate is clarified and displayed by a sensitivity analysis that, for various possible magnitudes of potential bias, yields a corresponding interval of possible significance levels. Some observational studies are sensitive to very small unobserved biases, whereas others are insensitive to large biases. Here, existing sensitivity analyses for component hypotheses are used to generate sensitivity analyses for complex hypotheses tested by combining component significance levels. We apply the procedure to our study of the timing of discharges of premature babies from neonatal intensive care units, focusing on the possible impact of delayed discharge on use of unplanned medical care after discharge.

KEY WORDS: Equivalence test; Risk-set matching; Superiority–equivalence test.

1. INTRODUCTION

In randomized clinical trials, it is common to test a complex hypothesis that is logically composed from simpler hypotheses. One instance is equivalence testing, in which a parameter θ measures a difference in treatment effects, and there is a prespecified magnitude $\delta > 0$ such that if $|\theta| < \delta$, then the difference is too small to be of concern. To provide evidence in favor of equivalence, in this sense of equivalence, is to reject the null hypothesis that $H_0 : \theta \leq -\delta$ or $\theta \geq \delta$ in a test against the alternative hypothesis $H_A : -\delta < \theta < \delta$. The two-one-sided-test (TOST) procedure of Schuirmann (1981) and Westlake (1981) tests H_0 at level α by testing $H_1 : \theta \leq -\delta$ with one-sided significance level P_1, testing $H_2 : \theta \geq \delta$ with one-sided significance level P_2, and rejecting H_0 at level α if max $(P_1, P_2) \leq \alpha$. The TOST procedure for equivalence testing is the simplest member of several large classes of test procedures that combine significance levels $P_1, \ldots, P_K$ for hypotheses $H_1, \ldots, H_K$ to test one or more logical consequences of $H_1, \ldots, H_K$, including Berger's (1982) ''intersection–union tests'' (as discussed also by Berger and Hsu (1996)), Bauer and Kieser's (1996) ''unifying approach,'' and ''testing hypotheses in order'' as discussed in Rosenbaum (2008) (see also Lehmann (1952) and Koch and Gansky (1996)). As discussed in these references, the K significance levels $P_1, \ldots., P_K$ may be dependent, because these procedures for combining significance levels make use of properties of the marginal distributions of P_k, $k = 1, \ldots., K$, rather than the joint distribution of $P_1, \ldots, P_K$. Use of randomization in an experiment ensures the validity of the individual significance levels, $P_1, \ldots, P_K$, (Fisher 1935; Cox and Reid 2000), and hence also the validity of the combined procedure.

In contrast, in an observational study of treatment effects, randomization is not used to assign treatments, so treated and control subjects may differ systematically prior to treatment in terms of covariates, and differing outcomes under different treatments may result from these pretreatment differences rather than from effects caused by the treatments. For pretreatment covariates that were measured, say $\mathbf{x}$, adjustments may be made, for instance by matching on $\mathbf{x}$ or by covariance adjustment for $\mathbf{x}$ (Cochran 1965). Invariably there is concern that some important covariate, say u, may not have been measured, so adjustments for $\mathbf{x}$ alone may fail to render the groups comparable (Rosenbaum 1991). Formally, even after adjustment, the significance levels $P_1, \ldots, P_K$ may not be valid tests of hypotheses $H_1, \ldots, H_K$ about treatment effects, so the combined procedure also may be invalid. For instance, unobserved bias might lead us to reject falsely the hypothesis of inequivalence, $H_0 : \theta \leq -\delta$ or $\theta \geq \delta$, leading to a mistaken conclusion that two treatments have similar effects when, in fact, they do not.

A sensitivity analysis asks how unobserved biases from u of various magnitudes might alter the conclusions of the study. The first sensitivity analysis by Cornfield et al. (1959) asked about the magnitude of unobserved bias that would need to be present to upset the conclusion from several observational studies that heavy smoking causes lung cancer. They concluded that an unobserved binary covariate u would need to be a near-perfect predictor of lung cancer and nine times more common among heavy smokers than nonsmokers to explain away the observed association between smoking and lung cancer as a

Paul R. Rosenbaum is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340 (E-mail: *rosenbaum@stat.wharton.upenn.edu*). Jeffrey H. Silber is Professor, Department of Pediatrics, University of Pennsylvania, Philadelphia, PA 19104-6340 (E-mail: *silberj@wharton.upenn.edu*). Supported by a grant from the Methodology, Measurement and Statistics Program of the U.S. National Science Foundation and by grant R40MC00236 from the Maternal and Child Health Bureau of the U.S. Public Health Service. We thank Gabriel J. Escobar, MD, Division of Research, Northern California Kaiser Permanente Medical Care Program, Oakland, CA, for his assistance in conducting this research.

© 2009 American Statistical Association
Journal of the American Statistical Association
June 2009, Vol. 104, No. 486, Applications and Case Studies
DOI 10.1198/jasa.2009.0016

bias due to u. Gastwirth (1992) extends the method in Cornfield et al. (1959), and Wang and Krieger (2006) argue that causal inferences are more sensitive to unobserved binary covariates than to other covariates with the same standard deviation.

A general approach to sensitivity analysis that is similar in spirit to the approach taken by Cornfield et al. (1959) measures the association between u and treatment assignment with an odds ratio parameter, Γ, and for each value of Γ determines observable sharp upper and lower bounds, $\overline{P}_{\Gamma k} \leq P_k \leq \overline{\overline{P}}_{\Gamma k}$, on unobservable significance levels P_k adjusted for u for tests of treatment effects (see Rosenbaum 1987, 1993, 2007a,b). If $\overline{\overline{P}}_{\Gamma k} \leq \alpha$, then a bias from u of magnitude Γ could not plausibly explain the rejection of H_k at level α. The bounds, $\overline{P}_{\Gamma k} \leq P_k \leq \overline{\overline{P}}_{\Gamma k}$, are sharp in the sense that they each are attained for some u, so the bounds cannot be narrowed except by assuming more about the unobserved u. Other methods of sensitivity analysis in observational studies are discussed by Cornfield et al. (1959); Rosenbaum and Rubin (1983); Gastwirth (1992); Marcus (1997); Lin, Psaty, and Kronmal (1998); Robins, Rotnitzky, and Scharfstein (1999); Copas and Eguchi (2001); Imbens (2003); Yu and Gastwirth (2005); and Tan (2006). For several applications of sensitivity analyses, see Aakvik (2001), Normand et al. (2001); Diprete and Gangl (2004); and Silber et al. (2005).

Suppose that we have calculated separate sensitivity bounds $\overline{P}_{\Gamma k} \leq P_k \leq \overline{\overline{P}}_{\Gamma k}$ for hypotheses H_k, $k = 1, \ldots, K$. What can be said about the sensitivity analysis for a method that combines significance levels to test one or more logical consequences of these hypotheses? An aspect of this question is that, although each bound $\overline{P}_{\Gamma k} \leq P_k \leq \overline{\overline{P}}_{\Gamma k}$ is sharp, being attained for some u, there may be no unobserved covariate u such that all K upper bounds are attained simultaneously. Nonetheless, for one large class of procedures, including intersection–union tests and tests of hypotheses in order, taking $P_k = \overline{\overline{P}}_{\Gamma k}$ for $k = 1, \ldots, K$ yields a sharp bound on the combined procedure; this is the conclusion of Proposition 1. In contrast, the sensitivity analysis for the Bonferroni inequality is not strict; as shown in Section 4.5, its bound is conservative, not sharp.

In Section 2, we discuss the observational study of neonatal intensive care units (NICUs) that motivated this work (Silber et al. 2008). In Section 3, we provide a brief review of procedures that test a complex hypothesis by decomposition in randomized experiments. Section 4 discusses sensitivity analyses for these procedures when used in observational studies, which are applied to a neonatal study in Section 5.

2. IS THERE BENEFIT FROM A FEW EXTRA DAYS IN AN NICU?

2.1 Prompt or Delayed Discharge Upon Achieving Functional Maturity

Premature newborn babies are kept in the NICU, where they are closely monitored, until they have matured sufficiently to go home. Typically, a baby must achieve various functional milestones and maintain them for several days before discharge. These milestones include maintenance of body temperature, coordinated sucking, swallowing and breathing while taking an adequate volume of feeding, sustained pattern of weight gain, and maturity and stability in cardiorespiratory function. Although virtually all babies must meet and maintain these functional milestones before discharge, there is some variability in how long babies remain in the hospital after these milestones have been maintained for several days. This variability raises a question: Are these additional days in the hospital that some babies receive and others do not of benefit to babies who receive them? Or would the costs of a longer hospital stay be better spent in some other way? Neonatologists often express concern that facilities outside hospitals are inadequate to assist parents of premature babies, so they keep babies an extra few days, but an alternative is to spend the cost of the additional hospital stay improving outpatient services for these babies. Are the extra days some babies receive of any benefit to them?

We studied this question using data provided by Kaiser Permanente Medical Care Program (KPMCP), a large, established, respected provider of health care primarily in California. KPMCP provided detailed longitudinal data on 1,402 premature babies born at KPMCP's hospitals who were discharged alive from the hospital. For premature infants, the traditional timescale of maturity is the baby's postmenstrual age (PMA), which starts the clock, not at birth, but at the end of the mother's last menstrual period prior to pregnancy. It is typically impossible to determine the date of conception, but PMA serves as a proxy for "age from conception." A full-term baby is born with a PMA of 39 weeks, whereas all babies in this study were born with a PMA of 34 weeks or less. Throughout this discussion, "time" refers to PMA—that is, to the baby's maturity. The data recorded some fixed variables, such as a baby's birthweight, Apgar score, PMA at birth, mother's income, mother's age, mother's race, and various medical problems at birth. During the hospital stay, the data also recorded daily scores of functional maturity (as described earlier), daily current weight, and daily PMA. At some point, the baby is discharged home.

Consider the $\binom{1,402}{2}$ possible pairs of two babies. Set aside the pairs in which both babies were discharged at the same PMA. For the remaining pairs, one baby was discharged earlier, with a lower PMA, than the other baby. On the day the earlier baby was discharged, we examined the covariates for the late baby, even though this was not the discharge day for the late baby, and we calculated a distance, defined later, that measured how similar the two babies were on the day the early baby was discharged. We wanted to find pairs of babies such that one baby was discharged earlier than the other, but the two babies looked very similar on the day the early baby was discharged, so that one could easily imagine either of these two babies would go home on that day.

Consider one baby. As noted earlier, on the day that this baby was discharged, the baby typically had achieved the functional milestones and had maintained them for several days. On the day this baby, the "early baby," went home, we found another baby who looked similar but who did not go home; this is the "late baby." On the day the early baby went home, the late baby had also achieved the functional milestones and had maintained them for several days, and on that day, the two matched babies were similar in many other respects. By KPMCP's typical practice, this late baby might have gone home on the day the early baby went home, but the late baby stayed typically a few days longer. The matching used a time-dependent propensity score (Li, Propert, and Rosenbaum 2001; Lu 2005), in which Cox's proportional hazards model was used to predict the day of discharge, where the model included both fixed and time-dependent variables, such as the daily values of the nine milestones. On the day the early baby went home, the two matched babies looked similar in that they had, on that day, a very similar estimated hazard of discharge, based on Cox's model. In addition to matching on the time-dependent propensity score, the matching procedure sought close matches on key covariates and current values of the time-dependent milestones.

As documented in detail in Silber et al. (2008), the early and late babies were similar on the day the early baby went home, but the late baby was measurably more mature on the late baby's own day of discharge. For instance, on the day of the early baby's discharge, the early babies weighed, on average, 2,153 g, whereas the late babies weighed, on average, 2,148 g —a difference of 5 g on average—but the late babies weighed, on average, 2,231 g, or 4% more (83 g more), a few days later on their own day of discharge. Similarly, late babies had been out of the incubator, off oxygen, and were feeding well for longer on their own day of discharge, but were similar to early babies when the early babies were discharged. The early and late babies were similar in terms of temporally fixed covariates, such as weight at birth, gestational age at birth, SNAP-II score at birth, gender, and various medical problems, such as necrotizing enterocolitis; and their mothers were similar in terms of age, race, ethnicity, income, marital status, and parity.

Technical details of the matching is briefly described here (see also Silber et al. 2008). In optimal nonbipartite matching, the 1,402 babies are divided into 701 nonoverlapping pairs of two distinct babies to minimize the sum of the 701 covariate distances within the 701 pairs. Derigs (1988) provided FORTRAN code for optimal nonbipartite matching, which Lu (2005) has made available from inside R. See Lu et al. (2001) and Lu and Rosenbaum (2004) for other uses of nonbipartite matching. The input to the matching algorithm is essentially a symmetric 1,402 × 1,402 distance matrix giving the distance between each pair of babies. We want the matched babies to be similar at the PMA when the early baby was discharged. If baby i and baby j were discharged at the same PMA, their distance was ∞. If baby i was discharged before baby j, so baby i would be the "early baby" if paired with baby j, then the distance between i and j describes baby i and baby j at the earlier PMA when baby i was discharged. The distance had several components, which are described in Silber et al. (2008), but essentially a caliper penalized large differences on the time-dependent propensity score, whereas the Mahalanobis distance sought babies who looked close in terms of important covariates. So, on the day the early baby went home, the late baby looked similar, and might have gone home, but actually stayed a few more days, and the late baby was measurably more mature (e.g., older, heavier) on the later day of the late baby's actual discharge. There are 701 pairs of two babies, an early baby and a late baby, who appeared similar on the day the early baby was discharged.

KPMCP keeps track of services used, such as days in the NICU, postdischarge visits to the emergency room, and so on. We used literature-based reference prices to convert emergency and sick-baby services into dollar costs, so two babies receiving the same services have the same costs by this formula. For a baby who has not yet been discharged, costs are dollars and cents, nothing more; the baby might be getting healthier as the costs increase. The situation changes upon discharge. A healthy baby will not have further hospital or sick-baby costs during the 6 months after discharge, so something has gone wrong if the baby has postdischarge costs—say, for a visit to the emergency room or a readmission to the intensive care unit. Both costs are important, but their meanings differ. For instance, one might reasonably judge an additional $1,000 spent during the original hospital stay as preferable if it prevented a subsequent $500 visit to the emergency room, even though the latter option is less expensive. For this reason, much of our analysis separates the two costs.

By definition, all the babies were discharged alive, but 5 of the 1,402 babies subsequently died after discharge. How should these five deaths be reflected in postdischarge costs? We use postdischarge costs as a measure of health problems after discharge, and for this reason, we record the five deaths as infinite costs. In our nonparametric analyses, an infinite cost is no different from any sufficiently large finite cost. See Rosenbaum (2006) for a formal discussion of this issue.

In thinking about the deaths, it may be useful to compare their frequency with the U.S. infant mortality rate in 2004, as recorded by the National Center for Health Statistics (2007). After discharge, during the subsequent 6 months, 5 of 1,402 babies died in this study, or 3.6 per thousand babies. For the United States as a whole in 2004, the infant mortality rate during the first year was 6.8 per thousand. It should be emphasized that these rates are not strictly comparable in several respects: (i) the U.S. rate is for the first year, whereas the study rate is for approximately 6 months after discharge; (ii) unlike the study, the U.S. rate consists of mostly full-term babies, but also includes some babies who died without being discharged from the hospital; and (iii) all KPMCP babies have health insurance, whereas many U.S. babies do not. With these several caveats firmly in mind, the mortality rate in this group does not seem strikingly high given that all babies in this study were premature and spent a significant time in the NICU.

Were the costly extra few days in the hospital that the late baby received of any value in reducing subsequent health problems? Or would the cost of those extra days have been better spent improving outpatient services for premature babies?

2.2 Elementary Analyses: Descriptive Statistics and Randomization Inferences

We measured cost in two time intervals: an initial or first interval from the early baby's discharge until the late baby's discharge, and a subsequent or second interval from late baby's discharge to 6 months after the early baby's discharge; both babies are observed for the same 6-month period defined by PMA. During the first interval, the early baby is home and the late baby is in the hospital. Both babies have been sent home by the start of the second interval. If the costs during the second interval were lower for the late baby, this might support keeping babies a few extra days. If the costs during the second interval were equivalent, it would be reasonable to find a better way to use the hospital costs accrued by the late baby during the first interval.

As one might expect, and as seen in Table 1, the distribution of costs is extremely right skewed. Although the babies are paired, Table 1 presents quantiles of the two marginal distributions, using the default definition of a quantile in R. The extreme right tail is of particular interest in the last two columns of Table 1, because these costs accrue after a baby has been discharged, so the long right tail identifies babies who had substantial health problems after discharge. The same information is displayed in the boxplots in Figure 1, but the log scale, necessary for plotting, takes attention away from the important right tail. Incidentally, the distribution of matched-pair differences during second-period costs is remarkably symmetric, albeit long tailed (see Fig. 2). In Table 1, during the first period, all late babies accrue substantial costs simply by virtue of being in the hospital. All but 8 of the 701 early babies ($8/701 = 1.1\%$) accrued less than $300 in costs during the first period, and those eight babies had costs ranging from $1,422 to $9,574. During the second period, the costs look similar. The middle of the distribution of total costs is dominated by the period 1 costs, whereas the right tail of the distribution of total costs is dominated by the period 2 costs for a small number of babies with extreme costs. There were three deaths among the early babies and two deaths among the late babies. Clearly, it is not possible to say much of a conclusive nature about the deaths alone, aside from saying that the death rate was low for both early and late babies; however, more than 6% of each group had significant health problems after discharge, in the form of death or second-period costs more than $5,000.

Table 2 presents standard randomization inferences, which are not strictly appropriate here because, although matched for important covariates, the discharge day was not randomly assigned to one baby in a pair. The first three rows of Table 2 are based on Wilcoxon's signed rank test for an additive effect. The typical difference in total costs (late-minus-early) is $5,016, and this is almost entirely from the first period; whereas, during the second period, the typical difference in cost is $17 with a 95% confidence interval of [−20, 56].

The final row of Table 2 refers to Stephenson's (1981) generalization of Wilcoxon's signed rank test, with his $m = 5$. As discussed in Rosenbaum (2007a), Stephenson's (1981) ranks are responsive to dramatic effects confined to a small fraction of the population—here, to bad outcomes for a small number of babies (see also Conover and Salsburg 1988). Stephenson's (1981) test is one sided, so Table 2 reports twice the smaller of the two one-sided significance levels. Even if one focuses attention on the highest costs, with deaths counting as infinite costs, there is no indication that the costs differed during the second period.

There is, of course, a natural concern about Table 2. Imagine an experiment in which one baby in each pair was picked at random to stay in the hospital a few extra days. If the data in Table 2 had come from this imagined randomized experiment, then the results would be quite convincing as they stand: The extra days increased costs in the hospital, but did not materially reduce subsequent costs. In reality, however, neonatologists decided to send one baby home later than the other. Why was this? We did find that a baby who reaches maturity on a weekend is likely to have discharge delayed until a subsequent weekday, and also there is some variation between hospitals in discharge patterns for most babies. Nonetheless, it is also possible that the neonatologists knew more about the two babies, or their mothers, than is available in the recorded data, and that in terms of an unmeasured covariate u, the sicker baby, or the baby with the less capable mother, was more likely to be kept in the hospital for an extra few days. It is also possible that these extra days had a beneficial effect on this sicker baby, reducing subsequent costs after discharge. Moreover, it is even possible that these two patterns—a sicker baby receiving a beneficial treatment—canceled out, to create the appearance of no effect on subsequent costs during period 2 in Table 2. What would an unobserved covariate u have to be like to create the appearance of a $17 effect when the true effect is quite beneficial?

Table 1. Calculated costs in dollars for 701 pairs of one early baby and one late baby

	Total	First + Second	First Interval		Second Interval	
Quantile	Early	Late	Early	Late	Early	Late
0 = 0%	0	2,698	0	2,562	0	0
1/4 = 25%	198	3,426	0	2,959	193	197
1/2 = 50%	326	5,011	0	4,387	307	332
3/4 = 75%	590	7,901	0	6,132	563	642
$7/8 \doteq 88\%$	2,357	10,399	34	8,551	2,239	2,934
$15/16 \doteq 94\%$	5,366	13,185	56	9,556	5,335	5,515
$31/32 \doteq 97\%$	10,917	20,093	112	10,459	10,917	13,080
$63/64 \doteq 98\%$	21,401	33,085	157	10,638	20,460	27,696
$127/128 \doteq 99\%$	41,018	49,599	1,665	10,792	36,631	38,452
1 = 100%	∞	∞	9,574	67,283	∞	∞
Deaths					3 = 0.4%	2 = 0.3%

NOTE: Matched babies appeared to be similar on the day the early baby was discharged; however, the late baby was discharged a few days later. The first interval or initial period is the time from discharge of the early baby to discharge of the late baby, so the early baby is home and the late baby is in the hospital. The second interval or subsequent period is the remainder of the first 6 months after the discharge of the early baby—a period when both babies are home.

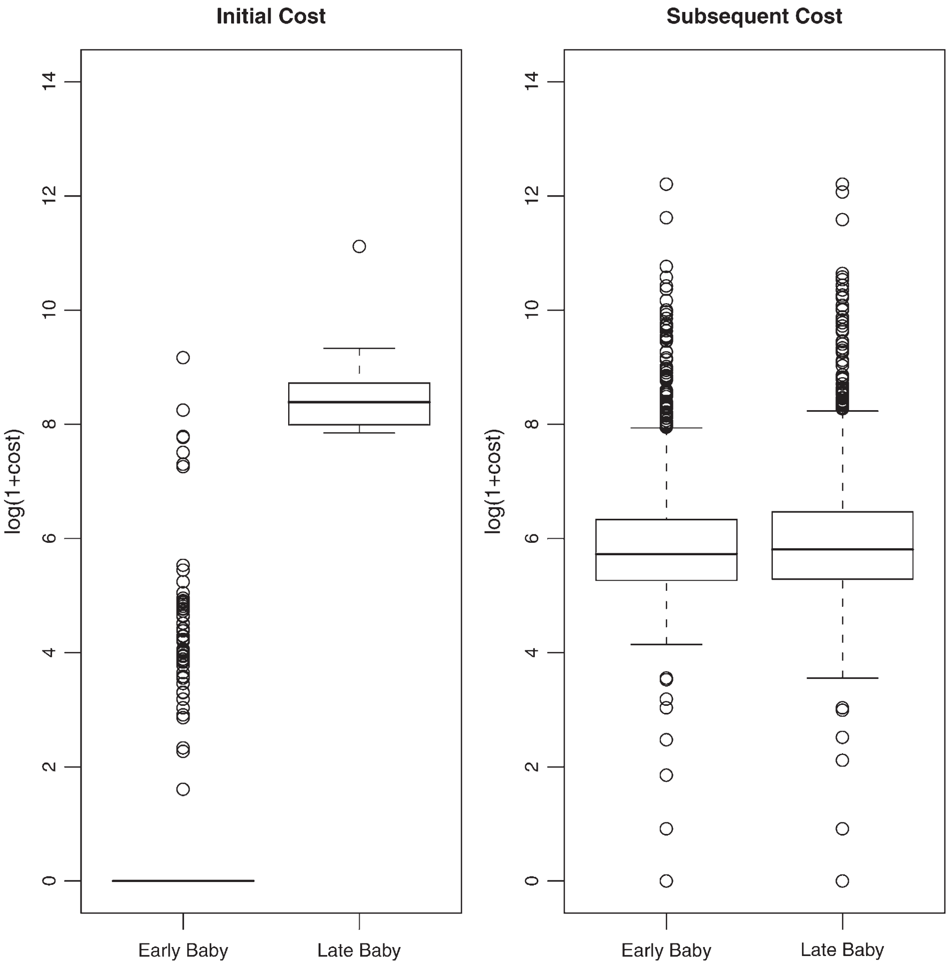

Figure 1. Initial and subsequent cost in dollars for 701 pairs of an early baby and a late baby. Scale is log(1 + cost). The early and late baby appeared similar on the day the early baby was discharged, but the late baby stayed in the hospital a few more days. The initial cost covers the time between the discharge of the early baby and the discharge of the late baby, while the early baby was at home. The subsequent cost covers the remainder of the 6 months after the discharge of the early baby, when both babies were home. To permit plotting, the five deaths are recorded as \$200,000—a round number somewhat larger than the largest second-period cost.

In the discussion of virtually every observational study, the possibility is raised that observed associations are the result of an unobserved covariate u and not an effect caused by the treatment. This possibility is not entirely open-ended: It is constrained, on one side, by the nature of the observed associations; and constrained, on the other side, by the logic of what an unobserved covariate would have to be like to produce such observed associations. A sensitivity analysis directs attention back from an abstract, open-ended possibility to a quantitative appraisal of the observed data. What would u have to be like to produce Table 2—that is, to produce a \$4,940 estimated difference in first-period costs and a \$17 estimated difference in second-period costs?

3. REVIEW: TESTS OF COMPLEX HYPOTHESES BY DECOMPOSITION

This section briefly reviews tests that decompose one or more decisions about a complex hypothesis into tests of simpler hypotheses. A general procedure is described, and then specific procedures are shown to be particular instances. Proposition 1 refers to the general procedure, and therefore to each of the specific cases.

There are K groups of hypotheses, $k = 1, \ldots, K$, L_k hypotheses $H_{k\ell}$ in the kth group, $\lambda = 1, \ldots, L_k$, and $L_+ = \sum_{k=1}^{K} L_k$ hypotheses in total. For each hypothesis $H_{k\ell}$, there is a valid significance level, $P_{k\ell}$, so that if $H_{k\ell}$ is true, then Pr $(P_{k\ell} \leq \alpha) \leq \alpha$ for each $\alpha \in [0, 1]$. Write $\mathcal{H} = \langle \{H_{11}, \ldots, H_{1,L_1}\}, \ldots, \{H_{K1}, \ldots, H_{K,L_K}\} \rangle$ for the ordered sequence of groups of hypotheses. Within each group k, there are logical relationships between the L_k hypotheses $\{H_{k1}, \ldots, H_{k,L_k}\}$; specifically, they are sequentially exclusive in the following sense: If all the hypotheses $\{H_{11}, \ldots, H_{1,L_1}\}, \ldots, \{H_{k-1,1}, \ldots, H_{k-1,L_{k-1}}\}$ are false, then at most one of the hypotheses in $\{H_{k1}, \ldots, H_{k,L_k}\}$ is true. Starting with $k = 1$, step k of the decomposition procedure is as follows:

Step k. Test each of the L_k hypothesis $H_{k\ell}$ in group k, rejecting $H_{k\ell}$ if $P_{k\ell} \leq \alpha$. If any hypothesis in $\{H_{k1}, \ldots, H_{k,L_k}\}$ is not rejected, then stop, and do not test hypotheses $H_{k'\ell}$ with $k' > k$. Otherwise, if all hypotheses in $\{H_{k1}, \ldots, H_{k,L_k}\}$ are rejected, then perform step $k + 1$.

In Rosenbaum (2008), it is shown that, for sequentially exclusive hypotheses, the chance that this procedure tests and rejects at least one true hypothesis is at most α. The proof resembles the proof for closed testing given by Marcus, Peritz,

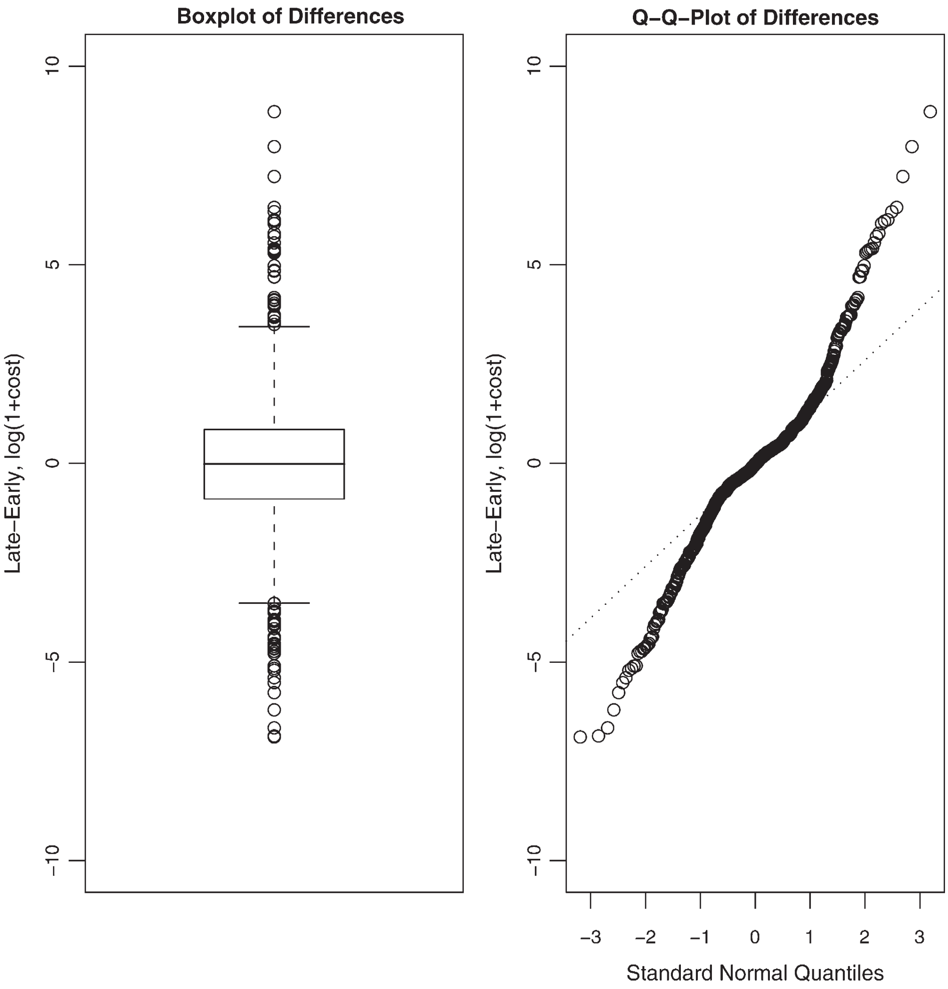

Figure 2. Matched-pair differences in second-period costs, on log(1 + cost) scale. The differences appear symmetric about zero, but even on the log scale are long tailed relative the normal distribution. To permit plotting, the five deaths are recorded as $200,000—a round number somewhat larger than the largest second-period cost. The dotted line in the quantile–quantile plot is fitted to the quartiles.

and Gabriel (1976), and for dose finding given by Bauer (1997) and Hsu and Berger (1999), although the procedure and the relationships among the hypotheses are different. Particular cases of the decomposition procedure follow.

Case 1: Equivalence Testing. As discussed in Section 1, in the TOST procedure, the null hypothesis $H_0: \; |\theta| \geq \delta$ or equivalently $H_0 : \theta \leq -\delta$ or $\theta \geq \delta$ is decomposed into $H = \langle\{H_{11}, H_{12}\}\rangle$, where $H_{11} : \theta \leq -\delta$ and $H_{12} : \theta \geq \delta$ are sequentially exclusive because at most one of H_{11} and H_{12} is true. The decomposition procedure is to reject $H_{1\ell}$ if $P_{1\ell} \leq \alpha$ for $\ell = 1, 2$ and, as a logical consequence, to reject $H_0 : \theta \leq -\delta$ or $\theta \geq \delta$ if both $P_{11} \leq \alpha$ and $P_{12} \leq \alpha$, in parallel with the TOST procedure.

Case 2: Intersection–Union Tests. Berger's (1982) intersection–union principle rejects $H_0 : H_1 \vee H_2 \vee \cdots \vee H_K$ at level α if $\max\{P_1, \ldots, P_K\} \leq \alpha$ (see also Lehmann 1952). This is the decomposition procedure with $\mathcal{H} = \langle\{H_1\}, \ldots, \{H_K\}\rangle$. The TOST procedure is a special case (see Berger and Hsu 1996).

Case 3: Superiority–Equivalence. With two outcomes, as in Section 2, we may ask: Is the new treatment superior with respect to the first outcome (say, $\theta_1 < 0$) and nearly equivalent on a second outcome (say, $|\theta_2| < \delta$) with specified $\delta > 0$? That is, the null hypothesis asserts $H_0 : (\theta_1 \leq 0) \vee (|\theta_2| \geq \delta)$, so rejecting H_0 provides evidence in favor of superiority on the first outcome and near equivalence on the second, which is evidence in favor of $H_A : (\theta_1{>}0) \wedge (|\theta_2| < \delta)$. Now H_0 is decomposed into $\mathcal{H} = \langle\{H_{11}\}, \{H_{21}, H_{22}\}\rangle$, where $H_{11} : \theta_1 \leq 0$, $H_{21} : \theta_2 \leq -\delta$, and $H_{22} : \theta_2 \geq \delta$, and H_{21} and H_{22} cannot both be true, so $\mathcal{H}$ is sequentially exclusive. Then H_{11} is not rejected if $P_{11} > \alpha$, in which case H_{21} and H_{22} are not tested. However, if $P_{11} \leq \alpha$, then H_{11} is rejected, H_{21} and H_{22} are both tested, and $H_{2\ell}$ is rejected if $P_{2\ell} \leq \alpha$, $\ell = 1, 2$, and H_0 is rejected if all three component hypotheses H_{11}, H_{21}, and H_{22} are rejected. Although formulated differently, Bauer and Kieser (1996) and Tamhane and Logan (2004) both discuss hypotheses that combine superiority and equivalence.

Case 4: Superiority–Superiority. With two outcomes, we may ask: Is the new treatment superior with respect to both outcomes? In this case, $H_0 : (\theta_1 \leq 0) \vee (\theta_2 \leq 0)$ and $H_A : (\theta_1{>}0) \wedge (\theta_2{>}0)$ (see Lehmann 1952). Define $H_{11} : \theta_1 \leq 0 \wedge \theta_2 \leq 0$, $H_{21} : \theta_1 \leq 0$, and $H_{22} : \theta_2 \leq 0$, so that $\mathcal{H} = \langle\{H_{11}\}, \{H_{21}, H_{22}\}\rangle$ is sequentially exclusive, because if H_{11} is false, then H_{21} and H_{22} cannot both be true. The procedure fails to reject H_{11} and stops if $P_{11} > \alpha$; otherwise, if $P_{11} \leq \alpha$, it rejects H_{11}, rejects $H_{2\ell}$ if $P_{2\ell} \leq \alpha$, $\ell = 1, 2$, and rejects H_0 if H_{11},

Table 2. Differences in costs in dollars for 701 matched pairs: (late baby) – (early baby)

Method	Total	First Period	Second Period
Hodges–Lehmann estimate	5,016	4,940	17
95% Confidence interval	[4,714, 5,235]	[4,485, 5,103]	[−20, 56]
p-Value from Wilcoxon's test	<10^{-10}	<10^{-10}	0.38
p-Value from Stephenson's test	<10^{-10}	<10^{-10}	0.21

NOTE: Deaths are recorded as infinite costs.

H_{21}, and H_{22} are all rejected. Similarly, if $\theta_\ell \geq 0$, $\ell = 1, 2$ is assumed a priori, then $H_0 : (\theta_1 = 0) \vee (\theta_2 = 0)$ can be tested against $H_A : (\theta_1 > 0) \wedge (\theta_2 > 0)$ using the sequentially exclusive partition $H = \langle\{H_{11}\}, \{H_{21}, H_{22}\}\rangle$, where $H_{11} : \theta_1 = \theta_2 = 0$ and $H_{2\ell} : \theta_\ell = 0$, $\ell = 1, 2$. This last procedure resembles the method in Lehmacher, Wassmer, and Reitmeir (1991, Fig. 2) that uses the closed testing approach of Marcus, Peritz, and Gabriel (1976).

The procedure may be represented by a function $\boldsymbol{\rho}\,(\cdot)$, with L_+ coordinates $\rho_{k\ell}(\cdot)$, $k = 1, \ldots, K$, $\ell = 1, \ldots, L_k$. Here, $\rho \; : \; [0,1]^{L_+} \rightarrow \{0,1\}^{L_+}$ maps the vector of L_+ significance levels into an L_+ dimensional vector of 1's and 0's, indicating whether particular hypotheses were tested and rejected. Specifically, let $\mathbf{p} = \left(p_{11}, \ldots, p_{K,L_K}\right) \in [0,1]^{L_+}$. For $k = 1, \ldots, K$, $\ell = 1, \ldots, L_k$, define the function $\rho_{k\ell} \; : \; [0,1]^{L_+} \rightarrow \{0,1\}$ as follows: $\rho_{k\ell}\,(\mathbf{p}) = 1$ if and only if $p_{k'\ell'} \leq \alpha$, for $k' = 1, \ldots, k - 1$, $\ell = 1, \ldots, L_k{}'$, and $p_{k\ell} \leq \alpha$. Let $\mathbf{P} = \left(P_{11}, \ldots, P_{K,L_K}\right)$ be the vector containing the L_+ significance levels; then the previous procedure tests and rejects $H_{k\ell}$ if and only if $\rho_{k\ell}\,(\mathbf{P}) = 1$; otherwise, if $\rho_{k\ell}\,(\mathbf{P}) = 0$, then the procedure either does not test $H_{k\ell}$ or tests $H_{k\ell}$ but does not reject it; in either case, $H_{k\ell}$ is not rejected.

4. SENSITIVITY ANALYSIS FOR DECOMPOSITION TESTS

4.1 Notation: Strata, Treatment Assignments, and Treatment Effects

There are $S \geq 1$ strata, $s = 1, \ldots, S$, defined by pretreatment covariates, with $n_s \geq 2$ subjects in stratum s, $i = 1, \ldots, n_s$, of whom m_s receive treatment, with $1 \leq m_s < n_s$, denoted $Z_{si} = 1$; and $n_s - m_s$ receive control, denoted $Z_{si} = 0$, so that $m_s = \sum_{i=1}^{n_s} Z_{si}$. In Section 2, there are $S = 701$ pairs, $n_s = 2$, with one late baby in each pair, $m_s = 1$, $s = 1, \ldots, 701$. Write $\mathbf{Z} = \left(Z_{11}, \ldots, Z_{S,n_S}\right)^T$*for the* $N = \sum n_s$ dimensional vector of treatment assignments, and write Ω for the set of possible values, $\mathbf{z}$, of $\mathbf{Z}$, so $\mathbf{z} \in \Omega$ if z_{ij} is 1 or 0 and $m_s = \sum_{i=1}^{n_s} z_{si}$ for each s. Write $|A|$ for the number of elements in a finite set A, so $|\Omega| = \prod_{s=1}^{S} \binom{n_s}{m_s}$. In Section 2, $N = 1{,}402$, $|\Omega| = \binom{2}{1}^S = 2^{701}$.

The ith subject in stratum j has two potential vector responses—namely, $\mathbf{r}_{Tsi}$ if treatment is assigned, $Z_{sj} = 1$; and $\mathbf{r}_{Csi}$ if control is assigned, $Z_{sj} = 0$, with observed response $\mathbf{R}_{si} = Z_{si}\mathbf{r}_{Tsi} + (1 - Z_{si})\,\mathbf{r}_{Csi}$ (see Neyman 1923 and Rubin 1974). In Section 2, $\mathbf{r}_{Tsi}$ is bivariate, recording the period 1 and period 2 cost of baby si if this baby is sent home late, and $\mathbf{r}_{Csi}$ is the pair of costs for the same baby if sent home early. The effect of the treatment on this subject is a comparison of $\mathbf{r}_{Tsi}$ and $\mathbf{r}_{Csi}$, such as $\mathbf{r}_{Tsi} - \mathbf{r}_{Csi}$, but because $\mathbf{r}_{Tsi}$ and $\mathbf{r}_{Csi}$ are never jointly observed, statements about treatment effects are inferences from the observable $(\mathbf{R}_{si}, Z_{si})$. Write $\mathbf{R}$, $\mathbf{r}_T$, and $\mathbf{r}_C$ for the matrices with N rows containing, respectively, the $\mathbf{R}_{si}$, $\mathbf{r}_{Tsi}$, and $\mathbf{r}_{Csi}$.

Each subject has an observed covariate $\mathbf{x}_{si}$ and an unobserved covariate u_{si}. The strata control for the observed covariate (that is, the strata are homogeneous in the observed covariate), so $\mathbf{x}_{si} = \mathbf{x}_{sj}$ for $1 \leq i < j \leq n_s$, $s = 1, \ldots, S$; but, it is not possible to control for the unobserved covariate, so typically $u_{si} \neq u_{sj}$. Write $\mathcal{F} = \{(\mathbf{r}_{Tsi}, \mathbf{r}_{Csi}, \mathbf{x}_{si}, u_{si}),\ i = \infty, \ldots, n_s,\ s = 1, \ldots, S\}$. Also, write $\mathcal{Z}$ for the event that $m_s = \sum_{i=1}^{n_s} Z_{si}, s = 1, \ldots, S$, or equivalently the event $\mathbf{Z} \in \Omega$. In a randomized experiment, the treatment assignment $\mathbf{Z}$ is picked at random from Ω, so that the design forces the event $\mathcal{Z}$ to occur with $\Pr(\mathbf{Z} = \mathbf{z} \,|\, \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ for each $\mathbf{z} \in \Omega$.

4.2 Randomization Inference in Randomized Experiments

Fisher (1935) emphasized that randomization forms the "reasoned basis for inference" about treatment effects in randomized experiments, in the following sense. The L_+ hypotheses $H_{k\ell}$ refer to different features of the response $(\mathbf{r}_{Tsi}, \mathbf{r}_{Csi})$. In the simplest case, $H_{k\ell}$ might refer only to the pth coordinate of the response—say, $\left(r_{Tsip}, r_{Csip}\right)$—perhaps asserting that this coordinate is unaffected; $H_{k\ell} : r_{Tsip} = r_{Csip}$, $\forall s, i$, or perhaps asserting the effect adds τ_p, so $H_{k\ell} : r_{Tsip} - \tau_p = r_{Csip}$, $\forall s, i$. Let $f_{Tk\ell(\cdot)}$ and $f_{Ck\ell}(\cdot)$ be two specified functions, possibly but not necessarily the same function, and consider the hypothesis $H_{k\ell} : f_{Tk\ell}(\mathbf{r}_{Tsi}) = f_{Ck\ell}(\mathbf{r}_{Csi}),\ \ \forall s,\ \ i$. Define $f_{k\ell}(\mathbf{R}_{si}, Z_{si}) = f_{Tk\ell}(\mathbf{r}_{Tsi})$ if $Z_{si} = 1$ and $f_{k\ell}(\mathbf{R}_{si}, Z_{si}) = f_{Ck\ell}(\mathbf{r}_{Csi})$ if $Z_{si} = 0$, noting that $f(\mathbf{R}_{si}, Z_{si})$ can always be computed from the observed data, because $\mathbf{R}_{si} = \mathbf{r}_{Tsi}$ if $Z_{si} = 1$ and $\mathbf{R}_{si} = \mathbf{r}_{Csi}$ if $Z_{si} = 0$. Write $\mathbf{F}_{k\ell} = \{f_{k\ell}(\mathbf{R}_{11}, Z_{11}),\ \ldots, f_{k\ell}\left(\mathbf{R}_{S,n_S}, Z_{S,n_S}\right)\}^T$ and $\mathbf{f}_{Ck\ell} = \left\{f_{Ck\ell}(\mathbf{r}_{C11}), \ldots, f_{Ck\ell}\left(\mathbf{r}_{CS,n_S}\right)\right\}^T$ for the N-dimensional vectors, so that *if* $H_{k\ell} : f_{Tk\ell}\ (\mathbf{r}_{Tsi}) = f_{Ck\ell}(\mathbf{r}_{Csi})$ were true, then $\mathbf{F}_{k\ell} = \mathbf{f}_{Ck\ell}$ is a function of $\mathcal{F}$ and is fixed by conditioning on $\mathcal{F}$. If $t_{k\ell}(\mathbf{Z}, \mathbf{F}_{k\ell})$ is a test statistic and $H_{k\ell} : f_{Tk\ell}(\mathbf{r}_{Tsi}) = f_{Ck\ell}(\mathbf{r}_{Csi})$, $\forall s$, i were true, then the null distribution of $t_{k\ell}(\mathbf{Z}, \mathbf{F}_{k\ell}) = t_{k\ell}(\mathbf{Z}, \mathbf{f}_{Ck\ell})$ is its randomization or permutation distribution

$$\Pr\{t_{k\ell}(\mathbf{Z}, \mathbf{f}_{Ck\ell}) \geq v \,|\, \mathcal{F}, \mathcal{Z}\} = \frac{|\{\mathbf{z} \in \Omega : t_{k\ell}(\mathbf{z}, \mathbf{f}_{Ck\ell}) \geq v\}|}{|\Omega|}. \tag{1}$$

For instance, in a familiar manner (e.g., Lehmann 1986, Section 5.14), the sharp null hypothesis $H_{k\ell} : r_{Tsip} - \tau_p = r_{Csip}$ of a specified additive effect τ_p is tested by calculating the significance level from (1) from $f_{k\ell}(\mathbf{R}_{si}, Z_{si}) = R_{sip} - \tau_p Z_{si}$, which equals $f_{Ck\ell}(\mathbf{r}_{Csi}) = r_{Csip}$ if $H_{k\ell}$ is true, and a $1 - \alpha$ confidence set for τ_p is obtained as the set of τ_p not rejected by this test. For certain hypotheses and test statistics, say for $H_{k\ell} : r_{Tsip} - \tau_p = r_{Csip}$ in matched pairs with $t_{k\ell}(\mathbf{Z}, \mathbf{f}_{Ck\ell})$ equal to Wilcoxon's signed rank statistic in Section 2, rejecting $H_{k\ell}$ in a one-tailed test may imply rejecting $H_{k\ell} : r_{Tsip} - \tau_p^* = r_{Csip}$ for all $\tau_p^* \leq \tau_p$, yielding one-sided confidence sets for τ_p that are half lines, or two-sided confidence sets that are intervals.

In a randomized experiment, (1) yields valid significance levels for use in the decomposition procedure in Section 3. In contrast, in an observational study, the distribution of treatment assignments $\Pr(\mathbf{Z} = \mathbf{z} \,|\, \mathcal{F}, \mathcal{Z})$ is not created by random assignment and is typically unknown. In particular, treatments Z_{ij} may tend to vary with unobserved covariates u_{ij} or the potential responses $(\mathbf{r}_{Tsi}, \mathbf{r}_{Csi})$ of subjects to treatments, so that (1) no longer yields a correct significance level for testing $H_{k\ell}$.

4.3 Sensitivity Analysis for a Single Significance Level

A sensitivity analysis in an observational study asks how departures from random treatment assignment, $\Pr(\mathbf{Z} = \mathbf{z} \,|\, \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$, of various magnitudes might alter the study's conclusions. A simple model for sensitivity analysis imagines that, in the population, subjects have independent treatment assignment with unknown probabilities, $\pi_{si} = \Pr(Z_{si} = 1|\mathcal{F})$ such that two subjects i and j who will end up in the same stratum because they have the same observed covariates—say, $\mathbf{x}_{si} = \mathbf{x}_{sj}$—may differ in theirs odds of treatment by at most a factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{si}\left(1-\pi_{sj}\right)}{\pi_{sj}(1-\pi_{si})} \leq \Gamma, \ \forall s, i, j, \tag{2}$$

and then returns attention to Ω by conditioning on $m_s = \sum_{i=1}^{n_s} Z_{si}$ for $s = 1, \ldots, S$ or equivalently conditioning on the event $\mathcal{Z}$. It is not difficult to verify that this is precisely the same as assuming there is an unobserved covariate u_{si} such that

$$\Pr(\mathbf{Z} = \mathbf{z} \,|\, \mathcal{F}, \mathcal{Z}) = \frac{\exp(\gamma \mathbf{z}^T \mathbf{u})}{\sum_{\mathbf{b} \in \Omega} \exp\left(\gamma \mathbf{b}^T \mathbf{u}\right)},$$
$$\text{for } \mathbf{z} \in \Omega, \text{with } 0 \leq u_{sj} \leq 1, \forall s, i, \tag{3}$$

where $\mathbf{u} = \left(u_{11}, \ldots, u_{S,n_S}\right)^T$ and $\gamma = \log(\Gamma) \geq 0$ (see Rosenbaum 2002, Section 4.2), where taking $u_{si} = \{\log(\pi_{si}) - \min_{1 \leq j \leq n_s} \log\left(\pi_{sj}\right)\}/\gamma$ establishes the equivalence of (2) and (3). Write $\mathcal{U} = [0, 1]^N$ for the N-dimensional unit cube. For each $\mathbf{u} \in \mathcal{U}$, (3) is distribution of $\mathbf{Z}$ on Ω. If $\Gamma = 1$ so $\gamma = 0$, then in (2), subjects with the same $\mathbf{x}$ have the same chance of receiving the treatment, and (3) becomes the randomization distribution, $\Pr(\mathbf{Z} = z \,|\, \Phi, Z) = 1/|\Omega|$. For each $\Gamma > 1$, the distribution of treatment assignments, $\Pr(\mathbf{Z} = Z \,|\mathcal{F}, \mathcal{Z})$, is unknown because u_{si} is unknown, but the departure from $1/|\Omega|$ is limited by the value of Γ. For each $\mathbf{u} \in \mathrm{U}$ and each Γ, (3) yields a significance level—say, $P_{\Gamma k\ell,\mathbf{u}}$—for $H_{k\ell} : f_{Tk\ell}(\mathbf{r}_{Tsi}) = f_{Ck\ell}(\mathbf{r}_{Csi})$, $\forall s$, i, because, if $H_{k\ell}$ is true, then $t_{k\ell}(\mathbf{Z}, \mathbf{F}_{k\ell}) = t_{k\ell}(\mathbf{Z}, \mathbf{f}_{Ck\ell})$, and so

$$P_{\Gamma k\ell,\mathbf{u}} = \Pr\{t_{k\ell}(\mathbf{Z}, \mathbf{f}_{Ck\ell}) \geq v \,|\, \mathcal{F}, \mathcal{Z}\}$$
$$= \frac{\sum_{\mathbf{z} \in \Omega} [t_{k\ell}(\mathbf{Z}, \mathbf{f}_{Ck\ell}) \geq v] \exp(\gamma \mathbf{z}^T \mathbf{u})}{\sum_{\mathbf{b} \in \Omega} \exp\left(\gamma \mathbf{b}^T \mathbf{u}\right)}$$

where $[t_{k\ell}(\mathbf{Z}, \mathbf{f}_{Ck\ell}) \geq v] = 1$ if $t_{k\ell}(\mathbf{Z}, \mathbf{f}_{Ck\ell}) \geq v$ and equals zero otherwise. For several values of $\Gamma \geq 1$, a sensitivity analysis determines sharp bounds $\overline{P}_{\Gamma k\ell} \leq P_{\Gamma k\ell,\mathbf{u}} \leq \overline{\overline{P}}_{\Gamma k\ell}$ on $P_{\Gamma k\ell,\mathbf{u}}$ in (4), yielding bounds on significance tests, and by inversion, bounds confidence intervals and point estimates (Rosenbaum 2002, Section 4). The bounds are sharp, not conservative, in the sense that each bound is attained for some $\mathbf{u} \in \mathcal{U}$; so, the bounds can be improved only by assuming more about the unobserved $\mathbf{u}$. For instance, in matched pairs, the upper bound for Wilcoxon's signed rank statistic is obtained from the distribution of the sum $\overline{\overline{T}}$ of S independent random variables, taking the value 0 with probability $1/(1 + \Gamma)$ and value s with probability $\Gamma/(1 + \Gamma)$, $s = 1, \ldots, S$, yielding the usual distribution of the signed rank statistic when $\Gamma = 1$. This also provides the sensitivity analysis for the associated Hodges–Lehmann point estimate (Rosenbaum 1993), and a similar approach works for M estimates (Rosenbaum 2007b).

Although Wilcoxon's signed rank test suffices in the example in Section 2, the randomization inferences in Section 4.2 and the sensitivity analyses in Section 4.3 are quite general (see Rosenbaum (2002, Section 2–5) for detailed development of most of the following situations). For instance, for binary responses, the methods include Fisher's exact test for a 2 × 2 table (where $S = 1$), McNemar's test for paired binary responses (where $n_s = 2$, $m_s = 1$, for $s = 1, \ldots, S$), and the Mantel–Haenszel test for a 2 × 2 × S table. For continuous responses, the methods apply to Wilcoxon's signed rank and rank sum tests, to the Hodges–Lehmann aligned rank test, to permutation tests associated with M estimates (Rosenbaum 2007b), and to studies with multivariate outcomes or doses of treatment (Rosenbaum 1997). In particular, in Case 4 in Section 3, the multivariate signed rank statistic in Rosenbaum (1997) might be used in the sensitivity analysis for H_{11}. For censored survival times, the methods apply to the permutation distributions of the log-rank and Gehan statistics, and to the O'Brien and Fleming test for censored matched pairs. An example illustrating a variety of randomization inferences is discussed by Small, Ten Have, and Rosenbaum (2008).

4.4 Sensitivity Analysis and Decomposition: A Strict Procedure

The decomposition procedure for randomized experiments in Section 3 was expressed in terms of an L_+ dimensional function, $\boldsymbol{\rho}(\mathbf{P})$, of the L_+ dimensional vector of significance levels, $\mathbf{P}$, where $\rho_{k\ell}(\mathbf{P}) = 1$ if the procedure tested and rejected $H_{k\ell}$, and $\rho_{k\ell}(\mathbf{P}) = 0$ if the procedure did not test $H_{k\ell}$ or tested but did not reject $H_{k\ell}$. In contrast, in the sensitivity analysis, the vector $\mathbf{P}_{\Gamma,\mathbf{u}} = \left\{P_{\Gamma 11,\mathbf{u}}, \ldots, P_{\Gamma K,L_K,\mathbf{u}}\right\}^T$ derived from (4) is unknown because $\mathbf{u}$ is unknown. There is a vector of upper bounds, $\overline{\overline{\mathbf{P}}}_\Gamma = \left(\overline{\overline{P}}_{\Gamma 11}, \ldots, \overline{\overline{P}}_{\Gamma K,L_K}\right)$, so that $P_{\Gamma k\ell,\mathbf{u}} \leq \overline{\overline{P}}_{\Gamma k\ell}$ for each k, ℓ and all $\mathbf{u} \in \mathrm{U}$. Do the bounds on individual significance levels yield a useful bound on the procedure as a whole? If so, are the bounds sharp or conservative?

For each k, ℓ there is a $\mathbf{u} \in \mathcal{U}$ such that $P_{\Gamma k\ell,\mathbf{u}} = \overline{\overline{P}}_{\Gamma k\ell}$, but typically this worst $\mathbf{u}$ will vary with k and ℓ, so typically there is no $\mathbf{u} \in \mathcal{U}$ that can produce $\overline{\overline{\mathbf{P}}}_\Gamma$—that is, no single $\mathbf{u} \in \mathcal{U}$ such that $P_{\Gamma k\ell,\mathbf{u}} = \overline{\overline{P}}_{\Gamma k\ell}$ for all k and ℓ. Despite this, and perhaps surprisingly, Proposition 1 shows that the decomposition procedure applied to the upper bounds, $\rho_{k\ell}\left(\overline{\overline{\mathbf{P}}}_\Gamma\right)$, does provide a sharp bound on the procedure applied separately at each $\mathbf{u} \in \mathcal{U}$—that is, on $\min_{\mathbf{u} \in \mathcal{U}} \rho_{k\ell}\left(\mathbf{P}_{\Gamma,\mathbf{u}}\right)$. That is, for each fixed $\Gamma \geq 1$, if $\boldsymbol{\rho}(\cdot)$ is evaluated at the (typically unattainable) joint bound $\overline{\overline{\mathbf{P}}}_\Gamma$, then for every $H_{k\ell}$ we learn whether there is some $\mathbf{u} \in \mathcal{U}$ such that, if this $\mathbf{u}$ were the true $\mathbf{u}$, then the decomposition procedure would fail to reject $H_{k\ell}$, with $\rho_{k\ell}(\mathbf{P}_{\Gamma,\mathbf{u}}) = 0$, or alternatively whether $H_{k\ell}$ is rejected for every $\mathbf{u} \in \boldsymbol{\mathcal{U}}$ with $\rho_{k\ell}(\mathbf{P}_{\Gamma,\mathbf{u}}) = 1$.

Proposition 1. For each $\Gamma \geq 1$, and for $k = 1, \ldots, K$, $\ell = 1, \ldots, L_k$,

$$\min_{\mathbf{u} \in \mathcal{U}} \rho_{k\ell}(\mathbf{P}_{\Gamma,\mathbf{u}}) = \rho_{k\ell}\left(\overline{\overline{\mathbf{P}}}_{\Gamma}\right). \tag{5}$$

Remark. Because $\boldsymbol{\rho}(\cdot)$ returns an L_+ dimensional vector with binary coordinates $\rho \ : \ [0,1]^{L_+} \to \{0,1\}^{L_+}$, expression (5) is equivalent to saying that (i) if $\rho_{k\ell}\left(\overline{\overline{\mathbf{P}}}_{\Gamma}\right) = 1$, then $\rho_{k\ell}(\mathbf{P}_{\Gamma,\mathbf{u}}) = 1$, $\forall \mathbf{u} \in \mathcal{U}$, but (ii) if $\rho_{k\ell}\left(\overline{\overline{\mathbf{P}}}_{\Gamma}\right) = 0$, then $\exists\, \mathbf{u} \in \boldsymbol{\mathcal{U}}$ such that $\rho_{k\ell}(\mathbf{P}_{\Gamma,\mathbf{u}}) = 0$. In words, $H_{k\ell}$ is tested and rejected for all $\mathbf{u} \in \mathcal{U}$ if and only if $\rho_{k\ell}\left(\overline{\overline{\mathbf{P}}}_{\Gamma}\right) = 1$. As a practical matter, this says that the sensitivity analysis for the decomposition procedure is straightforward: Compute the bounds $\overline{\overline{P}}_{\Gamma k\ell}$ separately for each k, ℓ, and apply the function $\boldsymbol{\rho}(\cdot)$ to $\overline{\overline{\mathbf{P}}}_{\Gamma}$.

Proof. Recall that $\rho_{k\ell}(\mathbf{p})$ is a function of $p_{k\ell}$ and the first $\sum_{m=1}^{k-1} L_m$ coordinates of $\mathbf{p}$, but not of other coordinates of $\mathbf{p}$; specifically, if $p_{k'\ell'} \leq \alpha$ for all $k' < k$ and all ℓ', then $\rho_{k\ell}(\mathbf{p}) = 1$ if $p_{k\ell} \leq \alpha$ and $\rho_{k\ell}(\mathbf{p}) = 0$ if $p_{k\ell} > \alpha$, whereas, if any $p_{k'\ell'} > \alpha$ for any $k' < k$ and any ℓ', then $\rho_{k\ell}(\mathbf{p}) = 0$. Say that $H_{k'\ell'}$ is earlier than $H_{k\ell}$ if $k' < k$ or if $k' = k$ and $\ell' < \ell$. In this order, suppose that $H_{\imath\jmath}$ is the earliest hypothesis such that $\overline{\overline{P}}_{\Gamma\imath\jmath} > \alpha$. To prove (5), three cases are considered separately—namely, $k < \imath, k = \imath$, and $k > \imath$. (i) In the first case, if $k < \imath$, then for all ℓ, and for all $\mathbf{u}$ $\in$ U, we have $\alpha \geq \overline{\overline{P}}_{\Gamma k\ell} = \max_{\mathbf{w} \in \mathcal{U}} P_{\Gamma k\ell,\mathbf{w}} \geq P_{\Gamma k\ell,\mathbf{u}}$, so that

$$\rho_{k\ell}(\mathbf{P}_{\Gamma,\mathbf{u}}) = \rho_{k\ell}\left(\overline{\overline{\mathbf{P}}}_{\Gamma}\right) = 1 \text{ for } k < \imath, \text{for all } \ell, \text{and for all } \mathbf{u} \in \mathcal{U}, \tag{6}$$

and hence $\min_{\mathbf{u} \in \mathcal{U}} \rho_{k\ell}(\mathbf{P}_{\Gamma,\mathbf{u}}) = \rho_{k\ell}\left(\overline{\overline{\mathbf{P}}}_{\Gamma}\right)$, so (5) is true when $k < \imath$. (ii) Second, consider $k = \imath$. As shown in (i), if $k' < k = \imath$, then $\rho_{k'\ell}(\mathbf{P}_{\Gamma,\mathbf{u}}) = \rho_{k'\ell}\left(\overline{\overline{\mathbf{P}}}_{\Gamma}\right) = 1$ for all ℓ and for all $\mathbf{u} \in \boldsymbol{\mathcal{U}}$. Because of this and the property recalled in the first sentence of the proof, for $\ell = 1, \ldots, L_\imath$, $\rho_{\imath\ell}(\mathbf{P}_{\Gamma,\mathbf{u}}) = 1$ if $P_{\Gamma\imath\ell,\mathbf{u}} \leq \alpha$ and $\rho_{\imath\ell}(\mathbf{P}_{\Gamma,\mathbf{u}}) = 0$ if $P_{\Gamma 1\ell,\mathbf{u}} > \alpha$. Similarly, $\rho_{\imath\ell}\left(\overline{\overline{\mathbf{P}}}_{\Gamma}\right) = 1$ if $\overline{\overline{P}}_{\Gamma\imath\ell} \leq \alpha$ and $\rho_{\imath\ell}\left(\overline{\overline{\mathbf{P}}}_{\Gamma}\right) = 0$ if $\overline{\overline{P}}_{\Gamma\imath\ell} > \alpha$. From this, it follows that $\min_{\mathbf{u} \in \mathcal{U}} \rho_{\imath\ell}(\mathbf{P}_{\Gamma,\mathbf{u}}) = 1$ if and only if $\max_{\mathbf{u} \in \mathcal{U}} P_{\Gamma\imath\ell,\mathbf{u}} \leq \alpha$, whereas $\rho_{\imath\ell}\left(\overline{\overline{\mathbf{P}}}_{\Gamma}\right) = 1$ if and only if $\overline{\overline{P}}_{\Gamma\imath\ell} \leq \alpha$, but these events are the same because $\max_{\mathbf{u} \in \mathcal{U}} P_{\Gamma\imath\ell,\mathbf{u}} = \overline{\overline{P}}_{\Gamma\imath\ell}$, so (5) is true when $k = \imath$. (iii) Finally, consider $k > \imath$. Because $\overline{\overline{P}}_{\Gamma\imath\jmath} > \alpha$, by what has just been established in (i) and (ii), it follows that $\min_{\mathbf{u} \in \mathcal{U}} \rho_{\imath\jmath}(\mathbf{P}_{\Gamma,\mathbf{u}}) = \rho_{\imath\jmath}\left(\overline{\overline{\mathbf{P}}}_{\Gamma}\right) = 0$. Let $\mathbf{v} \in \boldsymbol{\mathcal{U}}$ be such that $P_{\Gamma\imath\jmath,\mathbf{v}} = \max_{\mathbf{u} \in \mathcal{U}} P_{\Gamma\imath\jmath,\mathbf{u}} = \overline{\overline{P}}_{\Gamma\imath\jmath} > \alpha$; then for $k > \imath$, $0 = \rho_{k\ell}(\mathbf{P}_{\Gamma,\mathbf{v}})$, so $0 = \min_{\mathbf{u} \in \mathcal{U}} \rho_{k\ell}(\mathbf{P}_{\Gamma,\mathbf{u}})$, and also $0 = \rho_{k\ell}\left(\overline{\overline{\mathbf{P}}}_{\Gamma}\right)$, so (5) is true when $k > \imath$. ■

4.5 Sensitivity of the Bonferroni Inequality: A Conservative Procedure

To gain insight into Proposition 1, consider a different procedure for testing many hypotheses for which the conclusion of Proposition 1 fails to hold. The simplest procedure for testing many hypotheses uses the Bonferroni inequality, and in a sensitivity analysis it would not reject any of the L_+ hypotheses if

$$\max_{\mathbf{u} \in \mathcal{U}} \min_{k,\ell} P_{\Gamma k\ell,\mathbf{u}} > \frac{\alpha}{L_+}. \tag{7}$$

However, the L_+ separate upper bounds, $\overline{\overline{P}}_{\Gamma k\ell} = \max_{\mathbf{u} \in \mathcal{U}} P_{\Gamma k\ell,\mathbf{u}}$ do not determine whether (7) is true; rather, they are a conservative, not strict, guide to whether (7) is true, because, by a familiar theorem (e.g., Karlin 1992, vol. II, Lemma 1.3.1),

$$\max_{\mathbf{u} \in \mathcal{U}} \min_{k,\ell} P_{\Gamma k\ell,\mathbf{u}} \leq \min_{k,\ell} \max_{\mathbf{u} \in \mathcal{U}} P_{\Gamma k\ell,\mathbf{u}} = \min_{k,\ell} \overline{\overline{P}}_{\Gamma k\ell}, \tag{8}$$

where strict inequality is possible. That is, one can have $\min_{k,\ell} \overline{\overline{P}}_{\Gamma k\ell} > \alpha/L_+$, so use of the separate upper bounds, $\overline{\overline{P}}_{\Gamma k\ell}$, leads to acceptance of all L_+ hypotheses, and yet $\min_{k,\ell} P_{\Gamma k\ell,\mathbf{u}} \leq \alpha/L_+$ for all $\mathbf{u} \in \boldsymbol{\mathcal{U}}$, so that no $\mathbf{u} \in \boldsymbol{\mathcal{U}}$ would lead to acceptance of all L_+ hypotheses. On the other hand, by (8), if $\min_{k,\ell} \overline{\overline{P}}_{\Gamma k\ell} \leq \alpha/L_+$, then (7) is false; moreover, if $\overline{\overline{P}}_{\Gamma k\ell} \leq \alpha/L_+$, then $\max_{\mathbf{u} \in \mathcal{U}} P_{\Gamma k\ell,\mathbf{u}} \leq \alpha/L_+$ and $H_{k\ell}$ would be rejected for every $\mathbf{u} \in \boldsymbol{\mathcal{U}}$. In short, applying the Bonferroni inequality to the separate upper bounds, $\overline{\overline{P}}_{\Gamma k\ell}$, is conservative, possibly failing to reject some hypotheses that should be rejected, whereas by Proposition 1, applying the decomposition procedure to the separate upper bounds, $\overline{\overline{P}}_{\Gamma k\ell}$, rejects precisely the hypotheses that should be rejected.

5. EQUIVALENCE AND DIFFERENCE IN NICU'S

In Section 2, pairs of premature babies were compared who appeared similar at a certain moment when one baby was discharged home and the other remained in the hospital for a few more days. Table 1 showed that the initial hospital costs and total costs were much higher for babies discharged later, and there was no indication of lower postdischarge emergency or sick-baby costs for these babies when compared with the early babies. Analyses in Table 2 that pretended the matched pairs had come from a randomized experiment suggested a negligible difference in postdischarge costs during the second period, with a point estimate of \$17 and a 95% confidence interval of [−\$20, \$56]. The worry was that the matching failed to control for some variable u that the neonatologist may have used in determining when to discharge a baby. What would u have to be like to alter the naive impression from Table 2?

We address this question using a sensitivity analysis for equivalence and difference, as discussed in Section 3 and as implemented using Proposition 1 applied to Wilcoxon's signed rank test for the $S = 701$ pairs. The hypotheses are expressed in terms of additive treatment effects, discussed in Section 4.2, specifically τ_1 for period 1 costs, when the late baby is in the hospital and the early baby is home, and τ_2 for the remainder of the 6-month period after discharge of the late baby. The null hypothesis asserts $H_0 : (\tau_1 \leq 0) \vee (|\tau_2| \geq \delta)$ for some specified $\delta > 0$, so rejection of H_0 provides strong evidence that the late baby cost more during the initial period, $\tau_1 > 0$, and differed by less than δ in emergency and sick costs during the second period, $-\delta < \tau_2 < \delta$. As in Section 3, this hypothesis is decomposed into $\mathcal{H} = \langle\{H_{11}\}, \{H_{21}, H_{22}\}\rangle$, where $H_{11} : \tau_1 \leq 0$, $H_{21} : \tau_2 \leq -\delta$, and $H_{22} : \tau_2 \geq \delta$, and rejection of H_{11} is a prerequisite for testing H_{21} and H_{22}; then H_0 is rejected if H_{11}, H_{21}, and H_{22} are all rejected.

We consider three choices of δ. In the data, an uneventful day in the hospital for these babies cost approximately \$1,000. We considered $\delta = \$500$, which is substantially less than a day in the hospital; $\delta = \$1{,}000$, which is about the cost of a day in the hospital; and $\delta = \$2{,}500$, which is (from Table 1) roughly

Table 3. Sensitivity analysis for difference of initial cost (τ_1) and equivalence of subsequent costs (τ_2)

Γ	$H_0 : \tau_1 \leq 0 \vee \|\tau_2\| \geq \$500$	$H_0 : \tau_1 \leq 0 \vee \|\tau_2\| \geq \$1,000$	$H_0 : \tau_1 \leq 0 \vee \|\tau_2\| \geq \$2,500$
1	<0.00001	<0.00001	<0.00001
2	0.002	<0.00001	<0.00001
3	0.90	0.14	<0.00001
5	1.00	1.00	0.002

half the \$5,000 estimated cost of a later discharge. As emphasized in Section 2, costs that build up before discharge are just costs, whereas emergency or sick costs after discharge indicate that something has gone wrong, regardless of whether that something was avoidable—that is, the costs in the two periods mean different things. On this basis, one might reasonably argue that $\delta = \$500$ is more appropriate than $\delta = \$2{,}500$ as a definition of equivalence.

Consider testing $H_0 : (\tau_1 \leq 0) \vee (|\tau_2| \geq \$500)$ using randomization tests—that is, using (1) or using (4) with $\Gamma = 1$—that is, using the Wilcoxon signed rank test in a conventional manner three times to test $H_{11} : \tau_1 \leq 0$, $H_{21} : \tau_2 \leq -\$500$, and $H_{22} : \tau_2 \geq \$500$. In this case, $P_{11} < 0.00001$, so H_{11} is rejected at $\alpha = 0.05$, H_{21} and H_{22} are both tested with $P_{21} < 0.00001$ and $P_{22} < 0.00001$, so H_0 is also rejected.

A bias of magnitude $\Gamma = 2$ means that two matched babies might differ in terms of an unobserved u such that one baby is twice as likely as the other to be discharged late. This u might have any possible relationship with period 1 or period 2 costs. For $\Gamma = 2$, the same procedure could be carried through using (4) to test H_0 for any one fixed $\mathbf{u} \in \mathcal{U}$, and Proposition 1 provides a straightforward way to determine whether there is any $\mathbf{u} \in \mathcal{U}$ that would lead to acceptance of H_0. Using the upper bound $\overline{\overline{T}}$ in Section 4.3 with $\Gamma = 2$ gives $\overline{\overline{P}}_{211}$<0.00001, $\overline{\overline{P}}_{221}$<0.00001, $\overline{\overline{P}}_{222}$<0.0020, so $H_0 : (\tau_1 \leq 0) \vee (|\tau_2| \geq \$500)$ would be rejected for every $\mathbf{u} \in \mathcal{U}$ when $\Gamma = 2$.

Table 3 summarizes the sensitivity analysis for several Γ and δ. For all cases in Table 3, $\overline{\overline{P}}_{\Gamma 11} < 0.00001$, so H_{21} and H_{22} are tested. A bias of magnitude $\Gamma = 3$ could lead to acceptance of $H_0 : (\tau_1 \leq 0) \vee (|\tau_2| \geq \$500)$, but a bias of magnitude $\Gamma = 5$ would lead to rejection of $H_0 : (\tau_1 \leq 0) \vee (|\tau_2| \geq \$2,500)$ for every $\mathbf{u} \in \mathcal{U}$. In this case, the conclusions from the naive analysis in Table 2 are insensitive to moderately large biases, and only a very large bias, much larger than $\Gamma = 5$, could mask a net cost savings, $\tau_2 < -\$2{,}500$, from discharging late, which is about half the estimated savings from discharging early.

In short, discharging later is associated with substantially higher costs in the hospital, yet there is no indication of subsequently better outcomes or lower costs for babies discharged later. Although this pattern of associations could be produced by bias from an unobserved covariate, that covariate would need to more than double the odds of late discharge to mask a \$500 difference in second-period costs.

6. SUMMARY AND DISCUSSION

Proposition 1 provides a simple method of sensitivity analysis for complex hypotheses, such as equivalence–difference hypotheses, that can be decomposed into several simple hypotheses, as in Section 3. Essentially, if one can perform a sensitivity analysis for each of the component hypotheses, then the sensitivity analysis for the composite hypothesis follows immediately. The method applies to various types of complex hypotheses and outcomes, as discussed in Section 3 and Section 4.3.

[Received December 2007. Revised May 2008.]

REFERENCES

Aakvik, A. (2001), "Bounding a Matching Estimator: The Case of a Norwegian Training Program," *Oxford Bulletin of Economics and Statistics,* 63, 115–143.

Bauer, P. (1997), "A Note on Multiple Testing Procedures in Dose Finding," *Biometrics,* 53, 1125–1128.

Bauer, P., and Kieser, M. (1996), "A Unifying Approach for Confidence Intervals and Testing of Equivalence and Difference," *Biometrika,* 83, 934–937.

Berger, R. L. (1982), "Multiparameter Hypothesis Testing and Acceptance Sampling," *Technometrics,* 24, 295–300.

Berger, R. L., and Hsu, J. C. (1996), "Bioequivalence Trials, Intersection–Union Tests and Equivalence Confidence Sets," *Statistical Science,* 11, 283–319.

Cochran, W. G. (1965), "The Planning of Observational Studies of Human Populations," *Journal of the Royal Statistical Society* Ser. A128, 134–155.

Conover, W. J., and Salsburg, D. S. (1988), "Locally Most Powerful Tests for Detecting Treatment Effects When Only a Subset of Patients Can be Expected to 'Respond' to Treatment," *Biometrics,* 44, 189–196.

Copas, J., and Eguchi, S. (2001), "Local Sensitivity Approximations for Selectivity Bias," *Journal of the Royal Statistical Society* Ser. B, 63, 871–896.

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

Cox, D. R., and Reid, N. (2000), *Theory of the Design of Experiments,* New York: Chapman and Hall/CRC.

Derigs, U. (1988), "Solving Nonbipartite Matching Problems by Shortest Path Techniques," *Annals of Operations Research,* 13, 225–261.

Diprete, T. A., and Gangl, M. (2004), "Assessing Bias in the Estimation of Causal Effects," *Sociological Methodology,* 34, 271–310.

Fisher, R. A. (1935), *Design of Experiments,* Edinburgh Oliver and Boyd.

Gastwirth, J. L. (1992), "Methods for Assessing the Sensitivity of Statistical Comparisons in Title VII Cases to Omitted Variables," *Jurimetrics,* 33, 19–34.

Hsu, J. C., and Berger, R. L. (1999), "Stepwise Confidence Intervals Without Multiplicity Adjustment for Dose–Response and Toxicity Studies," *Journal of the American Statistical Association,* 94, 468–475.

Imbens, G. W. (2003), "Sensitivity to Exogeneity Assumptions in Program Evaluation," *The American Economic Review,* 93, 126–132.

Karlin, S. (1992), *Mathematical Methods and Theory in Games, Programming and Economics,* New York: Dover.

Koch, G. G., and Gansky, S. A. (1996), "Statistical Considerations for Multiplicity in Confirmatory Protocols," *Drug Information Journal,* 30, 523–533.

Lehmacher, W., Wassmer, G., and Reitmeir, P. (1991), "Procedures for Two-Sample Comparisons With Multiple Endpoints Controlling the Experimentwise Error Rate," *Biometrics,* 47, 511–521.

Lehmann, E. L. (1952), "Testing Multiparameter Hypotheses," *Annals of Mathematical Statistics,* 23, 541–552.

———. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: Wiley.

Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998), "Assessing the Sensitivity of Regression Results to Unmeasured Confounders," *Biometrics,* 54, 948–963.

Li, Y. P., Propert, K. J., and Rosenbaum, P. R. (2001), "Balanced Risk Set Matching," *Journal of the American Statistical Association,* 96, 870–882.

Lu, B. (2005), "Propensity Score Matching With Time-Dependent Covariates," *Biometrics,* 61, 721–728; http://cph.osu.edu/divisions/biostatistics/.

Lu, B., and Rosenbaum, P. R. (2004), "Optimal Matching With Two Control Groups," *Journal of Computational and Graphical Statistics,* 13, 422–434.

Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. R. (2001), "Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse," *Journal of the American Statistical Association,* 96, 1245–1253.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976), "On Closed Testing Procedures With Special Reference to Ordered Analysis of Variance," *Biometrika,* 63, 655–660.

Marcus, S. M. (1997), "Using Omitted Variable Bias to Assess Uncertainty in the Estimation of an AIDS Education Treatment Effect," *Journal of Educational and Behavioral Statistics,* 22, 193–201.

National Center for Health Statistics. (2007), "Overall Infant Mortality Rate in U.S. Largely Unchanged, " www.cdc.gov/nchs.

Neyman, J. (1923, 1990), "On the Application of Probability Theory to Agricultural Experiments," *Statistical Science,* 5, 463–480.

Normand, S. L. T., Landrum, N. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., and McNeil, B. J. (2001), "Validating Recommendations for Coronary Angiography Following Acute Myocardial Infarction in the Elderly," *Journal of Clinical Epidemiology,* 54, 387–398.

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,* eds. E. Halloran and D. Berry, New York: Springer, pp. 1–94.

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

———. (1991), "Discussing Hidden Bias in Observational Studies," *Annals of Internal Medicine,* 115, 901–905.

———. (1993), "Hodges–Lehmann Point Estimates in Observational Studies," *Journal of the American Statistical Association,* 88, 1250–1253.

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

———. (2002), *Observational Studies,* New York: Springer.

———. (2006), "Comment on a Paper by Donald B. Rubin: The Place of Death in the Quality of Life," *Statistical Science,* 21, 313–316.

———. (2007a), "Confidence Intervals for Uncommon but Dramatic Responses to Treatment," *Biometrics,* 63, 1164–1171.

———. (2007b), "Sensitivity Analysis for M-Estimates, Tests and Confidence Intervals in Matched Observational Studies," *Biometrics,* 63, 456–464.

———. (2008), "Testing Hypotheses in Order," *Biometrika,* 95, 248–252.

Rosenbaum, P., and Rubin, D. (1983), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," *Journal of the Royal Statistical Society* Ser. B, 45, 212–218.

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

Schuirmann, D. L. (1981), "On Hypothesis Testing to Determine if the Mean of a Normal Distribution Is Contained in a Known Interval," *Biometrics,* 37, 617.

Silber, J. H., Lorch, S. L., Rosenbaum, P. R., Medoff-Cooper, B., Bakewell-Sachs, S., Millman, A., Mi, L., Even-hoshan, O., Escobar, G. E. (2009), "Additional Maturity at Discharge and Subsequent Health Care Costs," *Health Services Research,* to appear.

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

Small, D., Ten Have, T., and Rosenbaum, P. R. (2008), "Randomization Inference in a Group-Randomized Trial of Treatments for Depression: Covariate Adjustment, Noncompliance and Quantile Effects," *Journal of the American Statistical Association,* 103, 271–279.

Stephenson, W. R. (1981), "A General Class of One-Sample Nonparametric Test Statistics Based on Subsamples," *Journal of the American Statistical Association,* 76, 960–966.

Tamhane, A., and Logan, B. (2004), "A Superiority–Equivalence Approach to One-Sided Tests on Multiple Endpoints in Clinical Trials," *Biometrika,* 91, 715–727.

Tan, Z. (2006), "A Distributional Approach for Causal Inference Using Propensity Scores," *Journal of the American Statistical Association,* 101, 1619–1637.

Wang, L., and Krieger, A. M. (2006), "Causal Conclusions Are Most Sensitive to Unobserved Binary Covariates," *Statistics in Medicine,* 25, 2257–2271.

Westlake, W. J. (1981), "Response to Kirkwood," *Biometrics,* 37, 591–593.

Yu, B. B., and Gastwirth, J. L. (2005), "Sensitivity Analysis for Trend Tests: Application to the Risk of Radiation Exposure," *Biostatistics,* (Oxford, England), 6, 201–209.