Building a Stronger Instrument in an Observational Study of Perinatal Care for Premature Infants

Mike BAIOCCHI, Dylan S. SMALL, Scott LORCH, and Paul R. ROSENBAUM

An instrument is a random nudge toward acceptance of a treatment that affects outcomes only to the extent that it affects acceptance of the treatment. Nonetheless, in settings in which treatment assignment is mostly deliberate and not random, there may exist some essentially random nudges to accept treatment, so that use of an instrument might extract bits of random treatment assignment from a setting that is otherwise quite biased in its treatment assignments. An instrument is weak if the random nudges barely influence treatment assignment or strong if the nudges are often decisive in influencing treatment assignment. Although ideally an ostensibly random instrument is perfectly random and not biased, it is not possible to be certain of this; thus a typical concern is that even the instrument might be biased to some degree. It is known from theoretical arguments that weak instruments are invariably sensitive to extremely small biases; for this reason, strong instruments are preferred. The strength of an instrument is often taken as a given. It is not. In an evaluation of effects of perinatal care on the mortality of premature infants, we show that it is possible to build a stronger instrument, we show how to do it, and we show that success in this task is critically important. We also develop methods of permutation inference for effect ratios, a key component in an instrumental variable analysis.

KEY WORDS: Design sensitivity; Effect ratio; Instrumental variable; Nonbipartite matching; Observational study; Optimal matching; Sensitivity analysis.

1. INTRODUCTION: MOTIVATION, EXAMPLE, AND DATA

1.1 Regionalization of Intensive Care for Premature Infants: Does It Save Lives?

Hospitals vary in their ability to care for premature infants. The American Academy of Pediatrics recognizes six levels of neonatal intensive care units (NICUs) of increasing technical expertise and capability: 1, 2, 3A, 3B, 3C, 3D, and regional centers, 4. The term "regionalization of care" refers to a policy that suggests or requires that high-risk mothers deliver at hospitals with greater capabilities. In other words, within a region, mothers are to be sorted into hospitals of varied capability based on the risks faced by the newborn, rather than on haphazard circumstances, such as affiliation or proximity. Regionalized perinatal systems were developed in the 1970s, when NICUs began to save infants with birth weight <1500 g. In the 1990s, however, NICU services began to diffuse from regional centers to community hospitals. Regionalization might reduce infant mortality by bringing together the sickest babies and the most capable hospitals; however, regionalization might not reduce infant mortality because the sorting by risk might be too inaccurate to affect health, or the capabilities of high-level NICUs might fail to deliver better outcomes.

In the current paper, we focus on whether delivering high risk infants at more capable NICUs reduces mortality. This is one key component in the evaluation of regionalized perinatal systems. More precisely, if a high-risk mother delivers at a less capable hospital, is her baby at greater risk of death? In a highly abstract world remote from the world that we inhabit, a randomized experiment could settle that question, with high-risk mothers assigned at random to hospitals of varied capabilities. In the world that we actually do inhabit, in which medical decisions are happily constrained by considerations of sound judgment, ethics, and patient preferences, such an experiment is not possible. We need to make some reasonable sense of the data that we can obtain. There is a basic difficulty, however, that arises in many contexts in which the most intense and capable care is given to the sickest patients. If regionalization succeeded in sorting mothers by risk, then the highest-risk mothers would deliver at the most-capable hospitals. The mortality rates at the more-capable hospitals might be higher, not lower, than those the less-capable hospitals because their patient populations were sicker, even if the more-capable hospitals were saving lives. A naïve comparison of mortality rate by level of NICU would do little or nothing to clarify whether regionalization is or is not effective, because it would not estimate the effect on mortality of delivery at a more-capable hospital.

Here we take an old tactic and improve it. The old tactic exploits proximity. A high-risk mother is more likely to deliver at a hospital with a high-level NICU if such a hospital is close to home. A pregnancy may conclude with a certain urgency, and awareness of this possibility may lead the mother to want to avoid a long trip. If travel time to a hospital with a high-level NICU affected risk only if it altered whether the baby received care at that hospital, then the so-called "exclusion restriction" would be plausible. (See Angrist, Imbens, and Rubin 1996 for a discussion of the exclusion restriction.) If it were also true that the mother's risk was unrelated to geography, then proximity would be an instrument for care at a hospital with a high-level NICU. In point of fact, the mother's risk is related to geography, largely through socioeconomic factors that vary with geography; however, we attempted to control for this and other issues by matching for measured covariates.

Mike Baiocchi is Doctoral Student, Dylan S. Small is Associate Professor, and Paul R. Rosenbaum is Professor (E-mail: *rosenbaum@wharton.upenn.edu*), Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340. Scott Lorch is Assistant Professor of Pediatrics at the University of Pennsylvania School of Medicine and an Attending Physician in the Division of Neonatology at The Children's Hospital of Philadelphia. This work was supported by grant SES-0849370 from the Measurement, Methodology and Statistics Program of the U.S. National Science Foundation and a grant from the Agency for Healthcare Research and Quality.

Journal of the American Statistical Association
December 2010, Vol. 105, No. 492, Applications and Case Studies
DOI: 10.1198/jasa.2010.ap09490

Proximity would be a strong instrument for delivery at a hospital with a high-level NICU if proximity were typically decisive in determining where the mother delivered. Proximity would be a weak instrument if it were a minor factor among many others. (For a discussion of various issues that arise with weak instruments, see Bound, Jaeger, and Baker 1995 and Imbens and Rosenbaum 2005.)

Weak instruments are invariably sensitive to very small unobserved biases, so strong instruments are an aspect of strong evidence. Here bias refers to nonrandom assignment of the instrument. Small and Rosenbaum (2008) studied the relationship between the strength of a particular instrument and its sensitivity to unobserved biases. Their criterion was the power of a sensitivity analysis with an instrument, which is the probability that a study will reject a false null hypothesis when a specified magnitude of unobserved bias in the instrument is allowed for. (See Rosenbaum 2004, 2005a for general discussion of the power of a sensitivity analysis.) Consider two studies, one with a strong instrument and the other with a weak instrument. If we assume that the instrument was randomly assigned, then the problems caused by a weak instrument might be offset by a sufficiently large sample size. But Small and Rosenbaum (2008) showed that if we take into account the possibility that an instrument is not perfectly random, then the small study with a stronger instrument is likely to be more powerful (in terms of power of sensitivity analysis) than the vastly larger study with a weaker instrument; indeed, the power with a weak instrument might tend to 0 with increasing sample size for a magnitude of bias such that the power with a strong instrument is tending to 1. In this article we demonstrate that, through careful design, we can extract from a single large study with a weak instrument a more powerful, smaller study with a stronger instrument.

1.2 Data: Covariates, NICU Level, Travel Time, and Survival

The data describe all premature births in the Commonwealth of Pennsylvania in the years 1995–2004 plus the first 6 months of 2005; that is, approximately 200,000 births. The data combine information from birth and death certificates and the UB-92 form, which hospitals provide.

Regionalization is a policy that would alter the level of the NICU at which a high-risk mother would deliver; it is not aimed at improving prenatal care, and is not a sensible strategy for improving prenatal care. Because we are interested in comparing the effectiveness of neonatal care provided by different levels of NICUs, we consider variables determined before birth as covariates. To the greatest extent possible, we would like to compare babies who were similar at birth and received the same prenatal care but received neonatal care at NICUs of different levels. We do not want to confuse an effect of NICU level on perinatal care with an effect of prenatal care provided by someone else. These covariates include birth weight and gestational age, prenatal care, health insurance, congenital anomalies, and other variables listed in Table 1. If some other study were interested in the effects not of NICU level but rather of, say, prenatal care, then some of the variables that are pretreatment covariates in our study might be considered outcomes in that other study. This is true, for example, of birth weight, which is not materially affected by the NICU level but might be affected by prenatal care, for instance, by coaxing a mother to abstain from smoking.

Following Rogowski et al. (2004), we recorded a mother as having delivered at a low-level hospital ($D = 1$) if that hospital delivered an average of fewer than 50 preterm babies per year or if its NICU was below level 3A, or as having delivered at a high-level hospital ($D = 0$) if the hospital delivered at least 50 preterm babies per year and had a NICU of level 3A–3D or 4. We investigated whether delivery of a preterm infant at a low-level hospital increases the risk of death, and if so, by how much?

Travel time was determined using ArcView software (ESRI) as the time from the centroid of mother's zip code to the closest low-level and high-level hospitals. The degree of encouragement to deliver at a low-level hospital was the difference in these two travel times, high-minus-low; for brevity, this is termed the *excess travel time*. Excess travel time takes a negative value if the closest hospital has a high-level NICU. Distance strongly encourages the mother to deliver at a low-level hospital if the difference in travel time is positive and large.

Stop for a moment and think about Pennsylvania, a state with two large cities (Philadelphia and Pittsburgh), several medium-sized cities (e.g., Harrisburg, Allentown–Bethlehem), numerous small towns, and large remote rural areas. Although many small towns are served by small hospitals, some are not. The highly capable medical school of Pennsylvania State University is in Hershey, Pennsylvania, with farming communities on several sides. In Philadelphia, there are many hospitals, some within walking distance of one another; thus excess travel times are small, and excess travel time will rarely determine where the mother delivers. In a rural area, excess travel time may be a decisive factor. Of course, most people live in or near urban areas. The full study (for which the current analysis is a pilot study) will look at Pennsylvania, Missouri, and California as three representative states; however, we are interested in the effects of high-level NICUs on mortality in general, not specifically in these states. Pennsylvania yields an instrument, but perhaps Pennsylvania is not ideally structured as a state to answer our question. Should we take Pennsylvania as it is, or should we improve Pennsylvania to build a stronger instrument?

2. MATCHING TO CREATE STRONGER INSTRUMENTS

2.1 Fewer Pairs at Greater Distances

We used optimal nonbipartite matching to pair babies with similar covariates but different excess travel times. There were $2I$ babies. First, a discrepancy was defined between every pair of babies, yielding a $2I \times 2I$ discrepancy matrix. (Here the term "discrepancy" is used in place of the more common term "distance," to avoid confusing the covariate discrepancy with the geographic distance to a NICU.) An optimal nonbipartite matching then divided the $2I$ babies into I nonoverlapping pairs of two babies in such a way that the sum of the discrepancies within the I pairs was minimized. That is, two babies in the same pair were as similar as possible. Fortran code for a polynomial-time optimization algorithm was developed by Derigs (1988) and was made available inside R by Lu et al. (2009). (For statistical applications of optimal nonbipartite matching,

Table 1. Covariate balance and degree of encouragement in two matched comparisons. Nine rare congenital anomalies were balanced as well. (–St-diff– = absolute standardized difference. 1/0 means 1 = yes, 0 = no. Prenatal care month refers to month in which prenatal care begain. Mother's education scale is a six-point scale with high school graduate scored as 3 and college graduate scored as 5. For zip code/census data, fr = fraction of zip code.)

	Weaker instrument No sinks 99,174 pairs of two babies			Stronger instrument Sinks remove 50% of babies 49,587 pairs of two babies		
	Near mean	Far mean	–St-dif–	Near mean	Far mean	–St-dif–
Magnitude of encouragement						
Excess travel time to high-level NICU, minutes	4.48	17.98	0.78	0.86	35.08	1.97
Pregnancy and birth						
Covariates						
Birth weight, g	2582	2581	0.00	2584	2581	0.00
Gestational age, weeks	35.11	35.11	0.00	35.14	35.13	0.00
Gestational diabetes, 1/0	0.05	0.05	0.00	0.04	0.04	0.01
Prenatal care, month	2.31	2.30	0.01	2.22	2.20	0.02
Prenatal care missing	0.11	0.11	0.02	0.07	0.07	0.02
Single birth, 1/0	0.83	0.83	0.00	0.85	0.83	0.05
Parity	2.11	2.11	0.00	2.01	2.03	0.02
Mother						
Mother's age	28.15	28.10	0.01	27.99	27.66	0.05
Mother's education (scale)	3.71	3.70	0.01	3.72	3.65	0.06
Mother's education missing	0.03	0.02	0.02	0.01	0.01	0.00
White, 1/0	0.70	0.71	0.03	0.85	0.86	0.01
Black, 1/0	0.17	0.15	0.04	0.06	0.05	0.03
Asian, 1/0	0.01	0.01	0.01	0.01	0.00	0.01
Other race, 1/0	0.03	0.03	0.00	0.02	0.01	0.01
Race missing, 1/0	0.09	0.09	0.01	0.07	0.08	0.04
Mother's health insurance						
Fee for service, 1/0	0.21	0.21	0.00	0.24	0.25	0.01
HMO, 1/0	0.37	0.37	0.00	0.35	0.33	0.04
Federal/state, 1/0	0.30	0.30	0.00	0.30	0.31	0.04
Other, 1/0	0.10	0.10	0.00	0.10	0.09	0.00
Uninsured, 1/0	0.01	0.01	0.00	0.01	0.01	0.02
Mother's neighborhood (zip code/census)						
Zip code data missing	0.06	0.06	0.00	0.00	0.00	0.00
Income ($1000)	41	41	0.01	42	40	0.13
Below poverty (fr)	0.13	0.13	0.02	0.11	0.10	0.02
Home value ($1000)	95	96	0.02	97	97	0.02
Has high school degree (fr)	0.80	0.80	0.00	0.82	0.82	0.02
Has college degree (fr)	0.21	0.21	0.03	0.21	0.19	0.12
Rent (fr)	0.30	0.29	0.06	0.28	0.26	0.15

see Lu et al. 2001, Lu and Rosenbaum 2004, Lu 2005, and Rosenbaum 2005b, and for a different application in neonatology, see Rosenbaum and Silber 2009a and Silber et al. 2009.)

We contrast two such matchings. One matching is slightly compulsive: it must, absolutely must, use every baby (about 200,000 babies), even though this implies that many excess travel times are small, so the instrument is fairly weak. This compulsion is not justified by statistical theory, which unambiguously shows that the problems of weak instruments are often so severe that they outweigh large increases in sample size (Small and Rosenbaum 2008), so the compulsion has its origins elsewhere. The other matching uses about half of the babies (about 100,000), allowing pairs that are closely matched for covariates, yet with substantial differences in excess travel time. In the second matching, we have about 50,000 pairs of babies closely matched for covariates, one far from the nearest high-level NICU and the other much closer.

The second matching eliminates some babies in an optimal manner using "sinks" (see Lu et al. 2001). To eliminate e babies, e sinks are added to the data set before matching, with each sink at zero discrepancy to each baby and at infinite discrepancy to all other sinks. This yields a $(2I+e)\times(2I+e)$ discrepancy matrix. An optimal match will pair e babies to the e sinks in such a way as to minimize the total of the remaining discrepancies within $I-e/2$ pairs of $2I-e$ babies; that is, the best possible choice of e babies is removed. The second match eliminates about half of the babies.

The discrepancy matrix was built in several steps using standard devices. Because we are matching mothers from different parts of Pennsylvania, and because socioeconomic status varies from place to place, it is important to compare mothers from wealthy communities with other mothers from wealthy communities and to compare mothers from poor communities with other mothers from poor communities. The six census/zip code measures are intended to represent local socioeconomic status, but socioeconomic status is not six-dimensional. First, socioeconomic measures describing a zip code were summarized using their first two principal components. These two components were combined with individual-level data about mother and baby in calculating a Mahalanobis discrepancy between each pair of babies; see Rubin (1980). A small penalty (i.e., a positive number) was added to the discrepancy for each of the following circumstances for any pair of babies that (a) did not agree on the number of congenital disorders, (b) did not agree on black race, or (c) did not agree on whether zip code information was missing. Two independent observations drawn from the same L-variate multivariate normal distribution have an expected Mahalanobis discrepancy equal to $2L$, so that, speaking informally, a penalty that is typically of size 2 will double the importance of matching on a variable. Small penalties are used to secure balance for a few recalcitrant covariates, usually those that are most systematically out of balance (see Rosenbaum 2010, sec. 9.2 for a discussion). It is typical to adjust small penalities to secure the desired balance. Finally, a substantial penalty was added to the discrepancy between any pair of babies whose excess travel time differed in absolute value by at most Λ, with $\Lambda = 0$ in the first match described earlier and $\Lambda = 25$ minutes in the second match. Substantial (effectively infinite) penalties are used to enforce compliance with a constraint whenever compliance is possible, and also to minimize the extent of deviation from a constraint whenever strict compliance is not possible. This substantial penalty used a "penalty function," a continuous function that is 0 if the constraint is respected and rises rapidly as the magnitude of the violation of the constraint increases. (See Avriel 1976 for a discussion of penalty functions and Rosenbaum 2010, sec. 8.4 for a discussion of the use of penalty functions in matching.)

In fact, we matched exactly on three important covariates, year of birth and coarse categorical versions of birth weight and gestational age. This means that we split one large matching problem into several smaller matching problems, grouping the pairs into one study at the end. Along with ensuring exact matches on these three covariates, this allows a rather large matching problem (~200,000 babies) to be broken into several smaller problems that can be solved separately in the manner described earlier. Because the discrepancy matrix has size on the order of the square of the number of babies, and the algorithm has a worst-case time bound on the order of the cube of the number of babies, splitting the problem to produce an exact match drastically reduces the computational effort (see Rosenbaum 2010, sec. 9.3). Inside those exact match categories, we also used the continuous versions of birth weight and gestational age to obtain closer matches than required by the categories alone.

2.2 Two Matched Comparisons, One Stronger and One Weaker, in the Study of Regionalization of Perinatal Care

Table 1 shows the two matches in terms of covariate balance and difference in excess travel time. Keep in mind that we want pairs that are similar in terms of covariates and different in terms of excess travel time. Table 1 shows means and absolute standardized differences in means, that is, the absolute value of the difference in means divided by the standard deviation before matching. The match on the left uses all of the babies and forms 99,174 pairs of babies, requiring only that the paired babies have different excess travel times. The match on the right uses sinks in an effort to enforce a difference in excess travel time of at least 25 minutes, thereby yielding 49,587 pairs of babies.

In Table 1, the two matched comparisons are both well matched for covariates. One could not choose between the two matches based on comparability in terms of covariates. They differ in a few ways. By design, one match uses all of the babies, and the other match uses about half of the babies; other things being equal, this speaks in favor of the match with more babies, but other things are far from equal. By design, there is a larger difference in excess travel time in the match with fewer babies, $35.08 - 0.86 = 34.22$ minutes versus $17.98 - 4.48 = 13.50$ minutes, or almost 2 standard deviations (SDs) versus about 0.75 SD. Because we think that after matching on key covariates, variation in NICU level produced by proximity to the hospital is likely to have little to do with infant survival besides influencing the choice of NICU, we prefer a larger difference in travel time. Our parallel analyses will contrast the two matchings.

Figure 1 contrasts three matched comparisons, the two displayed in Table 1 and one additional comparison. In Figure 1, All-0 refers to using all of the babies requiring only a difference in excess travel time greater than 0, and Half-25 refers to using half of the babies requiring a difference in excess travel time of 25 minutes. The additional comparison is All-25, which matched all of the babies and tried to force a difference in excess travel time of 25 minutes. It is clear that All-25 is not acceptable as a match, because quite a few covariates are substantially out of balance, and the difference in mean travel time is 23.4 minutes, compared with 34.2 minutes for the Half-25 match. In particular, in the All-25 match, 24% of mothers near a high-level NICU were black, as opposed to 8% of those far away from a high-level NICU, and there also was a 0.5 SD difference in the fraction of mother's zip code that was below the poverty line. Something has to give; it is not possible to use all of the babies while making pairs that are both close on covariates and far apart on travel time.

For many of the covariates listed in Table 1, the two matched comparisons appear similar. For instance, for such key variables as birth weight and gestational age, the two matched comparisons are similar. There are some differences, however. For instance, in Pennsylvania, blacks are disproportionately in urban areas, so it is difficult to find a pair of blacks, one far from a high-level NICU, the other close; most blacks are not far from a high-level NICU. The smaller stronger match is about 5% black, whereas the larger weaker match is about 15% black. There are also smaller differences in health insurance. These differences would be critically important if describing Pennsylvania accurately were critically important, but there is nothing special about Pennsylvania—it was picked as one of three representative states. Moreover, the second match is much closer to a clean experiment in which something haphazard was often decisive for treatment assignment.

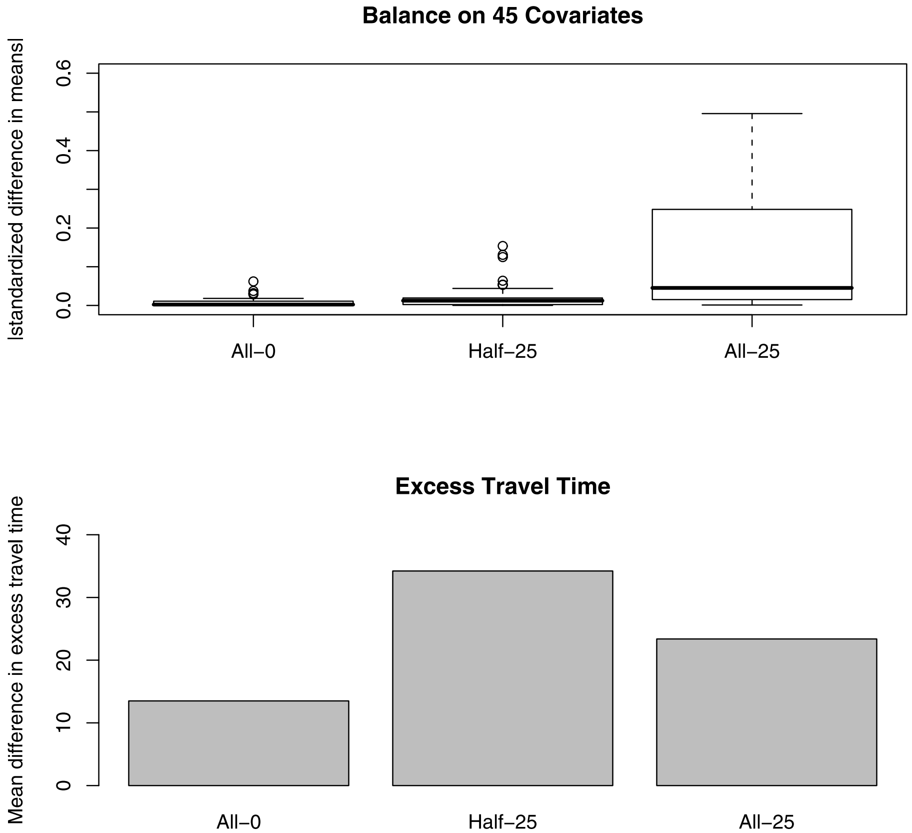

Figure 1. Comparison of three matched comparisons in terms of comparability on covariates and excess travel time. The match All-0 uses all of the babies but insists only on a nonzero difference in excess travel time. The match Half-25 uses half of the babies while trying to obtain at least a 25-minute difference in excess travel time. The match All-25 uses all of the babies while trying to obtain at least a 25-minute difference in excess travel time. Covariate balance is measured by the absolute standardized difference in covariate means. It is clear that All-25 is not an acceptable match; the imbalances in many covariates, including race and poverty, are quite large.

3. INFERENCE ABOUT EFFECT RATIOS

3.1 Notation: Treatment Effects, Treatment Assignments

There are I matched pairs, $i = 1, \ldots, I$, with two subjects, $j = 1, 2$, one treated subject and one control, or $2I$ subjects in total. If the jth subject in pair i receives the treatment, write $Z_{ij} = 1$, but if this subject receives the control, write $Z_{ij} = 0$, so $1 = Z_{i1} + Z_{i2}$ for $i = 1, \ldots, I$. In our study in Section 1, the matched pairs consist of one mother close to a high-level NICU (say control) and the one mother further away (say treated). Note that in this terminology, proximity is the "treatment," although our real interest is in the effect of delivering at a low-level versus high-level hospital. To emphasize, there are two matched samples in Table 1, and the notation can be understood as referring to either matched sample alone, but the relevant quantities and their meanings depend on which matched sample is under consideration.

The subscripts ij are bookkeeping labels and carry no information; all information about subjects is contained in observed or unobserved variables that describe them. (It is easy to construct noninformative labels; number the pairs i at random, then number the subjects j at random within each pair.) The matched pairs were formed by matching for an observed covariate $\mathbf{x}_{ij}$, but might have failed to control an unobserved covariate, u_{ij}, that is, $\mathbf{x}_{ij} = \mathbf{x}_{ik}$ for all i, j, k, but possibly $u_{ij} \neq u_{ik}$. This structure is in preparation for the inevitable comment or concern that the pairs in Table 1 look similar in terms of the variables in Table 1, but the table omits the specific covariate u_{ij}, which might bias the comparison. Write $\mathbf{u} = (u_{11}, u_{12}, \ldots, u_{I2})^T$ for the $2I$-dimensional vector.

For any outcome, each subject has two potential responses, one seen under treatment, $Z_{ij} = 1$, and the other seen under control, $Z_{ij} = 0$ (see Neyman 1923; Rubin 1974). In Section 1, speaking in this way of two potential responses entails imagining that a mother, ij, who lived either close to a high-level NICU ($Z_{ij} = 0$) or far from one ($Z_{ij} = 1$) might instead have lived in the opposite circumstances. What would have happened to a mother and her newborn had she lived either close to or far from a high-level NICU? Here there are two responses, (r_{Tij}, r_{Cij}) or (d_{Tij}, d_{Cij}), where r_{Tij} and d_{Tij} are observed from the jth subject in pair i under treatment, $Z_{ij} = 1$, whereas r_{Cij} and d_{Cij} are observed from this subject under control, $Z_{ij} = 0$. In Section 1, (r_{Tij}, r_{Cij}) indicates infant death (1 for dead, 0 for alive) and (d_{Tij}, d_{Cij}) indicates whether the mother delivered at a hospital *without* a high-level NICU (1 if yes, 0 if no). For instance, if $(r_{Tij}, r_{Cij}) = (1, 0)$ with $(d_{Tij}, d_{Cij}) = (1, 0)$ then (a) had the mother lived far from a high-level NICU ($Z_{ij} = 1$), she would not have delivered at a high-level NICU ($d_{Tij} = 1$), and her baby would have died ($r_{Tij} = 1$), but (b) had the mother lived near a high-level NICU ($Z_{ij} = 0$), then she would have delivered at a high-level NICU ($d_{Cij} = 0$), and her baby would have survived ($r_{Cij} = 0$).

Table 2. Magnitude of encouragement, level of NICU, and mortality in two matched comparisons. (–St-diff– = absolute standardized difference. 1/0 means 1 = yes, 0 = no.)

	Weaker instrument No sinks 99,174 pairs of two babies			Stronger instrument Sinks remove 50% of babies 49,587 pairs of two babies		
	Near mean	Far mean	–St-dif–	Near mean	Far mean	–St-dif–
			Magnitude of encouragement			
Excess travel time to high-level NICU, minutes	4.48	17.98	0.78	0.86	35.08	1.97
			Delivery at low-level NICU, D_{ij}			
Low-level NICU, 1/0	0.35	0.53	0.36	0.31	0.75	0.88
			Infant mortality, R_{ij}			
Dead, 1/0	0.0181	0.0198	0.01	0.0155	0.0194	0.03

The effects of treatment on a subject, $r_{Tij} - r_{Cij}$ or $d_{Tij} - d_{Cij}$, are not observed for any subject; that is, each mother lives either near to or far from a high-level NICU, and the fate of her baby under the opposite circumstance is not observed. However, $R_{ij} = Z_{ij} r_{Tij} + (1 - Z_{ij}) r_{Cij}$, $D_{ij} = Z_{ij} d_{Tij} + (1 - Z_{ij}) d_{Cij}$, and Z_{ij} are observed from each subject. Let $\mathcal{F} = \{(r_{Tij}, r_{Cij}, d_{Tij}, d_{Cij}, \mathbf{x}_{ij}, u_{ij}), i = 1, \dots, I, j = 1, 2\}$. Table 2 repeats the information from Table 1 about excess travel time and adds the information about the two outcomes NICU level and mortality. In the second match in Table 2, the difference in excess travel times is larger, with the consequence that more mothers far from high-level NICUs did not deliver at high-level NICUs; that is, the instrument is stronger.

Fisher's sharp null hypothesis of no treatment effect on (r_{Tij}, r_{Cij}) asserts that $H_0 : r_{Tij} = r_{Cij}$, for $i = 1, \dots, I$, $j = 1, 2$. In Section 1, this says that living close to a high-level NICU has no effect on perinatal mortality, even if proximity shifts some mothers to deliver at a hospital with a high-level NICU. If Fisher's null hypothesis were plausible, then it would be difficult to argue that regionalization of care is warranted.

In this article we make reference to the exclusion restriction, but we do not assume that it is true. The exclusion restriction asserts that $d_{Tij} = d_{Cij}$ implies $r_{Tij} = r_{Cij}$ (see Angrist, Imbens, and Rubin 1996). In Section 1, the exclusion restriction says that mother and baby are affected by proximity to a high-level NICU only if proximity to a high-level NICU changes the type of hospital at which the mother delivers. As we show, our analysis does not require the exclusion restriction, but a key parameter has an additional interpretation when the exclusion restriction is true.

A substantial distance between mother's home and the nearest high-level NICU is thought to "encourage" the mother to deliver at a less capable but presumably closer hospital. A mother with $(d_{Tij}, d_{Cij}) = (1, 0)$ is said to be a "complier," in the sense that she would deliver at a high-level NICU if one were close by ($d_{Cij} = 0$), but would deliver at a less-capable hospital if she lived far away ($d_{Tij} = 1$).

Write $|A|$ for the number of elements in a finite set A. Let $\mathbf{Z} = (Z_{11}, Z_{12}, \dots, Z_{I,2})^T$, and let Ω be the set containing the $|\Omega| = 2^I$ possible values $\mathbf{z}$ of $\mathbf{Z}$, so that $\mathbf{z} \in \Omega$ if $\mathbf{z} = (z_{11}, z_{12}, \dots, z_{I,2})^T$ with $z_{ij} = 0$ or $z_{ij} = 1$, $1 = z_{i1} + z_{i2}$ for $i = 1, \dots, I$. Write $\mathcal{Z}$ for the event that $\mathbf{Z} \in \Omega$. In a randomized experiment, $\mathbf{Z}$ is chosen at random from Ω, so that $\Pr(\mathbf{Z} = \mathbf{z} | \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ for each $\mathbf{z} \in \Omega$.

3.2 Effect Ratios

The effect ratio, λ, is the parameter

$$\lambda = \frac{\sum_{i=1}^{I} \sum_{j=1}^{2} (r_{Tij} - r_{Cij})}{\sum_{i=1}^{I} \sum_{j=1}^{2} (d_{Tij} - d_{Cij})}, \tag{1}$$

where it is implicitly assumed that $0 \neq \sum_{i=1}^{I} \sum_{j=1}^{2} d_{Tij} - d_{Cij}$. Here λ is a parameter of the finite population of $2I$ individuals whose data are recorded in $\mathcal{F}$, and because (r_{Tij}, r_{Cij}) and (d_{Tij}, d_{Cij}) are not jointly observed, λ cannot be calculated from observable data, so inference is required. Note that under Fisher's sharp null hypothesis of no effect H_0 in Section 3.1, $\lambda = 0$.

The effect ratio is the ratio of two average treatment effects. In a paired, randomized experiment, the mean of the treated-minus-control difference provides unbiased estimates of numerator and denominator effects separately, and under mild conditions, as $I \to \infty$, the ratio of these unbiased estimates is consistent for λ. The effect ratio measures the relative magnitude of two treatment effects, here the effect of distance on mortality compared with its effect on where mothers deliver. For instance, if $\lambda = 1/100$, then for every hundred mothers discouraged by distance from delivering at a hospital with a high-level NICU, there is one additional infant death. With no further assumptions, λ is both estimable in a randomized experiment and interpretable; however, the interpretation does not explicitly link the effects in the numerator and the effects in the denominator.

As discussed by Angrist, Imbens, and Rubin (1996), with additional assumptions such as the exclusion restriction and monotonicity, λ would be the average increase in mortality caused by delivering at a less-capable hospital among compliers, that is, mothers with $(d_{Tij}, d_{Cij}) = (1, 0)$, or mothers who would deliver at a low-level NICU if and only if no high-level NICU was close by. Our inferences are valid for λ whether or not the exclusion restriction lends this interpretation to λ. Here λ is unknown and is a function of $\mathcal{F}$.

3.3 Inference About an Effect Ratio in a Randomized Experiment

Consider the null hypothesis, $H_0^{(\lambda)} : \lambda = \lambda_0$. Here $H_0^{(\lambda)}$ is a composite hypothesis; there are many different finite populations $\mathcal{F}$ in which $H_0^{(\lambda)} : \lambda = \lambda_0$ is true. Recall that the size of a test of a composite hypothesis is the supremum over null hypotheses of the probability of rejection, and a valid test has size less than or equal to its nominal level. We test the hypothesis with the aid of the statistic

$$\begin{aligned} T(\lambda_0) &= \frac{1}{I}\sum_{i=1}^{I}\left\{\sum_{j=1}^{2} Z_{ij}(R_{ij}-\lambda_0 D_{ij}) - \sum_{j=1}^{2}(1-Z_{ij})(R_{ij}-\lambda_0 D_{ij})\right\} \\ &= \frac{1}{I}\sum_{i=1}^{I} V_i(\lambda_0), \quad \text{say,} \end{aligned} \tag{2}$$

where, because $R_{ij}-\lambda_0 D_{ij} = r_{Tij}-\lambda_0 d_{Tij}$ if $Z_{ij}=1$ and $R_{ij}-\lambda_0 D_{ij} = r_{Cij}-\lambda_0 d_{Cij}$ if $Z_{ij}=0$, we can write

$$V_i(\lambda_0) = \sum_{j=1}^{2} Z_{ij}(r_{Tij}-\lambda_0 d_{Tij}) - \sum_{j=1}^{2}(1-Z_{ij})(r_{Cij}-\lambda_0 d_{Cij}). \tag{3}$$

We define $y_{Tij,\lambda_0} = r_{Tij}-\lambda_0 d_{Tij}$, $y_{Cij,\lambda_0} = r_{Cij}-\lambda_0 d_{Cij}$, and

$$S^2(\lambda_0) = \frac{1}{I(I-1)}\sum_{i=1}^{I}\{V_i(\lambda_0)-T(\lambda_0)\}^2.$$

Propositions 1 and 2 state certain facts about the behavior of $T(\lambda_0)/S(\lambda_0)$ as a statistic for testing the composite hypothesis $H_0^{(\lambda)}: \lambda=\lambda_0$. More or less, under reasonable conditions, Propositions 1 and 2 state that the test works. The propositions are followed by several remarks that set these facts in either a historical or practical context. The central result of Section 3.3 is the inequality (10) on tail probabilities for $T(\lambda_0)/S(\lambda_0)$ when the composite hypothesis $H_0^{(\lambda)}: \lambda=\lambda_0$ is true. Because this is an inequality, not an equality, one might mistakenly believe that the use of (10) yields a conservative test of the composite hypothesis $H_0^{(\lambda)}: \lambda=\lambda_0$, and the remarks are largely intended to clarify why such a thought is indeed a mistake. The issue turns on the fact that the size of a test of a composite hypothesis is a supremum of the probability of false rejection over all simple null hypotheses contained in the composite null hypothesis. Because the inequality (10) is an equality for some simple null hypotheses within the composite null hypothesis, in large samples, a test that derives p-values from (10) has an actual size close to its nominal level; it is not conservative as a test of the composite hypothesis.

Proposition 1. In a randomized experiment with $\Pr(\mathbf{Z}=\mathbf{z}|\mathcal{F},\mathcal{Z}) = 1/|\Omega|$ for each $\mathbf{z}\in\Omega$, the $V_i(\lambda_0)$ are mutually independent given $\mathcal{F},\mathcal{Z}$, and

$$\begin{aligned} \mathrm{E}\{V_i(\lambda_0)|\mathcal{F},\mathcal{Z}\} &= \frac{1}{2}\big(y_{Ti1,\lambda_0}-y_{Ci1,\lambda_0}+y_{Ti2,\lambda_0}-y_{Ci2,\lambda_0}\big) \\ &= \mu_{i,\lambda_0}, \quad \text{say,} \end{aligned} \tag{4}$$

$$\begin{aligned} \mathrm{var}\{V_i(\lambda_0)|\mathcal{F},\mathcal{Z}\} &= \frac{1}{4}\big(y_{Ti1,\lambda_0}-y_{Ti2,\lambda_0}+y_{Ci1,\lambda_0}-y_{Ci2,\lambda_0}\big)^2 \\ &= \nu_{i,\lambda_0}, \quad \text{say,} \end{aligned} \tag{5}$$

$$\begin{aligned} \mathrm{E}\{T(\lambda_0)|\mathcal{F},\mathcal{Z}\} &= (\lambda-\lambda_0)\frac{1}{2I}\sum_{i=1}^{I}\sum_{j=1}^{2}(d_{Tij}-d_{Cij}) \\ &= \frac{1}{I}\sum_{i=1}^{I}\mu_{i,\lambda_0} = \overline{\mu}_{\lambda_0}, \quad \text{say,} \end{aligned} \tag{6}$$

$$\mathrm{var}\{T(\lambda_0)|\mathcal{F},\mathcal{Z}\} = \frac{1}{I^2}\sum_{i=1}^{I}\nu_{i,\lambda_0}, \tag{7}$$

$$\begin{aligned} &\mathrm{E}\{S^2(\lambda_0)|\mathcal{F},\mathcal{Z}\} - \mathrm{var}\{T(\lambda_0)|\mathcal{F},\mathcal{Z}\} \\ &\quad = \frac{1}{I(I-1)}\sum_{i=1}^{I}\big(\mu_{i,\lambda_0}-\overline{\mu}_{\lambda_0}\big)^2. \end{aligned} \tag{8}$$

Proof. Given $\mathcal{F},\mathcal{Z}$ in a randomized experiment, $\mathrm{E}(Z_{ij}) = 1/2$, so (4) and (5) follow from (3). The (Z_{i1}, Z_{i2}) in distinct matched pairs i are mutually independent, so the $V_i(\lambda_0)$ are independent, and (7) follows from this. Using this in (2) yields

$$\mathrm{E}\{T(\lambda_0)|\mathcal{F},\mathcal{Z}\} = \frac{1}{2I}\sum_{i=1}^{I}\sum_{j=1}^{2}\{(r_{Tij}-r_{Cij})-\lambda_0(d_{Tij}-d_{Cij})\},$$

so that (6) follows from the definition (1) of λ. Finally, (8) follows directly from the work of Gadbury (2001, sec. 3) with, for instance, his $X_i = (y_{Ti1,\lambda_0}+y_{Ti2,\lambda_0})/2$, $\epsilon_i = (y_{Ti2,\lambda_0}-y_{Ti1,\lambda_0})/2$.

For large I, the hypothesis $H_0^{(\lambda)}: \lambda=\lambda_0$ is tested by comparing $T(\lambda_0)/S(\lambda_0)$ to the standard normal cumulative distribution, $\Phi(\cdot)$. In the limiting argument here, with $I\to\infty$, there is no sampling of pairs from a population, but instead random treatment assignment is being applied to an ever-larger number, I, of pairs (e.g., Welch 1937). A moment's thought reveals that $T(\lambda)/S(\lambda)$ might not converge in distribution to $\Phi(\cdot)$ if, as pairs are added to the experiment, these new pairs became increasingly unstable (as they would, for instance, if the r_{Tij}'s were sampled independently from a Cauchy distribution). Proposition 2 is substantially more general than anything needed for the present work, because in the example the I inputs to $T(\lambda)/S(\lambda)$ share a finite support and have bounded moments of all orders. In particular, condition (9) permits the matched sets to become increasingly unstable as I increases but limits the rate at which this happens. In Proposition 2, it would be sufficient that I increased without bound over a set of values $I_1 < I_2 < I_3 < \cdots$, not necessarily $1, 2, \ldots$, with ρ_I and δ_I fixed.

Proposition 2. Consider a sequence of ever-larger paired randomized experiments, $(\mathcal{F}_I, \mathcal{Z}_I)$, where as the number I of pairs increases, $I\to\infty$, both $\rho_I = \frac{1}{2I}\sum_{i=1}^{I}\sum_{j=1}^{2}(r_{Tij}-r_{Cij})$ and $\delta_I = \frac{1}{2I}\sum_{i=1}^{I}\sum_{j=1}^{2}(d_{Tij}-d_{Cij})$ remain fixed at $\overline{\rho}$ and $\overline{\delta}$, with $\overline{\delta} > 0$. Write $\overline{\lambda} = \overline{\rho}/\overline{\delta}$. With $\vartheta_{Ii} = E\{|V_i(\overline{\lambda}) - \mu_{i,\overline{\lambda}}|^3|\mathcal{F}_I,\mathcal{Z}_I\}$ and $\kappa_{Ii} = E[\{V_i(\overline{\lambda})\}^4|\mathcal{F}_I,\mathcal{Z}_I]$, assume that

$$\begin{aligned} 0 = \limsup_{I\to\infty}\frac{\sum_{i=1}^{I}\vartheta_{Ii}}{(\sum_{i=1}^{I}\nu_{i,\overline{\lambda}})^{3/2}} \quad &\text{and} \\ \sum_{i=1}^{I}\kappa_{Ii} = o(I^2) \quad &\text{as } I\to\infty. \end{aligned} \tag{9}$$

Then, for each $k > 0$,

$$\limsup_{I\to\infty} \Pr\left\{\frac{T_I(\bar{\lambda})}{S_I(\bar{\lambda})} \le -k \,\middle|\, \mathcal{F}_I, \mathcal{Z}_I\right\} \le \Phi(-k) \qquad \text{and}$$
$$\limsup_{I\to\infty} \Pr\left\{\frac{T_I(\bar{\lambda})}{S_I(\bar{\lambda})} \ge k \,\middle|\, \mathcal{F}_I, \mathcal{Z}_I\right\} \le \Phi(-k). \tag{10}$$

Proof. The proof depends on two observations. (a) First, observe that the right-side condition in (9) ensures that the weak law of large numbers (Serfling 1980, sec. 1.8C, p. 27) applies to $IS_I^2(\bar{\lambda})$, which, by (8), ensures that for all $\epsilon > 0$, $\delta > 0$, there exists an I^* such that for $I \ge I^*$, $\delta > \Pr[IS_I^2(\bar{\lambda}) - I\operatorname{var}\{T_I(\bar{\lambda})|\mathcal{F}_I, \mathcal{Z}_I\} < -\epsilon]$. In words, in a sufficiently large experiment, it is nearly certain that $IS_I^2(\bar{\lambda})$ does not greatly underestimate $I\operatorname{var}\{T_I(\bar{\lambda})|\mathcal{F}_I, \mathcal{Z}_I\} = (1/I)\sum_{i=1}^{I} \nu_{i,\bar{\lambda}} = \varsigma_I$, say. (b) Second, by Proposition 1, $0 = \mathrm{E}\{T_I(\bar{\lambda})|\mathcal{F}, \mathcal{Z}\} = (1/I)\sum_{i=1}^{I} \mu_{i,\bar{\lambda}}$ for all I. From Proposition 1, the $V_i(\bar{\lambda}) - \mu_{i,\bar{\lambda}}$ are independent with expectation 0 and variance $\nu_{i,\bar{\lambda}}$, so, given $\mathcal{F}_I, \mathcal{Z}_I$, the quantity $\sqrt{I}T_I(\bar{\lambda}) = (1/\sqrt{I})\sum_{i=1}^{I}\{V_i(\bar{\lambda}) - \mu_{i,\bar{\lambda}}\}$ has expectation 0 and variance $(1/I)\sum_{i=1}^{I} \nu_{i,\bar{\lambda}}$. Using a version of the central limit theorem (thm. 9.2 in Breiman 1968, p. 186), the left-side condition in (9) implies that $\sqrt{I}T_I(\bar{\lambda})/\sqrt{\varsigma_I}$ converges in distribution to the standard normal distribution as $I \to \infty$. Combining (a) and (b) yields (10).

Remarks 1 and 2 consider an older and simpler situation than the main topic of this article—namely, the situation where $d_{Tij} - d_{Cij} = 1$ for all ij, so that there are simply treated subjects with $D_{ij} = Z_{ij} = 1$ and controls with $D_{ij} = Z_{ij} = 0$; that is, everyone is a complier. Remarks 1 and 2 relate to an old disagreement between Fisher and Neyman about the appropriate definition of "no treatment effect." Fisher (1935) defined no effect as $H_0 : r_{Tij} = r_{Cij}$, for $i = 1, \ldots, I$, $j = 1, 2$. In contrast, Neyman (1935) defined "no treatment effect" as no effect on average, which is essentially the same as $H_0 : \lambda = 0$ when $d_{Tij} - d_{Cij} = 1$ for all ij. For the current discussion, the key point is that Neyman's $H_0 : \lambda = 0$ is a composite hypothesis that includes Fisher's hypothesis such that (10) holds as an equality when Fisher's hypothesis is true; thus a test using p-values derived from (10) is not conservative as a test of Neyman's composite hypothesis, because the nominal level is achieved for large I when Fisher's hypothesis is true.

Remark 1. Under Fisher's sharp null hypothesis of no effect, $H_0 : r_{Tij} = r_{Cij}$, for $i = 1, \ldots, I$, $j = 1, 2$, the effect ratio λ equals 0, and $\mu_{i,\lambda} = 0$, so there is equality in (8) and (10). In this case, $T(0)/S(0)$ is the permutational t-statistic for testing the null hypothesis of no effect, and Propositions 1 and 2 describe its moments and limiting distribution, so in this case, the results closely resemble results of Fisher (1935), Welch (1937), and Robinson (1973), among others.

Remark 2. If $d_{Tij} - d_{Cij} = 1$ for all ij, then λ in (1) is the average treatment effect, where the effect $r_{Tij} - r_{Cij}$ may vary from one subject to another. In this case, Propositions 1 and 2 describe the behavior of the permutational t-statistic in testing the composite hypothesis that the average treatment effect λ is some number λ_0. (In this case, there is a link to Neyman 1935 and Gadbury 2001.) If the treatment effect were an additive constant, $r_{Tij} - r_{Cij} = \lambda_0$ for all ij, then (a) $\mu_{i,\lambda_0} = 0$ for all i; (b) expression (8) equals 0 and there is equality in (10); (c) as $I \to \infty$, a test that rejects $H_0^c : r_{Tij} - r_{Cij} = \lambda_0$ for all ij when $T_I(\lambda_0)/S_I(\lambda_0) \ge k$ has size $\Phi(-k)$; and (d) because H_0^c is one of the hypotheses in the composite hypothesis about the average treatment effect, $H_0^{(\lambda)} : \lambda = \lambda_0$, as $I \to \infty$, the size of the test of the composite hypothesis tends to $\Phi(-k)$.

Remark 3 is parallel to Remarks 1 and 2 except for the removal of the restriction that $d_{Tij} - d_{Cij} = 1$. In particular, within the composite hypothesis $H_0^{(\lambda)} : \lambda = \lambda^*$, there is a specific hypothesis (11) such that equality holds in (10).

Remark 3. The model that asserts that the effect of the treatment Z_{ij} on (r_{Tij}, r_{Cij}) is proportional to its effect on (d_{Tij}, d_{Cij}) asserts that there is a λ^* such that

$$r_{Tij} - r_{Cij} = \lambda^*(d_{Tij} - d_{Cij}) \quad \text{for } i = 1, \ldots, I, j = 1, 2, \tag{11}$$

and in this case λ in (1) equals λ^* and $\mu_{i,\lambda} = 0$, so with $\lambda_0 = \lambda^*$ expression (8) equals 0, and there is equality in (10). So, as in Remark 2, because (11) is one of the hypotheses in the composite hypothesis $H_0^{(\lambda)} : \lambda = \lambda^*$, as $I \to \infty$, the size of the test that rejects when $T_I(\lambda)/S_I(\lambda) \ge k$ tends to $\Phi(-k)$.

In a randomized clinical trial, say, we genuinely randomize treatment assignment, but the patients in the trial are not a random sample from a population. Remarks 4 and 5 connect Propositions 1 and 2 to random samples from an infinite population, as opposed to randomized treatment assignment in a finite population. In particular, there is a sense, admittedly informal, in which the inequality in (10) would be an equality if we were sampling an infinite population. Importantly, Proposition 2 shows that $T_I(\widetilde{\lambda})/S_I(\widetilde{\lambda})$ yields appropriate inferences without the fanciful notion that randomized experiments are performed on a random sample from a population. Also, Remark 5 shows that the common linear structural equation (12) is a special case of the hypothesis (11), which is a special case of the composite hypothesis $H_0^{(\lambda)} : \lambda = \lambda^*$.

Remark 4. Imagine that $\mathcal{F}$ was obtained by sampling a superpopulation of matched pairs such that (a) distinct pairs are mutually independent; (b) within pairs, subjects are exchangeable but perhaps not independent; (c) the distribution of $(r_{Tij}, r_{Cij}, d_{Tij}, d_{Cij}, \mathbf{x}_{ij}, u_{ij})$ is the same for all ij; (d) $(r_{Tij}, r_{Cij}, d_{Tij}, d_{Cij})$ have expectations and variances; and (e) $E(d_{Tij}) - E(d_{Cij}) > 0$. Then write $\widetilde{\lambda} = E(r_{Tij} - r_{Cij})/E(d_{Tij} - d_{Cij})$. In this superpopulation, the effect ratio λ_I based on a sample of I pairs in Proposition 2 is a random variable that converges in probability to $\widetilde{\lambda}$ as $I \to \infty$. Also, in the superpopulation (i.e., without conditioning on $\mathcal{F}$), the quantity $V_i(\widetilde{\lambda})$ has expectation 0 and constant variance $\sigma^2 = E\{V_i(\widetilde{\lambda})^2\}$, so that $I \cdot S^2(\widetilde{\lambda})$ converges in probability to σ^2. Also, unconditionally, the $V_i(\widetilde{\lambda})$ are iid, so $T_I(\widetilde{\lambda})/S_I(\widetilde{\lambda})$ converges in distribution to $\Phi(\cdot)$. This is an alternative view of the approximation (10).

Remark 5. The most basic view of instrumental variables links them to a linear structural equation,

$$R_{ij} = \theta_i + \lambda^* D_{ij} + \varepsilon_{ij} \quad \text{with } \varepsilon_{ij} \perp\!\!\!\perp Z_{ij}, \tag{12}$$

and the current remark relates structural equations to Propositions 1 and 2. Unlike a regression, in a linear structural equation (12), it is imagined that if D_{ij} were changed to $D_{ij} + \delta$, then R_{ij} would change to $R_{ij} + \delta\lambda^*$, in accordance with (12). In (12), θ_i is a fixed, unknown matched-pair parameter linking observations in the same pair. In $T(\lambda_0)$, differencing eliminates θ_i. Contrast setting $D_{ij} = d_{Tij}$ with response $R_{ij} = r_{Tij}$, say, and $D_{ij} = d_{Cij}$ with response $R_{ij} = r_{Cij}$, say, in (12). Then, using (12), it follows that $r_{Tij} - r_{Cij} = \lambda^*(d_{Tij} - d_{Cij})$, so that (11) holds, and once again, λ in (1) equals λ^*, $\mu_{i,\lambda} = 0$, and expression (8) equals 0, and there is equality in (10). In this case $T_I(\widetilde{\lambda})/S_I(\widetilde{\lambda})$ is similar to the Anderson–Rubin (1949) statistic and the solution to $T_I(\widetilde{\lambda}) = 0$ is the Wald (1940) estimator.

3.4 Application to the Study of Perinatal Care

Recall that the effect ratio, λ, is the ratio of the increase in mortality to the increase in use of a low-level NICU that occurs with increased distance to a high-level NICU. Under the exclusion restriction, λ is the effect on mortality among mothers who would change the level of the NICU depending on their distance from a high-level NICU. Recall from Table 2 that the infant mortality rate for mothers far from a high-level NICU was on the order of 2%. Among mothers who would switch from a low-level NICU to a high-level NICU if one were close, what is the estimated reduction in mortality?

In Table 3, the 95% confidence interval (CI) for λ is the solution to $T(\lambda_0)/S(\lambda_0) = \pm 1.96$, and the point estimate is the solution to $T(\lambda_0)/S(\lambda_0) = 0$. In Table 3, the point estimates from the two matched samples are similar, but the CI is shorter with a stronger instrument. (This is not the principal reason for preferring a stronger instrument; see Section 4.)

The point estimate, 0.0090, is substantial, almost half the infant mortality for mothers living far from a high-level NICU. The lower endpoint of the 95% CI from the strong instrument, 0.0057, is also substantial; it is more than one-quarter of the infant mortality for mothers living far from a high-level NICU.

It is natural to ask how Table 3 compares with two-stage least squares applied to all the babies, with excess travel time as an instrument for a low-level NICU. It should be emphasized that two-stage least squares is not strictly appropriate here, for several reasons. Using all of the babies means that most mothers live in or near urban areas, and excess travel time rarely decides where mother delivers, so it is a weak instrument in this case. Two-stage least squares can give misleading answers with a weak instrument (Bound, Jaeger, and Baker 1995), whereas this problem does not arise with pivotal methods of the type in Section 3.3 (see Imbens and Rosenbaum 2005). Moreover, both R_{ij} and D_{ij} are binary, but two-stage least squares ignores this, producing 4965 negative predicted values for D_{ij} and 4236 predicted values for D_{ij} that are >1; also, 97,035 babies (49%) have negative predicted probabilities of death in the second stage. Conceivably, negative probabilities of death for half of the babies do no harm in two-stage least squares, but they are at least disconcerting, and perhaps worrisome. In contrast, in Section 3.3 binary responses are treated as binary responses. With these caveats in mind, two-stage least squares yields a point estimate of λ of 0.0083 and a 95% CI [0.0050, 0.0116], with length 0.0067; thus, compared with the strong instrument in Table 3, the two-stage least squares yields an estimated effect that is about 8% smaller (0.0083 vs. 0.0090) with a CI that is slightly longer.

The inferences in Table 3 assume that within pairs matched for covariates, living close to or far from a high-level NICU occurs at random; that is, $\Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ for each $\mathbf{z} \in \Omega$. In the next section, Section 4, we consider the possibility that this assumption is false.

4. SENSITIVITY ANALYSIS: WHAT IF THE INSTRUMENT IS NOT RANDOMLY ASSIGNED?

4.1 General Method: Quantifying Departures From Random Assignment

In previous sections, our inferences acted as if, within pairs matched for $\mathbf{x}_{ij}$, proximity to a high-level NICU is random, in the sense that $\Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ for each $\mathbf{z} \in \Omega$. The sensitivity analysis asks how unmeasured biases in assignment of proximity might alter these inferences. The sensitivity analysis imagines that before matching, mother ij had a probability $\pi_{ij} = \Pr(Z_{ij} = 1|\mathcal{F})$ of living far from a high-level NICU with independent assignments for distinct mothers, and two mothers, say ij and ij', who might be matched because they have the same observed covariates, $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$, might differ in their odds of living far from a high-level NICU by at most a factor of $\Gamma \geq 1$, so

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1 - \pi_{ij'})}{\pi_{ij'}(1 - \pi_{ij})} \leq \Gamma, \quad \text{for all } i, j, j', \text{ with } \mathbf{x}_{ij} = \mathbf{x}_{ij'}. \quad (13)$$

Then the distribution of $\mathbf{Z}$ is returned to Ω by conditioning on the event $\mathcal{Z}$ that $\mathbf{Z} \in \Omega$. It is straightforward to show that this sensitivity model is exactly equivalent to assuming that for $\mathbf{z} \in \Omega$,

$$\Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathcal{Z}) = \frac{\exp(\gamma \mathbf{z}^T \mathbf{u})}{\sum_{\mathbf{b} \in \Omega} \exp(\gamma \mathbf{b}^T \mathbf{u})} \quad \text{with } \mathbf{u} \in [0, 1]^{2I}, \quad (14)$$

where $\gamma = \log(\Gamma)$. [See Rosenbaum 1995, sec. 1.2, 2002, sec. 4.2 for the quick, elementary steps demonstrating the equivalence of (13) and (14), and see Wang and Krieger 2006 for related discussion.] If $\Gamma = 1$, so $\gamma = 0$, then $\pi_{ij} = \pi_{ij'}$ whenever $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$ in (13) and $\Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ in (14) is the randomization distribution. For fixed $\Gamma > 1$, the $\pi_{ij} = \Pr(Z_{ij} = 1|\mathcal{F})$ are unknown to a bounded degree, so that an inference quantity, such as a p-value or an estimate, is unknown but confined to an interval. For several values of Γ, a sensitivity analysis computes the range of possible inferences (say, the range

Table 3. Inference about the effect ratio λ under the assumption of random assignment of excess travel time within pairs matched for covariates

	Weaker instrument 99,174 pairs of two babies			Stronger instrument 49,587 pairs of two babies	
Point estimate		0.0092			0.0090	
95% CI	0.0036		0.0148	0.0057		0.0123
Length of 95% CI		0.0112			0.0066	

of possible p-values), thereby indicating the magnitude of bias that would need to be present to alter the qualitative conclusions reached assuming random assignment.

As noted in Section 3.1, Fisher's sharp null hypothesis of no treatment effect on (r_{Tij}, r_{Cij}) asserts that $H_0: r_{Tij} = r_{Cij}$, for all ij. As noted in Section 3.2, if H_0 were true, then the effect ratio λ would be 0, $I \cdot T(0)$ equals $\sum_{i=1}^{I}\{\sum_{j=1}^{2} Z_{ij} r_{Cij} - \sum_{j=1}^{2}(1 - Z_{ij}) r_{Cij}\}$, and the randomization distribution of $T(0)$ yields the same p-values for testing Fisher's null hypothesis H_0 as the permutational t-test (e.g., Welch 1937), and it was used in Section 3.3 to test $H_0^{(\lambda)}: \lambda = 0$. If Fisher's H_0 were true, then standard methods of sensitivity analysis might be applied to $T(0)$. (See Rosenbaum 1987, 1991, 2002, secs 4.4–5, 2007; Rosenbaum 1999 for a sensitivity analysis with an instrument; and, e.g., Gastwirth 1992, Marcus 1997, Lin, Psaty, and Kronmal 1998, Robins, Rotnitzky, and Scharfstein 1999, Copas and Eguchi 2001, Imbens 2003, and Small 2007 for alternative methods of sensitivity analysis.)

4.2 Application to the Study of Regionalization of Perinatal Care

In the case of matched pairs with binary responses, as in Section 1, say that pair i is discordant if it contains exactly one death, $R_{i1} + R_{i2} = 1$, and let $I^* \le I$ be the number of discordant pairs and $\mathcal{D}$ be the set of the indices i of the I^* discordant pairs, so that $|\mathcal{D}| = I^*$. If Fisher's sharp null hypothesis of no effect, $H_0: r_{Tij} = r_{Cij}$ for all ij, were true, then the number of pairs with $R_{i1} + R_{i2} = 0$, $R_{i1} + R_{i2} = 1$, and $R_{i1} + R_{i2} = 2$ would be determined by $\mathcal{F}$, and hence fixed by conditioning on $\mathcal{F}$, but whether or not the one death in a discordant pair is a treated death (i.e., whether $\sum_{j=1}^{2} Z_{ij} R_{ij}$ equals 1 or 0) is not a function of $\mathcal{F}$ alone, and is determined by the treatment assignment Z_{ij} within discordant pairs. In testing Fisher's H_0 in matched pairs with binary responses, the distribution of $T(0)$ under (14) receives a nondegenerate contribution from matched pair i only if the pair is discordant. In this case $T(0)$ is effectively the same as McNemar's statistic; that is, under H_0, as $\mathbf{z}$ varies over Ω, the statistic $T(0)$ is a linear function of the number of deaths, T^*, among treated subjects in discordant pairs, $T^* = \sum_{i \in \mathcal{D}} \sum_{j=1}^{2} Z_{ij} R_{ij}$. In a randomized experiment under H_0, the randomization distribution of T^* is binomial with sample size I^* and probability of success 1/2. Under H_0, the bounds on p-values from (14) are provided by comparing T^* to two binomial distributions, one with sample size I^* and probability of success $\Gamma/(1+\Gamma)$ and the other with sample size I^* and probability of success $1/(1+\Gamma)$ (see Rosenbaum 1987, 1991, 2002, sec. 4 for detailed discussion).

Tables 4 and 5 display the data in the form used for McNemar's test. Specifically, these tables count pairs, and discordant pairs fall in the off-diagonal cells. In Table 4 there are $I^* = 554 + 748 = 1302$ discordant pairs, and the upper bound 0.037 on the one-sided p-value is obtained by comparing 748 deaths among distant mothers to the binomial with 1302 trials and probability $\Gamma/(1+\Gamma) = 1.22/(1+1.22)$ of an event. As is shown more clearly in Table 6, the two quoted values of Γ in Tables 4 and 5 ($\Gamma = 1.07$ and $\Gamma = 1.22$) are to two decimals the values of Γ in which the conventional 0.05 significance level is achieved. In Tables 4 and 5 the larger study with a weaker instrument is quite a bit more sensitive to unmeasured biases ($\Gamma = 1.07$ vs. $\Gamma = 1.22$), despite the larger sample size, which is precisely the prediction of statistical theory (Small and Rosenbaum 2008).

Table 4. Mortality in the 25-minute, 50-sinks match with 49,587 pairs. The upper bound on the one-sided p-value is 0.037 for $\Gamma = 1.22$

		Near high-level NICU $Z_{ij} = 0$	
		Alive, $R_{ij} = 0$	Dead, $R_{ij} = 1$
Far from high-level NICU, $Z_{ij} = 1$	Alive, $R_{ij} = 0$	48,070	554
	Dead, $R_{ij} = 1$	748	215

In brief, with a strong instrument in Table 4, results are sensitive to an unmeasured bias of magnitude $\Gamma > 1.22$, whereas with a weak instrument in Table 5, results are sensitive to an unmeasured bias of magnitude $\Gamma \ge 1.07$. To put this in perspective using techniques not described here, an unobserved covariate associated with a doubling of the odds of death and a doubling of the odds of delivering at a low-level NICU corresponds to $\Gamma = 1.25$, whereas an unobserved covariate associated with a doubling of the odds of death and a 25% increase in the odds of delivering at a low-level NICU corresponds to $\Gamma = 1.08$. [See Gastwirth, Krieger, and Rosenbaum 1998 and Rosenbaum and Silber 2009b for a detailed discussion of two correspondences between one-parameter (Γ) and two-parameter sensitivity analyses of the type just mentioned.]

Although Tables 4 and 5 pay attention to which mother in a pair has a greater excess travel time to a high-level NICU, they ignore the actual magnitude of the time. For the match with the stronger instrument, the mean difference is about 34 minutes, but this difference does vary from pair to pair. Presumably, the encouragement to deliver at a low-level NICU is greater if the excess travel time to a high-level NICU is 45 minutes rather than 25 minutes. Would the findings be different if we took into account the magnitude of the difference in excess travel time? This is a natural question to ask because one conventional method (two-stage least squares) does take into account such magnitudes. McNemar's test focuses on pairs discordant for infant mortality, relating mortality in these pairs to the binary indicator of proximity. Among randomization tests, a familiar test that takes into account magnitudes is Wilcoxon's signed-rank test applied within pairs discordant for mortality, where the test is applied to the difference in magnitude of excess travel time. Wilcoxon's test gives greater weight to a discordant pair if the difference in travel times is larger. (See Rosenbaum 1991, 2002, sec. 4 for details on the sensitivity analysis for Wilcoxon's test

Table 5. Mortality in the 0-minute, 0-sinks match, with 99,174 pairs. The upper bound on the one-sided p-value is 0.070 for $\Gamma = 1.07$ and 0.97 for $\Gamma = 1.22$

		Near high-level NICU $Z_{ij} = 0$	
		Alive, $R_{ij} = 0$	Dead, $R_{ij} = 1$
Far from high-level NICU, $Z_{ij} = 1$	Alive, $R_{ij} = 0$	96,044	1226
	Dead, $R_{ij} = 1$	1391	574

Table 6. Sensitivity analysis, unweighted and weighted, with a stronger and a weaker instrument. The table gives upper bounds on the one-sided p-value for testing no effect on mortality for a given value of Γ. In each column, the last p-value less than or equal to 0.05 is in **bold**

	Instrument			
	Weaker		Stronger	
Measure:	Mortality	Weighted	Mortality	Weighted
Γ	$n = 99{,}235$	$n = 99{,}235$	$n = 49{,}587$	$n = 49{,}587$
1	0.0006	0.0001	0.0000	0.0000
1.05	**0.0239**	0.0034	0.0000	0.0000
1.1	0.2147	**0.0346**	0.0001	0.0004
1.15	0.6348	0.1671	0.0021	0.0040
1.2	0.9238	0.4401	0.0177	0.0233
1.22	0.9681	0.5659	0.0352	**0.0414**
1.23	0.9804	0.6263	**0.0481**	0.0539
1.24	0.9884	0.6834	0.0644	0.0690
1.25	0.9933	0.7360	0.0845	0.0871

applied to pairs discordant for a binary outcome.) Table 6 displays four sensitivity analyses, two with the stronger instrument and two with the weaker instrument, and two using McNemar's test and two using a weighted test. In Table 6 the weighting is of some help to the weak instrument—it downweights pairs in which the difference in excess travel time is too small to influence hospital choice—but there is less sensitivity to unmeasured bias with a stronger instrument, despite the reduction in sample size.

5. DISCUSSION: WHAT CHANGES WHEN AN INSTRUMENT IS STRENGTHENED?

Pairing all of the babies in Pennsylvania using observed covariates yields 99,174 pairs and a weak instrument. Pairing about half of the babies in Pennsylvania using observed covariates and excess travel time yields 49,587 pairs and a much stronger instrument. Making an instrument stronger in this way changes a few things that must be noted; however, none of the changes are particularly worrisome, because they were produced in a known, algorithmic way using only observed covariates and travel time.

In the first instance, the population under study has changed slightly, but the changes are clearly indicated in Table 1, because these are the variables used to change the population. The biological aspects of babies and mothers are largely the same in the two matched samples, as are measures of education and income. Notable in Table 1 is the reduction in the proportion of blacks from 15% in the 99,174 pairs to about 5% in the 49,587 pairs. Why did this happen? Because most blacks in Pennsylvania live in or near urban areas, they are typically close to a hospital with a high-level NICU, and it is hard to pair them with blacks living far from high-level NICUs. The larger match also contains slightly more people who rent rather than own their homes, and slightly fewer mothers with fee-for-service health insurance (e.g., Blue Cross) and slightly more with a health maintenance organization (HMO). Within pairs, these covariates are balanced, but the population of pairs has shifted slightly. In brief, the smaller match is explicitly less often black and implicitly less often urban. In terms of the shift in the population, when building a stronger instrument, the investigator should describe and discuss the shift using, for instance, a table similar to Table 1.

In the second instance, if the instrument is stronger, then mothers are more likely to comply, and thus the meaning of a "complier" has changed. Importantly, we did not use compliance behavior in building the matched sample; rather, we used excess travel time, whether or not travel time influenced where a mother delivered. In the larger match, the average difference in travel time within pairs was less than 14 minutes, whereas in the smaller match, it was more than 34 minutes. Imagine being in labor with the knowledge that it will take an extra 34 minutes to reach a hospital with high-level NICU beyond the time it takes to reach a hospital with a low-level NICU. It is easy to imagine a mother who would comply in response to 34 extra minutes, but not to 14 extra minutes. It is not the mother that changes; rather, it is the incentive on an offer for compliance. To the extent that the Wald estimator estimates the average causal effect on compliers (Angrist, Imbens, and Rubin 1996), it is estimating an average over different groups of mothers with a strong instrument and a weak instrument. If one believes that the typical mother will comply for an extra 34 minutes but not for an extra 15 minutes, then the smaller match with a stronger instrument will be somewhat more likely to describe the effect for a typical mother. That is, the smaller match looks a little less like Pennsylvania than the larger match, but compliance behavior is normal behavior in the smaller match, and it is less common behavior in the larger match, so an average effect over compliers is an average over normal mothers in the smaller match and an average over somewhat unusual mothers in the larger match. We would prefer a study in which a strong incentive to comply was offered to some mothers and denied to others in an essentially random manner; the typical mother would then respond to the strong incentive.

6. SUMMARY: STRONGER INSTRUMENTS BY DESIGN

In Pennsylvania, excess travel time is a fairly weak instrument for delivery at a hospital with a low-level NICU, because most people live in or near urban areas, so they live close to several hospitals of varied capabilities. We could have accepted Pennsylvania as it is, accepting also a weak instrument, or could have searched for another state or cross-state region whose geography made excess travel time into a stronger instrument. Instead, we built a matched study in which very similar mothers and babies were paired with very different excess travel times; that is, we built a study with a stronger instrument. Theory from Small and Rosenbaum (2008) and the empirical results presented here support the conclusion that a smaller study with a strong instrument is preferable to a larger study with a weak instrument. Confidence intervals were shorter and conclusions were less sensitive to unmeasured biases in the smaller but stronger matched comparison.

[Received August 2009. Revised February 2010.]

REFERENCES

Anderson, T. W., and Rubin, H. (1949), "Estimations of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 20, 46–63. [1293]

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables" (with discussion), *Journal of the American Statistical Association*, 91, 444–455. [1285,1290,1295]

Avriel, M. (1976), *Nonlinear Programming*, NJ: Prentice Hall. [1288]

Bound, J., Jaeger, D. A., and Baker, R. M. (1995), "Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443–450. [1286,1293]

Breiman, L. (1968), *Probability*, Reading, MA: Addison Wesley. [1292]

Copas, J., and Eguchi, S. (2001), "Local Sensitivity Approximations for Selectivity Bias," *Journal of the Royal Statistical Society, Ser. B*, 63, 871–896. [1294]

Derigs, U. (1988), "Solving Nonbipartite Matching Problems by Shortest Path Techniques," *Annals of Operations Research*, 13, 225–261. [1286]

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd. [1292]

Gadbury, G. L. (2001), "Randomization Inference and the Bias of Standard Errors," *The American Statistician*, 55, 310–313. [1291,1292]

Gastwirth, J. L. (1992), "Methods for Assessing the Sensitivity of Statistical Comparisons Used in Title VII Cases to Omitted Variables," *Jurimetrics*, 33, 19–34. [1294]

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998), "Dual and Simultaneous Sensitivity Analysis for Matched Pairs," *Biometrika*, 85, 907–920. [1294]

Imbens, G. W. (2003), "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, 93, 126–132. [1294]

Imbens, G., and Rosenbaum, P. R. (2005), "Robust, Accurate Confidence Intervals With a Weak Instrument: Quarter of Birth and Education," *Journal of the Royal Statistical Society, Ser. A*, 168, 109–126. [1286,1293]

Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998), "Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies," *Biometrics*, 54, 948–963. [1294]

Lu, B. (2005), "Propensity Score Matching With Time-Dependent Covariates," *Biometrics*, 61, 721–728. [1287]

Lu, B., and Rosenbaum, P. R. (2004), "Optimal Matching With Two Control Groups," *Journal of Computational and Graphical Statistics*, 13, 422–434. [1287]

Lu, B., Greevy, R., Xu, X., and Beck, C. (2009), "Optimal Nonbipartite Matching and Its Statistical Applications," *The American Statistician*, to appear. [1286]

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. [1287]

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. [1294]

Neyman, J. (1923, 1990), "On the Application of Probability Theory to Agricultural Experiments," *Statistical Science*, 5, 463–480. [1289]

——— (1935), "Statistical Problems in Agricultural Experimentation," *Journal of the Royal Statistical Society*, 2 (Supplement), 107–180. [1292]

Robins, J. M., Rotnitzky, A., and Scharfstein, D. (1999), "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference," in *Statistical Models in Epidemiology*, eds. E. Halloran and D. Berry, New York: Springer, pp. 1–94. [1294]

Robinson, J. (1973), "The Large Sample Power of Permutation Tests for Randomization Models," *The Annals of Statistics*, 1, 291–296. [1292]

Rogowski, J. A., Horbar, J. D., Staiger, D. O., Kenny, M., Carpenter, J., and Geppert, J. (2004), "Indirect vs Direct Hospital Quality Indicators for Very Low-Birth-Weight Infants," *Journal of the American Medical Association*, 291, 202–209. [1286]

Rosenbaum, P. R. (1987), "Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies," *Biometrika*, 74, 13–26. [1294]

——— (1991), "Sensitivity Analysis for Matched Case-Control Studies," *Biometrics*, 47, 87–100. [1294]

——— (1995), "Quantiles in Nonrandom Samples and Observational Studies," *Journal of the American Statistical Association*, 90, 1424–1431. [1293]

——— (1999), "Using Combined Quantile Averages in Matched Observational Studies," *Applied Statistics*, 48, 63–78. [1294]

——— (2002), *Observational Studies* (2nd ed.), New York: Springer-Verlag. [1293,1294]

——— (2004), "Design Sensitivity in Observational Studies," *Biometrika*, 91, 153–164. [1286]

——— (2005a), "Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies," *The American Statistician*, 59, 147–152. [1286]

——— (2005b), "An Exact, Distribution Free Test Comparing Two Multivariate Distributions Based on Adjacency," *Journal of the Royal Statistical Society, Ser. B*, 67, 515–530. [1287]

——— (2007), "Sensitivity Analysis for *m*-Estimates, Tests and Confidence Intervals in Matched Observational Studies," *Biometrics*, 63, 456–464. [1294]

——— (2010), *Design of Observational Studies*, New York: Springer. [1288]

Rosenbaum, P. R., and Silber, J. H. (2009a), "Sensitivity Analysis for Equivalence and Difference in an Observational Study of Neonatal Intensive Care Units," *Journal of the American Statistical Association*, 104, 501–511. [1287]

——— (2009b), "Amplification of Sensitivity Analysis in Observational Studies," *Journal of the American Statistical Association*, 104, 1398–1405. [1294]

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701. [1289]

——— (1980), "Bias Reduction Using Mahalanobis Metric Matching," *Biometrics*, 36, 293–298. [1288]

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley. [1292]

Silber, J. H., Lorch, S. L., Rosenbaum, P. R., Medoff-Cooper, B., Bakewell-Sachs, S., Millman, A., Mi, L., Even-Shoshan, O., and Escobar, G. E. (2009), "Additional Maturity at Discharge and Subsequent Health Care Costs," *Health Services Research*, 44, 444–463. [1287]

Small, D. (2007), "Sensitivity Analysis for Instrumental Variables Regression With Overidentifying Restrictions," *Journal of the American Statistical Association*, 102, 1049–1058. [1294]

Small, D., and Rosenbaum, P. R. (2008), "War and Wages: The Strength of Instrumental Variables and Their Sensitivity to Unobserved Biases," *Journal of the American Statistical Association*, 103, 924–933. [1286,1287,1294, 1296]

Wald, A. (1940), "The Fitting of Straight Lines If Both Variables Are Subject to Error," *Annals of Mathematical Statistics*, 11, 284–300. [1293]

Wang, L. S., and Krieger, A. (2006), "Causal Conclusions Are Most Sensitive to Unobserved Binary Covariates," *Statistics in Medicine*, 25, 2256–2271. [1293]

Welch, B. L. (1937), "On the z-Test in Randomized Blocks and Latin Squares," *Biometrika*, 29, 21–52. [1291,1292,1294]