2010

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

Mike Baiocchi
University of Pennsylvania

Dylan Small
University of Pennsylvania

Scott A. Lorch
University of Pennsylvania

Paul R. Rosenbaum
University of Pennsylvania

Recommended Citation

Baiocchi, M., Small, D., Lorch, S. A., & Rosenbaum, P. R. (2010). Building a Stronger Instrument in an Observational Study of Perinatal Care for Premature Infants. *Journal of the American Statistical Association, 105* (492), 1285-1296. http://dx.doi.org/10.1198/jasa.2010.ap09490

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

Abstract

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.

Keywords

design sensitivity, effect ratio, instrumental variable, nonbipartite matching, observational study, optimal matching, sensitivity analysis

Disciplines

Statistics and Probability

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

Mike Baiocchi, Dylan S. Small, Scott Lorch, Paul R. Rosenbaum[1]

University of Pennsylvania, Philadelphia

Abstract. An instrument is a random nudge towards acceptance of a treatment which affects outcomes only to the extent that it affects acceptance of the treatment. In settings in which treatment assignment is mostly deliberate and not random, there may nonetheless exist some essentially random nudges to accept treatment, so that use of an instrument may 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 they are often decisive in influencing treatment assignment. Although one hopes that an ostensibly random instrument is perfectly random and not biased, it is not possible to be certain of this, so a typical concern is that even the instrument is biased to some degree. It is known from theoretical arguments that weak instruments are invariably sensitive to extremely small biases, so 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. Also, we develop methods of permutation inference for effect ratios, a key component in an instrumental variable analysis.

Keywords: Design sensitivity; effect ratios; instrumental variable; nonbipartite matching; observational study; optimal matching; sensitivity analysis.

[1] *Address for correspondence:* Department of Statistics, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 USA. 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. 15 February 2010 E-mail: rosenbaum@stat.wharton.upenn.edu.

1 Introduction: Motivation; Example; 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, namely 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 based on haphazard circumstances such as affiliation or proximity. Regionalized perinatal systems were developed in the 1970's when NICUs began to save infants with birth weight under 1500 grams. In the 1990's, however, neonatal intensive care 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. Regionalization might not reduce infant mortality in any of several ways: the sorting by risk might be too inaccurate to affect health, or the capabilities of high level NICU's 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 we inhabit, a randomized experiment could settle that question, with high risk mothers assigned at random to hospitals of varied capabilities. In the world we inhabit, a world in which medical decisions are happily constrained by considerations of sound judgement, ethics and patient preferences, such an experiment is not possible. We need to make some

reasonable sense of data we can obtain. There is, however, a basic difficulty, one 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, the highest risk mothers would deliver at the most capable hospitals. The mortality rates at the most capable hospitals might be higher, not lower, than the mortality rates at less capable hospitals because their patient populations are sicker, even if the more capable hospitals were saving lives. A naïve comparison of mortality rates by level of NICU will 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.

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 there is one close to home. A pregnancy may conclude with a certain urgency, and awareness of this possibility may lead 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 receives care at that hospital, then the so-called 'exclusion restriction' would be plausible; see Angrist, Imbens and Rubin (1996) for discussion of the 'exclusion restriction'. If it were also true that mother's risk is unrelated to geography, proximity would be an instrument for care at a hospital with a high level NICU. In point of fact, mother's risk is related to geography, largely through socioeconomic factors that vary with geography, but we attempt to control for this issue and many others by matching for measured covariates.

Proximity would be a strong instrument for delivery at a hospital with a high level NICU if proximity were typically decisive in determining where mother delivered. Proximity would be a weak instrument if it were a minor factor among many others. For 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, 2005) for general discussion of the power of a sensitivity analysis. Consider two studies, one with a strong instrument, the other with a weak instrument. If one assumed that the instrument was randomly assigned, then the problems caused by a weak instrument might be offset by a sufficiently large sample size. However, Small and Rosenbaum showed that if one takes account of 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 may tend to zero with increasing sample size for a magnitude of bias such that the power with a strong instrument is tending to one. In this paper, we demonstrate that for a single large study with a weak instrument, we can, by careful design, extract from it 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 six months of 2005; that is, approximately 200,000 births. The data combine information from birth and death certificates and a form, UB-92, that hospitals provide.

Regionalization is a policy that would alter the level of the neonatal intensive care unit

(NICU) at which a high risk mother would deliver; it is neither aimed at improving prenatal care, nor is it a sensible strategy for improving prenatal care. Because we are interested in comparing the effectiveness of the neonatal care provided by different levels of NICUs on newborns, we regard variables that are determined prior to birth as covariates. To the extent possible, we would like to compare babies similar at birth who received the same prenatal care but who received neonatal care at NICUs of different levels. We do not want to confuse an effect of the level of the NICU 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 NICUs, but 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 a NICU but might be affected by prenatal care, for instance by coaxing a mother to abstain from smoking.

Following Rogowski et al. (2004), a mother is recorded as having delivered at a low level hospital ($D = 1$) if that hospital delivered an average of fewer than 50 premature babies per year or if its NICU is below level 3A, whereas, otherwise, she is recorded as having delivered at a high level hospital ($D = 0$) if that hospital delivers at least 50 premature babies per year and has a NICU of level 3A-3D or 4. We ask: Does delivery of a premature infant at a low level hospital increase risk of death, and if it does, then by how much?

Travel time was determined using ArcView software from ESRI, Inc, as the time from the centroid of mother's zip code to the closest low 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 the *excess travel time*. Excess travel time takes negative values if the closest hospital has a high level NICU. Distance strongly encourages

mother to deliver at a low level hospital if this difference in travel time is positive and large.

Stop for a moment and think about Pennsylvania. There are two large cities, Philadelphia and Pittsburgh, several medium size cities such as Harrisburg and 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. Inside Philadelphia, there are many hospitals often within walking distance of one another, so excess travel times are small, and excess travel time will rarely decide where mother delivers. In a rural area, excess travel time may be decisive. 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 are $2I$ babies. First, a discrepancy is defined between every pair of babies, yielding a $2I \times 2I$ discrepancy matrix. (The term 'discrepancy' is used in place of the more common term 'distance' to avoid confusion of covariate discrepancy with the geographic distance to a NICU.) An optimal nonbipartite matching then divides the $2I$ babies into I nonoverlapping pairs of two babies in such a way that the sum of

the discrepancies within the I pairs is minimized. That is, two babies in the same pair are 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, Greevy, Xu and Beck (2009). For statistical applications of optimal nonbipartite matching, see Lu, et al. (2001), Rosenbaum and Lu (2004), Lu (2005), and Rosenbaum (2005), 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 the babies (about 100,000 babies), permitting pairs which are closely matched for covariates, yet with substantial differences in excess travel time. In the second matching, we have about 50,000 pairs of two babies, closely matched for covariates, one far from the nearest high level NICU, 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, where each sink is 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 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 to other mothers from wealthy communities, and mothers from poor communities to 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 every pair of babies. A small penalty (i.e., a positive number) was added to the discrepancy for each of the following circumstances for any pair of babies which: (i) did not agree on the number of congenital disorders, (ii) did not agree on black race, (iii) 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 $2L$, 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 which are most systematically out of balance; see Rosenbaum (2010, §9.2) for 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, where $\Lambda = 0$ in the first match described above 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 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 zero if the constraint is respected and rises rapidly as the magnitude of the violation of the constraint increases; see Avriel (1976) for discussion of penalty functions and see Rosenbaum (2010, §8.4) for discussion of the use of penalty functions in matching.

In fact, we matched exactly on three important covariates. One was year of birth. The other two covariates that were exactly matched were 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. In addition to ensuring exact matches on these three covariates, this permits a rather large matching problem ($\sim 200,000$ babies) to be broken into several smaller problems that are solved separately in the manner indicated above. 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, §9.3) for discussion. Inside these exact match categories, we also used the continuous versions of birth weight and gestational age to obtain closer matches than the categories alone required.

2.2 Two matched comparisons, one stronger, the other 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. Remember, 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 the babies, forming 99,174 pairs of two 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 two babies.

In Table 1, the two matched comparisons are both well matched for covariates. One

Table 1: Covariate balance and degree of encouragement in two matched comparisons. Nine rare congenital anomalies were also balanced. (|st-diff| = absolute standardized difference. 1/0 means 1=yes, 0=no. Prenatal care month refers to month in which prenatal care began. 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|
Excess Travel Time to	Magnitude of Encouragement					
High Level NICU (minutes)	4.48	17.98	0.78	0.86	35.08	1.97
Covariates	Pregnancy and Birth					
Birth Weight (grams)	2,582	2,581	0.00	2,584	2,581	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

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 the babies and the other match uses about half the babies; other things being equal, that 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 versus about $\frac{3}{4}$ of a standard deviation. 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 aside from 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 zero, 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 in addition the difference in mean travel time is 23.4 minutes rather than 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% far away, and there was also a half-standard deviation 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 covariates in Table 1, the two matched comparisons look similar. For instance, for such key variables as birth weight and gestational age, the two matched com-

parisons are similar. There are 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.

3 Inference About Effect Ratios

3.1 Notation: treatment effects, treatment assignments

There are I matched pairs, $i = 1, \ldots, I$, with 2 subjects, $j = 1, 2$, one treated subject and one control, or $2I$ subjects in total. If the j^{th} subject in pair i receives the treatment, write $Z_{ij} = 1$, whereas 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 §1, the matched pairs consist of one mother close to a high level NICU (say control), the other further away (say treated). Notice that, in this terminology, proximity is the 'treatment,' although our real interest is in the effect of delivering at a low-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 upon which matched sample is under consideration.

The subscripts ij are book-keeping 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 may 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, $Z_{ij} = 1$, the other seen under control, $Z_{ij} = 0$; see Neyman (1923) and Rubin (1974). In §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}) and (d_{Tij}, d_{Cij}) where r_{Tij} and d_{Tij} are observed from j^{th} subject in pair i under treatment, $Z_{ij} = 1$, while r_{Cij} and d_{Cij} are observed from this subject under control, $Z_{ij} = 0$. In §1, (r_{Tij}, r_{Cij}) indicates infant death, 1 for dead, 0 for alive, and (d_{Tij}, d_{Cij}) indicates whether 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: (i) if mother had 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 (ii) if mother had 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 ($r_{Cij} = 0$).

The effects of the 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

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|
Excess Travel Time to	Magnitude of Encouragement					
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

every subject. Let $\mathcal{F} = \{(r_{Tij}, r_{Cij}, d_{Tij}, d_{Cij}, \mathbf{x}_{ij}, u_{ij}),\ i = 1, \ldots, I,\ j = 1, 2\}$. 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 NICU's did not deliver at high level NICU's; i.e., 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 $i = 1, \ldots, I,\ j = 1, 2$. In §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, it would be difficult to argue that regionalization of care is warranted.

In the current paper, 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 §1, the exclusion restriction says that mother and baby are affected by a high level NICU nearby only if proximity to a high level NICU changes the type of hospital in which mother delivers. As will be seen, 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" 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 she 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 $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{I,2})^T$, let Ω be the set containing the $|\Omega| = 2^I$ possible values $\mathbf{z}$ of $\mathbf{Z}$, so $\mathbf{z} \in \Omega$ if $\mathbf{z} = (z_{11}, z_{12}, \ldots, z_{I,2})^T$ with $z_{ij} = 0$ or $z_{ij} = 1$, $1 = z_{i1} + z_{i2}$ for $i = 1, \ldots, I$. Write $\mathcal{Z}$ for the event that $\mathbf{Z} \in \Omega$. In a randomized experiment, $\mathbf{Z}$ is picked at random from Ω, so $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ 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. Notice that under Fisher's sharp null hypothesis of no effect H_0 in §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 to

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 there was no high level NICU 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. The hypothesis will be tested with the aid of the statistic,

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

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 may write

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

Also, 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 say that the test works. The propositions are followed by several remarks that set these facts in either historical or practical contexts. The central result of §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 think that use of (10) yields a conservative test of the composite hypothesis $H_0^{(\lambda)} : \lambda = \lambda_0$, and the remarks are, in large part, 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 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* $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ *for each* $\mathbf{z} \in$

Ω, *the* $V_i(\lambda_0)$ *are mutually independent given* $\mathcal{F}$, $\mathcal{Z}$, *and*

$$\mathrm{E}\{V_i(\lambda_0) \mid \mathcal{F}, \mathcal{Z}\} = \frac{1}{2}\left(y_{Ti1,\lambda_0} - y_{Ci1,\lambda_0} + y_{Ti2,\lambda_0} - y_{Ci2,\lambda_0}\right) = \mu_{i,\lambda_0},\ say, \tag{4}$$

$$\mathrm{var}\{V_i(\lambda_0) \mid \mathcal{F}, \mathcal{Z}\} = \frac{1}{4}\left(y_{Ti1,\lambda_0} - y_{Ti2,\lambda_0} + y_{Ci1,\lambda_0} - y_{Ci2,\lambda_0}\right)^2 = \nu_{i,\lambda_0},\ say, \tag{5}$$

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

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

$$\mathrm{E}\left\{S^2(\lambda_0) \,\middle|\, \mathcal{F}, \mathcal{Z}\right\} - \mathrm{var}\{T(\lambda_0) \mid \mathcal{F}, \mathcal{Z}\} = \frac{1}{I(I-1)}\sum_{i=1}^{I}\left(\mu_{i,\lambda_0} - \overline{\mu}_{\lambda_0}\right)^2. \tag{8}$$

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

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

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

For large I, the hypothesis $H_0^{(\lambda)} : \lambda = \lambda_0$ will be 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 large 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 become increasingly unstable (as they would, for instance, if the r_{Tij} were sampled independently from a Cauchy distribution). Proposition 2 is substantially more general than anything needed for the current paper, 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 increase without bound over a set of values $I_1 < I_2 < I_3 \cdots$, not necessarily 1, 2, ..., 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* $\vartheta_{Ii} = E\left\{\left|V_i\left(\overline{\lambda}\right) - \mu_{i,\overline{\lambda}}\right|^3 \,\middle|\, \mathcal{F}_I,\, \mathcal{Z}_I\right\}$ *and* $\kappa_{Ii} = E\left[\left\{V_i\left(\overline{\lambda}\right)\right\}^4 \,\middle|\, \mathcal{F}_I,\, \mathcal{Z}_I\right]$, *assume that*

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

Then for each $k > 0$,

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

Proof. The proof depends upon two observations. (i) First observe that the right hand condition in (9) ensures that the weak law of large numbers (Serfling 1980, 1.8C, p. 27) applies to $I\,S_I^2\left(\overline{\lambda}\right)$ which, by (8), ensures that for all $\epsilon > 0$, $\delta > 0$, there exists an I^* such that for $I \ge I^*$, $\delta > \Pr\left[I\,S_I^2\left(\overline{\lambda}\right) - I\,\mathrm{var}\left\{T_I\left(\overline{\lambda}\right) \mid \mathcal{F}_I,\, \mathcal{Z}_I\right\} < -\epsilon\right]$; in words, in a sufficiently large experiment, it is nearly certain that $I\,S_I^2\left(\overline{\lambda}\right)$ does not much underestimate $I\,\mathrm{var}\left\{T_I\left(\overline{\lambda}\right) \mid \mathcal{F}_I,\, \mathcal{Z}_I\right\} = (1/I)\sum_{i=1}^{I}\nu_{i,\overline{\lambda}} = \varsigma_I$, say. (ii) By Proposition 1,

$0 = \mathrm{E}\left\{T_I\left(\overline{\lambda}\right) \mid \mathcal{F}, \mathcal{Z}\right\} = (1/I)\sum_{i=1}^{I}\mu_{i,\overline{\lambda}}$ for all I. From Proposition 1, the $V_i\left(\overline{\lambda}\right) - \mu_{i,\overline{\lambda}}$ are independent with expectation zero and variance $\nu_{i,\overline{\lambda}}$, so given $\mathcal{F}_I$, $\mathcal{Z}_I$, the quantity $\sqrt{I}T_I\left(\overline{\lambda}\right) = \left(1/\sqrt{I}\right)\sum_{i=1}^{I}\left\{V_i\left(\overline{\lambda}\right) - \mu_{i,\overline{\lambda}}\right\}$ has expectation 0 and variance $(1/I)\sum_{i=1}^{I}\nu_{i,\overline{\lambda}}$. Using a version of the central limit theorem (Theorem 9.2 in Breiman 1968, p. 186), the left condition in (9) implies $\sqrt{I}T_I\left(\overline{\lambda}\right)/\sqrt{\varsigma_I}$ converges in distribution to the standard Normal distribution as $I \to \infty$. Combining (i) and (ii) yields (10). ■

Remarks 3 and 4 consider an older and simpler situation than the main topic of the current paper, namely the situation in which $d_{Tij} - d_{Cij} = 1$ 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 3 and 4 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 which includes Fisher's hypothesis such that (10) holds as an equality when Fisher's hypothesis is true; hence, 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 3 *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 in Fisher (1935), Welch (1937) and Robinson (1973), among others.*

Remark 4 *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: (i) $\mu_{i,\lambda_0} = 0$ for all i, (ii) expression (8) equals zero and there is equality in (10), (iii) as $I \to \infty$, a test which rejects $H_0^c : r_{Tij} - r_{Cij} = \lambda_0$ for all ij when $T_I(\lambda_0)/S_I(\lambda_0) \geq k$ has size $\Phi(-k)$, and (iv) 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 5 is parallel to Remarks 3 and 4 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 5 *The model which 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*

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

and in this case λ in (1) equals λ^* and $\mu_{i,\lambda} = 0$, so with $\lambda_0 = \lambda^*$ expression (8) equals zero and there is equality in (10). So, as in Remark 4, 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 which rejects when $T_I(\lambda)/S_I(\lambda) \geq k$ 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 6 and 7 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 one were sampling an infinite population. Importantly, Proposition 2 shows that $T_I\left(\widetilde{\lambda}\right)/S_I\left(\widetilde{\lambda}\right)$ yields appropriate inferences without the fanciful notion that randomized experiments are performed on a random sample from a population. Also, Remark 7 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 6 *Imagine that $\mathcal{F}$ was obtained by sampling a superpopulation of matched pairs such that (i) distinct pairs are mutually independent, (ii) within pairs, subjects are exchangeable but perhaps not independent, (iii) the distribution of $(r_{Tij}, r_{Cij}, d_{Tij}, d_{Cij}, \mathbf{x}_{ij}, u_{ij})$ is the same for all ij, (iv) $(r_{Tij}, r_{Cij}, d_{Tij}, d_{Cij})$ have expectations and variances, (iv) $E(d_{Tij}) - E(d_{Cij}) > 0$; 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 $V_i\left(\widetilde{\lambda}\right)$ has expectation zero and constant variance $\sigma^2 = E\left\{V_i\left(\widetilde{\lambda}\right)^2\right\}$, so that $I \cdot S^2\left(\widetilde{\lambda}\right)$ converges in probability to σ^2. Also, unconditionally, the $V_i\left(\widetilde{\lambda}\right)$ are iid, so $T_I\left(\widetilde{\lambda}\right)/S_I\left(\widetilde{\lambda}\right)$ converges in distribution to $\Phi(\cdot)$. This is an alternative view of the approximation (10).*

Remark 7 *The most basic view of instrumental variables links them to a linear structural equation*

$$R_{ij} = \theta_i + \lambda^* D_{ij} + \varepsilon_{ij} \quad \textit{with} \quad \varepsilon_{ij} \underline{||} Z_{ij}, \tag{12}$$

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 accord 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 (11) holds, and once again,* λ *in (1) equals* λ^**,* $\mu_{i,\lambda} = 0$*, and expression (8) equals zero and there is equality in (10). In this case,* $T_I\left(\widetilde{\lambda}\right)/S_I\left(\widetilde{\lambda}\right)$ *is similar to the Anderson-Rubin (1949) statistic.*

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 NICU depending upon 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 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 confidence interval is shorter with a stronger instrument. (This is not the principal reason for preferring a stronger instrument; see §4.)

The point estimate, 0.0090, is substantial: it is almost half the infant mortality for mothers living far from a high level NICU. The lower endpoint of the 95% confidence interval 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

Table 3: Inference about the effect ratio λ under the assumption of random assignment of excess travel time within pairs matched for covariates. (CI = confidence interval)

	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	

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 §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 above one; also, 97,035 babies (49%) have negative predicted probabilities of death in the second stage. Conceivably, negative probabilities of death for half the babies do no harm in two-stage least squares, but they are at least disconcerting, perhaps worrisome. In contrast, in §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 95% interval [0.0050, 0.0116], with length 0.0067, so in comparison with the strong instrument in Table 3, the two-stage least squares yields an estimated effect that is about 8% smaller, 0.0083 versus 0.0090, with a confidence interval that is a tiny bit 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, $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ for each $\mathbf{z} \in \Omega$. In the next 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, inferences acted as if, within pairs matched for $\mathbf{x}_{ij}$, proximity to a high level NICU is random, in the sense that $\Pr(\mathbf{Z}=\mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ 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, prior to matching, mother ij had a probability $\pi_{ij} = \Pr(Z_{ij}=1 \mid \mathcal{F})$ 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, $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$, may differ in their odds of living far from a high level NICU by at most a factor of $\Gamma \geq 1$, so

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

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 $\mathbf{z} \in \Omega$,

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

where $\gamma = \log(\Gamma)$; see Rosenbaum (1995, §1.2; 2002, §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 $\Pr(\mathbf{Z}=\mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ in (14) is the randomization distribution. For fixed $\Gamma > 1$, the $\pi_{ij} = \Pr(Z_{ij}=1 \mid \mathcal{F})$ 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 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 §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 §3.2, if H_0 were true then the effect ratio λ is zero, $I \cdot T(0)$ equals $\sum_{i=1}^{I} \left\{ \sum_{j=1}^{2} Z_{ij} r_{Cij} - \sum_{j=1}^{2} (1 - Z_{ij}) r_{Cij} \right\}$, 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 §3.3 to test $H_0^{(\lambda)} : \lambda = 0$. If Fisher's H_0 were true, then standard methods of sensitivity analysis may be applied to $T(0)$; see Rosenbaum (1987; 1991; 2002, §4.4-5; 2007) and see Rosenbaum (1999) for a sensitivity analysis with an instrument.

For discussion of alternative methods of sensitivity analysis, see, for instance, Copas and Eguchi (2001), Gastwirth (1992), Imbens (2003), Lin, Psaty, and Kronmal (1998), Marcus (1997), Robins, Rotnitzky and Scharfstein (1999) and Small (2007).

4.2 Application to the study of regionalization of perinatal care

In the case of matched pairs with binary responses, as in §1, say that pair i is discordant if it contains exactly one death, $R_{i1} + R_{i2} = 1$, and let $I^* \leq I$ be the number of discordant pairs, and $\mathcal{D}$ be the set of the indices i of the I^* discordant pairs, so $|\mathcal{D}| = I^*$. 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 — that is, 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.

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	Alive, $R_{ij} = 0$	48070	554
NICU, $Z_{ij} = 1$	Dead, $R_{ij} = 1$	748	215

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 is 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	Alive, $R_{ij} = 0$	96044	1226
NICU, $Z_{ij} = 1$	Dead, $R_{ij} = 1$	1391	574

In testing Fisher's H_0 in matched pairs with binary responses, the distribution of $T\left(0\right)$ under (14) receives a nondegenerate contribution from matched pair i only if the pair is discordant, and in this case, $T\left(0\right)$ is effectively the same as McNemar's statistic; that is, under H_0, as $\mathbf{z}$ varies over Ω, the statistic $T\left(0\right)$ 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 $\Gamma/\left(1+\Gamma\right)$, the other with sample size I^* and probability of success $1/\left(1+\Gamma\right)$; see Rosenbaum (1987; 1991; 2002, §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 $\Gamma/\left(1+\Gamma\right) = 1.22/\left(1+1.22\right)$ of an event. As will become

clearer in Table 6, the two quoted values of Γ in Tables 4 and 5, namely $\Gamma = 1.07$ and $\Gamma = 1.22$, are to two decimals the values of Γ where 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 ($\Gamma = 1.07$ versus $\Gamma = 1.22$), despite the larger sample size, which is precisely the prediction of statistical theory (Small and Rosenbaum 2008).

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 $\Gamma \geq 1.07$. To put that in perspective using techniques not described in the current paper, 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 with $\Gamma = 1.25$, 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 with $\Gamma = 1.08$. See Gastwirth, Krieger and Rosenbaum (1998) and Rosenbaum and Silber (2009b) for detailed discussion of two correspondences between one parameter (Γ) and two parameter sensitivity analyses of the type just mentioned.

Tables 4 and 5 pay attention to which mother in a pair has a greater excess travel time to a high level NICU, but 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 account of 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 account of such magnitudes. McNemar's test focused 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 account of

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	Weaker	Stronger	Stronger
Measure	Mortality	Weighted	Mortality	Weighted
Γ	$n = 99235$	$n = 99235$	$n = 49587$	$n = 49587$
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

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, §4) for the details of the sensitivity analysis for Wilcoxon's test applied to pairs discordant for a binary outcome. Table 6 displays four sensitivity analyses, two with the stronger instrument, two with the weaker instrument, two using McNemar's test, two using a weighted test. In Table 6, the weighting is of some help to the weak instrument — it down-weights 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 the babies in Pennsylvania using observed covariates yields 99,174 pairs and a weak instrument. Pairing about half 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 which 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 quite well 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 slight more with an 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 it is less often urban. In terms of the shift in the population, when building a stronger instrument, the investigator should describe and discuss the shift, for instance with a table similar to Table 1.

In the second instance, if the instrument is stronger, mothers are more likely to comply, so the meaning of a 'complier' has changed. Importantly, we did not use their compliance behavior in building the matched sample; rather, we used excess travel time, whether or not travel time influenced where mother delivered. In the larger match, the average difference in travel time within pairs was less than 14 minutes, while 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 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 and a weak instrument. If one thought that the typical mother would comply for an extra 34 minutes but not for an extra 15 minutes, then the smaller match with a stronger instrument is 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. One could accept Pennsylvania as it is, accepting also a weak instrument, or one could search for another state or cross-state region whose geography made excess travel time into a stronger instrument. Instead of this, 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 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.

7 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.

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.

Avriel, M. (1976), *Nonlinear programming*, New Jersey: Prentice Hall.

Bound, J., Jaeger, D. A., 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.

Breiman, L. (1968) *Probability*. Reading, MA: Addison Wesley. Reprinted by SIAM.

Copas, J. and Eguchi, S. (2001), "Local sensitivity approximations for selectivity bias," *Journal of the Royal Statistical Society* B 63, 871-96.

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

Gadbury, G. L. (2001), "Randomization inference and the bias of standard errors," *American Statistician*, 55, 310-313.

Gastwirth, J. L. (1992), "Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables," *Jurimetrics* 33, 19-34.

Gastwirth, J. L. and Krieger, A. M. and Rosenbaum, P. R. (1998), "Dual and simultaneous sensitivity analysis for matched pairs," *Biometrika*, 85, 907–920.

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

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*, A, 168, 109-126.

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.

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.

Lu, B. and Rosenbaum, P. R. (2004), "Optimal matching with two control groups," *Journal of Computational and Graphical Statistics*, 13, 422-434.

Lu, B. (2005), "Propensity score matching with time-dependent covariates," *Biometrics*, 61, 721-728.

Lu, B., Greevy, R., Xu, X., and Beck, C. (2009), "Optimal nonbipartite matching and its statistical applications." *American Statistician*, to appear.

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.

Neyman, J. (1923, 1990), "On the application of probability theory to agricultural experiments," *Statistical Science*, 5, 463-480.

Neyman, J. (1935), "Statistical problems in agricultural experimentation," *Supplement to the Journal of the Royal Statistical Society*, 2, 107-180.

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

Robinson, J. (1973), "The large sample power of permutation tests for randomization models," Annals of Statistics, 1, 291-296.

Rogowski, J. A., Horbar, J. D., Staiger, D. O., Kenny, M., Carpenter, J., Geppert, J. (2004), "Indirect vs direct hospital quality indicators for very low-birth-weight infants," *Journal of the American Medical Association*, 291, 202-209.

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

Rosenbaum, P. R. (1991), "Sensitivity analysis for matched case-control studies," *Biometrics*, 47, 87-100.

Rosenbaum, P. R. (1999), "Using combined quantile averages in matched observational studies," *Applied Statistics*, **48**, 63-78.

Rosenbaum, P. R. (2002) *Observational Studies* (Second Edition). New York: Springer-Verlag.

Rosenbaum, P. R. (2004), "Design sensitivity in observational studies," *Biometrika*, 91, 153-164.

Rosenbaum, P. R. (2005a), "Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies," *American Statistician*, 59, 147-152.

Rosenbaum, P. R. (2005b), "An exact, distribution free test comparing two multivariate distributions based on adjacency," *Journal of the Royal Statistical Society*, B, 67, 515-530.

Rosenbaum, P. R. (2007), "Sensitivity analysis for m-estimates, tests and confidence intervals in matched observational studies," *Biometrics*, 63, 456-464.

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.

Rosenbaum, P. R. and Silber, J. H. (2009b), "Amplification of sensitivity analysis in observational studies," *Journal of the American Statistical Association*, 104, 1398–1405.

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

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

Rubin D. B. (1980), "Bias reduction using Mahalanobis metric matching," *Biometrics*, 36, 293-298.

Silber, J. H., Lorch, S. L., Rosenbaum, P. R., Medoff-Cooper, B., Bakewell-Sachs, S., Millman, A., Mi, L., Even-Shoshan, O., Escobar, G. E. (2009), "Additional maturity at discharge and subsequent health care costs," *Health Services Research*, 44, 444-463.

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

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.

Wald, A. (1940), "The fitting of straight lines if both variables are subject to error," *Annals of Mathematical Statistics*, 11, 284-300.

Wang, L. S. and Krieger, A. (2006), "Causal conclusions are most sensitive to unobserved binary covariates," *Statistics in Medicine*, 25, 2256-2271.

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

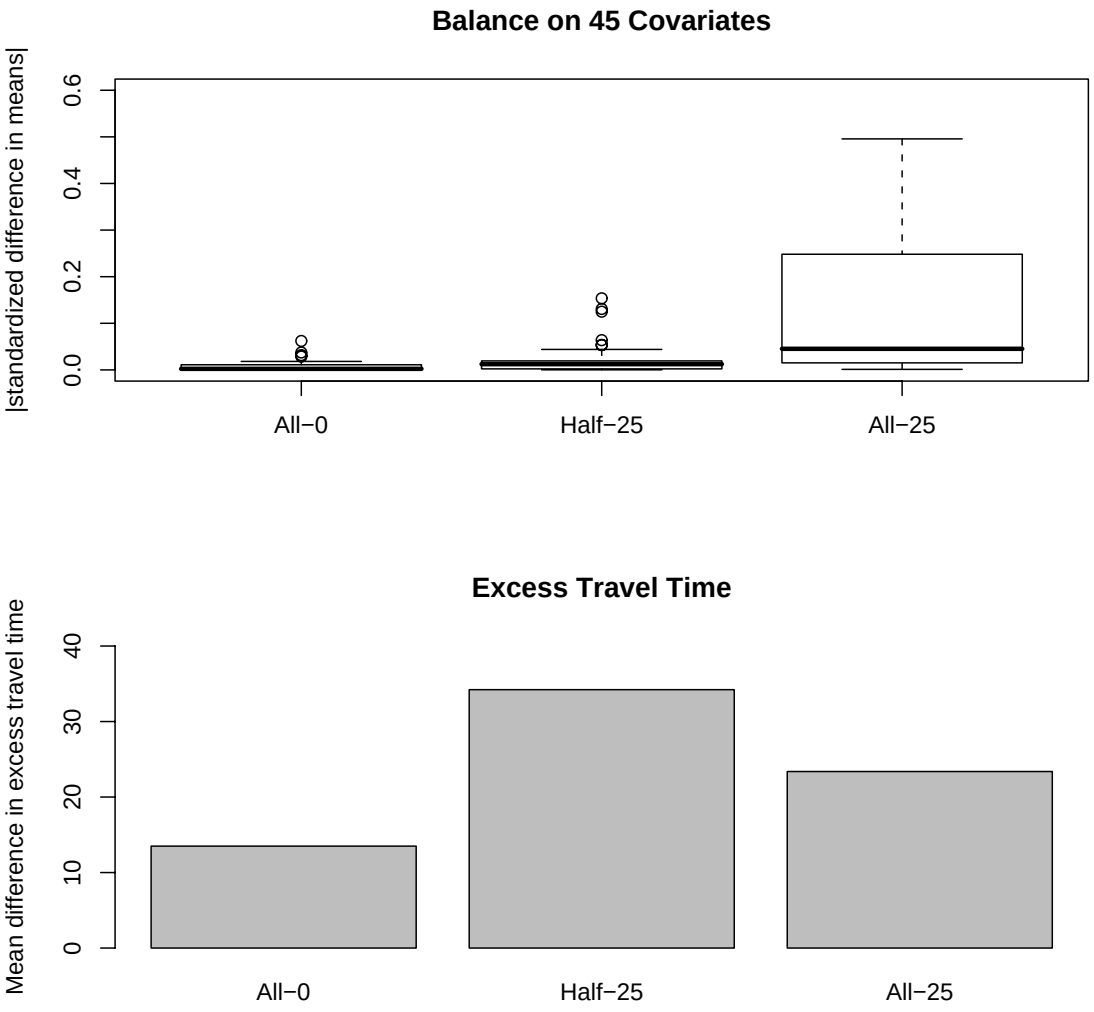

Figure 1: Comparison of three matched comparisons in terms of comparability on covariates and excess travel time. The match All-0 uses all of the babies, but insists only on a nonzero difference in excess travel time. The match Half-25 uses half the babies while trying to obtain at least a 25 minute 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.