Supplemental Information

TECHNICAL APPENDIX 1 Maternal Comorbid Conditions and Infant Congenital Anomalies Included in This Study

Comorbid Condition or Anomaly	ICD-9-CM Code
Disorders of placentation	641.0x, 641.1x, 641.2x
Chronic hypertension	642.0x, 642.1x, 642.2x
Cord abnormality	663.0x, 663.1x, 663.5x
Preterm labor	644.0x, 644.2x
Premature rupture of membranes	658.1x, 658.2x
Chorioamniotis	658.4x, 659.2x, 659.3x
Genitourinary tract infection	646.6x
Pregnancy-induced hypertension	642.4x, 642.5x, 642.7x
Oligohydramnios	658.0x
Blood transfusion	99.0, 99.00, 99.02, 99.03, 99.04
Amniocentesis	75.1
Cord prolapse	663.0x, 762.4
Diabetes mellitus	250.x 357.2 362.0x 366.41 648.0x
Gestational diabetes	648.8x
Renal disease	646.2x
Uterine rupture	665.0x 665.1x
OB shock hypotension	669.1x 669.2x
Thyroid disfunction	648.1x
Thrombosis	671.2x 671.3x 671.4x 671.5x
Congenital cardiac disease	648.5x
Eclampsia	642.6x
Placenta previa	641.0x 641.1x
Placenta abruption	641.2x
Hysterectomy	68.3 68.4 68.5x 68.6 68.7 68.8 68.9
Gastrointestinal malformation	756.70, 756.79, 750.3, 750.4, 750.5, 750.7, 750.8, 750.9, 751.1, 751.5, 751.8, 751.9, 560.2, 751.4, 751.0, 751.2, 751.3, 771.1, 751.61, 751.7, 751.60, 751.69
Genitourinary malformation	753.0, 753.12, 753.14, 753.15, 753.10, 753.19, 753.3, 753.4, 753.21, 753.22, 753.23, 753.29, 753.6, 753.7, 753.8, 753.9, 753.20, 756.71
Central nervous system malformation	741.0x, 741.9x, 742.0, 742.1, 742.2, 742.3, 742.4, 742.59, 742.8, 742.9
Pulmonary malformation	519.4, 553.3, 748.9, 750.6, 756.6, 748.3, 748.4, 748.60, 748.61, 748.69, 748.8
Cardiac malformation	746.3, 746.4, 424.1, 747.10, 747.21, 747.29, 747.11, 747.22, 746.81, 746.7, 425.3, 746.5, 424.0, 746.6, 746.84, 745.10, 745.19, 745.12, 746.85, 425.1, 745.3, 745.11, 745.0, 746.01, 746.83, 746.2, 746.09, 745.2, 746.1, 745.60, 745.61, 745.69, 746.82, 747.41, 747.42, 747.40, 747.49, 746.9, 746.89, 746.87
Skeletal malformation	756.50, 756.51, 756.55, 756.56, 756.59
Skin malformation	757.1
Chromosomal anomaly	758.3, 758.5, 758.89, 758.9, 759.89, 759.9 759.7 759.4
Other malformation	778.0, 759.6, 776.5

For each ICD-9-CM code, "x" represents any number at that specific digit location.

TECHNICAL APPENDIX 2

This technical appendix follows closely to the work done in Baiocchi et al (2010). This technical report serves as a foundation for intuition of the methods used in this article. Arguments are reproduced without proof, although care is taken to provide references. For a more complete discussion and development of these methods please consult Baiocchi et al (2010).

1 Matching to Create Stronger Instruments

1.1 Fewer Pairs at Greater Distances

We used optimal nonbipartite matching to pair infants with similar covariates but different excess travel times. There are $2I$ infants. First, a discrepancy is defined between every pair of infants, 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 $2I$ infants into I nonoverlapping pairs of 2 infants in such a way that the sum of the discrepancies within the I pairs is minimized. That is, 2 infants 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 et al (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).

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

The discrepancy matrix was built in several steps by using standard devices. Because we are matching mothers from different parts of the states, 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 mothers from poor communities with other mothers from poor communities. The 6 census/zip code measures are intended to represent local socioeconomic status, but socioeconomic status is not 6-dimensional. First, socioeconomic measures describing a zip code were summarized by using their first 2 principal components. These 2 components were combined with individual-level data about mother and infant in calculating a Mahalanobis discrepancy between every pair of infants. A small penalty (ie, a positive number) was added to the discrepancy for each of the following circumstances for any pair of infants 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 penalties to secure the desired balance. Finally, a substantial penalty was added to the discrepancy between any pair of infants 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 0 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 4 important covariates. The first covariate is state. Another was year of birth. The other 2 covariates that were exactly matched were coarse categorical versions of birth weight and gestational age. This means that we split 1 large matching problem into several smaller matching problems, grouping the pairs into 1 study at the end. In addition to ensuring exact matches on these 4 covariates, this process permits a rather large matching problem 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 infants and the algorithm has a worst case time bound on the order of the cube of the number of infants, 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 Inference About Effect Ratios

2.1 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 jth 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, the matched pairs consist of 1 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.

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 x_{ij}, but may have failed to control an unobserved covariate u_{ij}; that is, $x_{ij} = x_{ik}$ 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 look similar in terms of the observed variables that are reported in the tables showing the balance of the covariates, but the tables omit the specific covariate u_{ij}, which might bias the comparison. Write $\mathbf{u} = (u_{11}, u_{12}, \ldots u_{I2})$ for the $2I$-dimensional vector.

For any outcome, each subject has 2 potential responses, one seen under treatment, $Z_{ij} = 1$, the other seen under control, $Z_{ij} = 0$; see Neyman (1923) and Rubin (1974). In the current study of NICUs, speaking in this way of 2 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 2 responses, (r_{Tij}, r_{Cij}) and (d_{Tij}, d_{Cij}), where r_{Tij} and d_{Tij} are observed from jth 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 our study, (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 infant 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 infant 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 infant 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 every subject. Let $\mathcal{F} = \{(r_{Tij}, r_{Cij}, d_{Tij}, d_{Cij}, x_{ij}, u_{ij}), i = 1, \ldots, I, j = 1, 2\}$.

Fisher 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 our study, 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 null hypothesis were plausible, it would be difficult to argue that regionalization of care is warranted.

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 $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{I2})^T$, let Ω be the set containing the $|\Omega| = 2^I$ possible values $\mathbf{z}$ of $\mathbf{Z}$, so $\mathbf{z} \in \Omega$ if $\mathbf{z} = (z_{11}, z_{12}, \ldots, z_{I2})^T$ 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 $\Pr(\mathbf{Z} = \mathbf{z} | \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ for each $\mathbf{z} \in \Omega$.

2.2 Effect Ratios

The effect ratio, λ, is the parameter

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

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 the Fisher sharp null hypothesis of no effect H_0 in §2.1, $\lambda = 0$.

The effect ratio is the ratio of 2 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 2 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 100 mothers discouraged by distance from delivering at a hospital with a high-level NICU, there is 1 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 et al (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 $\mathcal{F}$.

2.3 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,

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

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

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

For large I, the hypothesis $H_0^{(\lambda)} : \lambda = \lambda_0$ will be tested by comparing $T(\lambda_0)/S(\lambda_0)$ with 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 (eg, Welch 1937).

2.4 Inference About Risk Ratios

The compliance class H_{ij} of subject ij describes ij's behavior under encouragement by the IV versus no encouragement. The possible compliance classes are H_{ij} = complier if $d_{Tij} = 1$, $d_{Cij} = 0$; H_{ij} = always taker if $d_{Tij} = 1$, $d_{Cij} = 1$; H_{ij} = never taker if $d_{Tij} = d_{Cij} = 0$; and H_{ij} = defier if $d_{Tij} = 0, d_{Cij} = 1$. We make the monotonicity assumption that there are no defiers as in Angrist et al (1996). Suppose pairs are sampled from a superpopulation and r_T, r_C are the potential outcomes of a randomly selected subject; R, D, Z are the observed response, treatment received and encouragement level of the randomly chosen subject; and H is the compliance class of the randomly chosen subject. Then,

$$E(r_T|H=\text{complier}) = E\left\{\frac{\frac{R\times D\times Z}{P(Z=1)} - \frac{R\times D\times(1-Z)}{P(Z=0)}}{\frac{D\times Z}{P(Z=1)} - \frac{D\times(1-Z)}{P(Z=0)}}\right\} \tag{4}$$

and

$$E(r_C|H=\text{complier}) = E\left\{\frac{\frac{R\times(1-D)\times(1-Z)}{P(Z=0)} - \frac{R\times(1-D)\times Z}{P(Z=1)}}{\frac{D\times Z}{P(Z=1)} - \frac{D\times(1-Z)}{P(Z=0)}}\right\} \tag{5}$$

The risk ratio for compliers is

$$RR_{\text{complier}} = \frac{E(r_T|H=\text{complier})}{E(r_C|H=\text{complier})}.$$

We estimate RR_{complier} by plugging the sample values of R, D, Z, $P(Z=1)$, $P(Z=0)$ into the right-hand sides of (4) and (5), and taking the ratio of $\hat{E}(r_T|H=\text{complier})$ to $\hat{E}(r_C|H=\text{complier})$. To find a confidence interval for RR_{complier}, we bootstrapped the pairs (ie, resampled the pairs with replacement) and found a bootstrap percentile confidence interval (Efron and Tibshirani, 1986).

3 REFERENCES

Anderson TW, Rubin H. Estimations of the parameters of a single equation in a complete system of stochastic equations. *Ann Math Stat*. 1949;20:46–63

Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables (with Discussion). *J Am Stat Assoc*. 1996;91:444–455

Avriel M. *Nonlinear Programming*. Englewood Cliffs, NJ: Prentice Hall; 1976

Baiocchi M, Small DS, Lorch S, Rosenbaum PR. Building a stronger instrument in an observational study of perinatal care for premature infants. *J Am Stat Assoc*. 2010;105(496)1496–1512

Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J Am Stat Assoc*. 1995;90:443–450

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

Copas J, Eguchi S. Local sensitivity approximations for selectivity bias. *J R Stati Soc B*. 2001;63:871–896

Derigs U. Solving nonbipartite matching problems by shortest path techniques. *Ann Operat Res*. 1988;13:225–261

Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall; 1986

Gadbury GL. Randomization inference and the bias of standard errors. *Am Stat*. 2001;55:310–313

Gastwirth JL. Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics*. 1992;33:19–34

Gastwirth JL, Krieger AM, Rosenbaum PR. Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika*. 1998;85:907–920

Imbens GW. Sensitivity to exogeneity assumptions in program evaluation. *Am Econ Rev*. 2003;93:126–132

Imbens G, Rosenbaum PR. Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *J R Stat Soc A*. 2005;168:109–126

Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*. 1998;54:948–963

Lu B. Propensity score matching with time-dependent covariates. *Biometrics*. 2005;61:721–728

Lu B, Greevy R, Xu X, Beck C. Optimal nonbipartite matching and its statistical applications. *Am Stat*. 2011;65(1):21–30

Lu B, Rosenbaum PR. Optimal matching with two control groups. *J Comput Graph Stat*. 2004;13:422–434

Lu B, Zanutto E, Hornik R, Rosenbaum PR. Matching with doses in an observational study of a media campaign against drug abuse. *J Am Stat Assoc*. 2001;96:1245–1253

Marcus SM. Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *J Educ Behav Stat*. 1997;22:193–201

Neyman J. On the application of probability theory to agricultural experiments. *Stat Sci*. 1923, 1990;5:463–480

Neyman J. Statistical problems in agricultural experimentation. *Suppl J R Stat Soc*. 1935;2:107–180

Robins JM, Rotnitzky A, Scharfstein D. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference. In: Halloran E, Berry D, eds. *Statistical Models in Epidemiology*. New York, NY: Springer; 1999:1–94.

Robinson J. The large sample power of permutation tests for randomization models. *Ann Stat*. 1973;1:291–296

Rogowski JA, Horbar JD, Staiger DO, Kenny M, Carpenter J, Geppert J. Indirect vs direct hospital quality indicators for very low-birth-weight infants. *JAMA*. 2004;291:202–209

Rosenbaum P R. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*. 1987;74:13–26

Rosenbaum PR. Sensitivity analysis for matched case-control studies. *Biometrics*. 1991;47:87–100

Rosenbaum PR. Using combined quantile averages in matched observational studies. *Appl Stat*. 1999;48:63–78

Rosenbaum PR. Observational Studies. 2nd ed. New York, NY: Springer-Verlag; 2002

Rosenbaum PR. Design sensitivity in observational studies. *Biometrika*. 2004;91:153–164

Rosenbaum PR. Heterogeneity and causality: unit heterogeneity and design sensitivity in observational studies. *Am Stat*. 2005a;59:147–152

Rosenbaum PR. An exact, distribution free test comparing two multivariate distributions based on adjacency. *J R Stat Soc B*. 2005b;67:515–530

Rosenbaum PR. Sensitivity analysis for m-estimates, tests and confidence intervals in matched observational studies. *Biometrics*. 2007;63:456–464

Rosenbaum PR. *Design of Observational Studies*. New York., NY: Springer; 2010

Rosenbaum PR, Silber JH. Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. *J Am Stat Assoc*. 2009a;104:501–511

Rosenbaum PR, Silber JH. Amplification of sensitivity analysis in observational studies. *J Am Stat Assoc*. 2009b;104:1398–1405

Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66:688–701

Rubin DB. Bias reduction using Mahalanobis metric matching. *Biometrics*. 1980;36:293–298

Silber JH, Lorch SL, Rosenbaum PR, et al. Additional maturity at discharge and subsequent health care costs. *Health Services Res*. 2009;44:444–463

Small D. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J Am Stat Assoc*. 2007;102:1049–1058

Small D, Rosenbaum PR. War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *J Am Stat Assoc*. 2008;103:924–933

Wald A. The fitting of straight lines if both variables are subject to error. *Ann Math Stat*. 1940;11:284–300

Wang LS, Krieger A. Causal conclusions are most sensitive to unobserved binary covariates. *Stat Med*. 2006;25:2256–2271

Welch BL. On the z-test in randomized blocks and Latin squares. *Biometrika*. 1937;29:21–52

RESULTS APPENDIX 1 Strength of Instrument in Pennsylvania, California, and Missouri, 1995–2005

	Quartiles for Instrument				
	1st	2nd	3rd	4th	Δ/SD[a]
Pennsylvania					
Differential travel time	(1.5)	3.3	13.5	38.6	2.26
Percent deliver at high-level NICU	79.8%	72.9%	45.6%	23.9%	1.12
Birth weight	2538	2550	2624	2630	0.13
Gestational age	35.0	35.0	35.2	35.3	0.11
Race					
White	55.5%	58.6%	82.4%	85.4%	0.65
Black	28.8%	26.5%	5.8%	4.8%	0.65
Asian	1.7%	1.5%	1.1%	0.5%	0.11
Other	4.2%	3.9%	3.3%	1.4%	0.16
Insurance status					
FFS	18.0%	17.0%	23.6%	25.4%	0.20
HMO	34.5%	37.8%	43.1%	30.3%	0.27
Public	35.5%	33.0%	22.0%	32.9%	0.29
Other	10.4%	10.6%	9.5%	8.9%	0.06
Uninsured	1.3%	1.1%	1.4%	1.8%	0.06
Singleton birth	83.7%	83.2%	81.6%	84.4%	0.07
SGA	17.4%	16.7%	14.9%	15.5%	0.07
Maternal comorbid conditions and complications of pregnancy					
Comorbid conditions					
Chronic HTN	1.80%	1.97%	1.48%	1.27%	0.05
Gestational diabetes	5.31%	5.12%	5.27%	5.11%	0.01
Diabetes mellitus	1.75%	1.79%	1.71%	1.97%	0.02
Renal disease	0.27%	0.30%	0.30%	0.30%	0.01
Congenital heart disease	0.12%	0.10%	0.14%	0.10%	0.01
Complications of pregnancy					
Preterm labor	46.43%	45.12%	43.53%	43.75%	0.06
PIH	10.51%	10.78%	10.09%	10.24%	0.02
PROM	19.12%	18.65%	17.81%	17.18%	0.05
Oligohydramnios	4.58%	4.08%	3.86%	3.10%	0.08
Disorders of placentation	5.66%	5.76%	5.92%	5.36%	0.02
California					
Differential travel time	(1.8)	1.2	4.2	26.3	1.49
Percent deliver at high-level NICU	79.60%	67.17%	63.72%	37.72%	0.86
Birth weight	2613	2622	2629	2633	0.02
Gestational age	35.4	35.4	35.4	35.4	0.02
Race					
White	61.11%	60.09%	62.69%	72.81%	0.27
Black	10.94%	8.27%	8.01%	4.77%	0.23
Asian	9.51%	11.05%	10.00%	7.16%	0.13
Other	16.82%	18.87%	17.70%	13.55%	0.14
Insurance status					
FFS	3.26%	3.30%	3.21%	5.52%	0.12
HMO	46.13%	43.33%	46.05%	45.52%	0.06
Public	46.50%	49.38%	46.80%	44.15%	0.10
Other	0.85%	0.67%	0.95%	1.42%	0.08
Uninsured	3.23%	3.28%	2.95%	3.36%	0.02
Singleton birth	89.45%	89.93%	89.84%	89.15%	0.03
SGA	10.53%	10.43%	10.37%	10.27%	0.01
Maternal comorbid conditions and complications of pregnancy					
Comorbid conditions					
Chronic HTN	1.04%	0.99%	0.97%	1.05%	0.01
Gestational diabetes	5.68%	5.66%	5.67%	5.52%	0.01
Diabetes mellitus	1.26%	1.30%	1.29%	1.09%	0.02
Renal disease	0.16%	0.16%	0.15%	0.18%	0.01
Congenital heart disease	0.04%	0.05%	0.05%	0.05%	0.01

RESULTS APPENDIX 1 Continued

	Quartiles for Instrument				
	1st	2nd	3rd	4th	Δ/SD[a]
Complications of pregnancy					
Preterm labor	27.50%	27.23%	26.64%	26.86%	0.02
PIH	6.73%	6.86%	6.83%	7.60%	0.03
PROM	10.10%	10.32%	9.96%	11.12%	0.04
Oligohydramnios	3.44%	3.40%	3.23%	2.99%	0.03
Disorders of placentation	4.12%	4.14%	4.08%	4.27%	0.01
Missouri					
Differential travel time	2.47	9.93	24.86	93.23	2.16
Percent deliver at high-level NICU	55.7%	22.6%	24.1%	10.1%	1.02
Birth weight	2803	2800	2820	2830	0.04
Gestational age	35.2	35.1	35.3	35.3	0.06
Race					
White	72.57%	62.83%	77.51%	91.87%	0.68
Black	22.73%	34.44%	20.74%	6.10%	0.70
Asian	3.37%	2.08%	1.07%	0.78%	0.19
Other	0.75%	0.47%	0.46%	1.11%	0.08
Insurance status					
FFS	30.18%	28.37%	29.79%	23.78%	0.14
HMO	25.72%	25.75%	26.34%	15.48%	0.26
Public	37.27%	34.50%	36.19%	53.11%	0.38
Other	3.26%	8.74%	5.10%	4.70%	0.24
Uninsured	3.15%	2.46%	2.42%	2.88%	0.05
Singleton birth	88.19%	89.48%	89.66%	90.71%	0.08
SGA	12.10%	11.96%	11.55%	11.65%	0.02
Maternal comorbid conditions and complications of pregnancy					
Comorbid conditions					
Chronic HTN	1.39%	1.54%	1.50%	1.13%	0.04
Gestational diabetes	4.71%	4.79%	4.23%	3.68%	0.05
Diabetes mellitus	1.21%	1.35%	1.37%	1.21%	0.01
Renal disease	0.26%	0.18%	0.28%	0.21%	0.02
Congenital heart disease	0.09%	0.04%	0.04%	0.05%	0.02
Complications of pregnancy					
Preterm labor	31.54%	28.69%	29.46%	28.71%	0.06
PIH	8.45%	9.28%	9.13%	8.54%	0.03
PROM	13.04%	12.24%	12.11%	11.14%	0.06
Oligohydramnios	4.85%	3.88%	3.87%	3.67%	0.06
Disorders of placentation	4.71%	4.52%	4.39%	4.02%	0.03

FFS, fee for service; HMO, health maintenance organization; SGA, small for gestational age; HTN, hypertension; PIH, pregnancy-induced hypertension; PROM, premature rupture of membranes

[a] Δ/SD is the standardized difference between the high-level NICU and other delivery hospital groups for a specific variable, defined as (difference in means between 2 groups of patients) ÷ (SD of entire cohort). A value <0.20 is considered adequate balance between groups.

RESULTS APPENDIX 2 Adjusted Rates of Complications at High-level NICUs and Other Delivery Hospitals by Using Matched-Paired Propensity Score Analyses, Pennsylvania, California, and Missouri, 1995–2005

	Pennsylvania		California		Missouri	
	RD[a]	RR[b]	RD[a]	RR[b]	RD[a]	RR[b]
In-hospital death	−0.6 (−1.6, 0.4)	0.93 (0.89–1.04)	−2.1 (−2.6 to −1.6)[c]	0.86 (0.81–0.88)[c]	−2.7 (−4.7 to −0.6)[c]	0.90 (0.83–0.98)[c]
Neonatal death	−0.5 (−1.5, 0.5)	0.95 (0.85–1.05)	−0.4 (−0.8 to 0)	0.96 (0.93–1.01)	0.2 (−1.7 to 2.1)	1.01 (0.92–1.10)
Preventable fetal death	−0.3 (−0.7, 0.1)	0.78 (0.58–1.02)	−1.7 (−1.9 to −1.4)[c]	0.61 (0.57–0.66)[c]	−2.9 (−3.9 to −1.9)[c]	0.50 (0.40–0.64)[c]
BPD	3.8 (2.9–4.7)[c]	1.54 (1.41–1.69)[c]	1 (0.7–1.3)[c]	1.18 (1.13–1.25)[c]	−5.4 (−7.3 to −3.5)[c]	0.79 (0.71–0.86)[c]
NEC	0.9 (0.2–1.6)[c]	1.26 (1.08–1.51)[c]	0.8 (0.6–1.1)[c]	1.36 (1.25–1.47)[c]	3.5 (2.2–4.7)[c]	1.62 (1.35–1.88)[c]
Fungal sepsis	3.9 (3.1–4.8)[c]	1.93 (1.66–2.22)[c]	0.9 (0.6–1.2)[c]	1.3 (1.22–1.41)[c]	2.3 (0.7–4.0)[c]	1.18 (1.04–1.32)[c]
Bacterial sepsis	2.4 (0.6–4.2)[c]	1.08 (1.03–1.15)[c]	3.4 (2.7–4)[c]	1.16 (1.14–1.20)[c]	34.1 (31.0–37.3)[c]	1.89 (1.78–2.00)[c]
ROP	1 (0.3–1.7)[c]	1.27 (1.07–1.51)[c]	3.4 (3.1–3.8)[c]	1.73 (1.64–1.83)[c]	10.8 (8.7–12.9)[c]	1.50 (1.37–1.62)[c]
Surgery for ROP	0.2 (−0.1 to 0.5)	1.42 (0.85–2.26)	0.7 (0.5–0.8)[c]	2.05 (1.80–2.39)[c]	0.1 (−0.8 to 1.0)	1.03 (0.82–1.26)
Laparotomy	0.3 (−0.3 to 0.8)	1.12 (0.88–1.38)	0.1 (−0.3 to 0.1)	0.94 (0.84–1.03)	2.3 (1.2–3.4)[c]	1.52 (1.29–1.82)[c]
Any IVH	3.2 (2.2–4.3)[c]	1.39 (1.27–1.52)[c]	2.5 (2.2–2.9)[c]	1.56 (1.49–1.64)[c]	3.1 (1.1–5.2)[c]	2.41 (1.50–3.36)[c]

[a] A positive RD indicates a higher rate at high-level NICUs compared with other delivery hospitals. A negative RD indicates a lower rate at high-level NICUs compared with other delivery hospitals.

[b] A RR >1 indicates a higher rate at high-level NICUs compared with other delivery hospitals. A RR <1 indicates a lower rate at high-level NICUs compared with other delivery hospitals.

[c] All results statistically significant at a $P < .05$ level.