Aporetic Conclusions When Testing the Validity of an Instrumental Variable

Fan Yang, José R. Zubizarreta, Dylan S. Small, Scott Lorch, Paul R. Rosenbaum[1]
University of Pennsylvania and Columbia University

Abstract: An instrument or instrumental variable is often used in an effort to avoid selection bias in inference about the effects of treatments when treatment choice is based on thoughtful deliberation. An instrument is a haphazard nudge to accept one treatment or another, where the push can affect outcomes only to the extent that it alters the treatment received. There are two key assumptions here: (i) the push is haphazard or essentially random once adjustments have been made for observed covariates, (ii) the push affects outcomes only by altering the treatment, the so-called "exclusion restriction." These assumptions are often said to be untestable; however, that is untrue if testable means checking the compatibility of assumptions with other things we think we know. A test of this sort may result in an aporia, that is, a collection of claims that are individually plausible but mutually inconsistent, without clear indication as to which claim is culpable for the inconsistency. We discuss this subject in the context of our on-going study of the effects of delivery by cesarean section on the survival of extremely premature infants of 23-24 weeks gestational age.

Keywords: Aporia; causal inference; instrumental variable; observational study.

1 Testing untestable assumptions in causal inference with instrumental variables

1.1 What is an instrument? What assumptions underlie their use?

An instrument is a haphazard nudge to accept a treatment where the nudge can affect the outcomes only to the extent that it alters the treatment received. The most basic example is Holland's (1988) randomized encouragement design, in which people are randomized to one of two groups, and members of one group are encouraged to adopt some health promoting behavior, say quit smoking, but the outcome, say an evaluation of lung tissue, might respond to a reduction in cigarettes consumed but not to encouragement to quit that leaves cigarette consumption unchanged. There are two key elements here. First, in the encouragement experiment, people are picked at random for encouragement — selection

[1] *Address for correspondence:* Department of Statistics, The Wharton School, University of Pennsylvania, Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 USA. E-mail: dsmall@wharton.upenn.edu. 18 September 2013.

does not just look haphazard, it is actually randomized — so the comparison of encouraged and unencouraged groups is equitable, not subject to biases of self-selection. Even in the randomized encouragement design, people who change their behavior, quit smoking, are a self-selected part of the encouraged and possibly unencouraged groups, so a comparison of quitters and others could be very biased: quitters may be more self-disciplined in all areas of their lives and may be more concerned with health promotion. The second element is that encouragement works, affects the outcome, only if it changes behavior, the so-called exclusion restriction. Stated informally in words, the instrumental variable (IV) estimate, the Wald estimate, attributes the entire difference in outcomes between the randomized encouraged and unencouraged groups to the greater change in behavior in the encouraged group, thereby avoiding biases of self-selection. If the encouraged group has a mean outcome that is one unit better than the mean in the unencouraged group, and if half of the encouraged quit while none of the unencouraged quit, then the Wald estimator claims the effects of quitting on those who quit when encouraged is two units, because encouragement only affected half of those who were encouraged. See Angrist, Imbens and Rubin (1996) for an equivalent formal statement.

So there are two key elements in the randomized encouragement design:

(i) encouragement is randomized,

(ii) encouragement affects only those individuals who change their behavior in response to encouragement, the exclusion restriction.

In the encouragement design, (i) is ensured by the use of randomization, and (ii) seems highly plausible because of what we think we know about the relationships that might exist between advice, behavior and lung tissue. Typical applications of the reasoning involving instruments are less compelling, because (i) is not ensured by actual random assignment, and (ii) is less firmly grounded in other things we think we know. In particular, (i) is typically rendered somewhat plausible by adjusting for visible differences in measured pretreatment covariates between encouraged and unencouraged groups, but of course this strategy may fail to control a covariate that was not measured. Typically, the encouraged and unencouraged groups are not formed by random assignment, but rather in a way that appears irrelevant and haphazard, but these appearances may deceive. Typically, the exclusion restriction seems plausible to anyone who cannot imagine a way encouragement could affect the outcome without altering the treatment, but this may simply reflect

inadequate imagining. So it is natural to want to test the assumptions that define an instrument.

Instruments are useful in two related senses. First, they are useful when one expects direct selection into the treated group to be severely biased. In the normal course of events, people who go to college are different from people who end their education with a high school degree, and they were different in high school — going to college costs money, requires tolerable academic performance in high school, requires ambition of a certain type. Therefore, one could not reasonably attribute the difference in earnings of college graduates and high school graduates to an effect of college education — the groups were not comparable before college. As an instrument in this case, Card (1995) used living in a town with a college, reasoning that people who live near a college may reduce expense by living at home, but merely growing up near a college if you do not go to college is unlikely to boost your income. Similarly, in the normal course of events, medical practice typically ensures that people who receive a particular medical treatment differ from those who do not, so one cannot estimate the effects of the treatment by comparing just any treated and untreated patients. Nonetheless, a widely used instrument exploits the fact that different hospitals may use that treatment more or less frequently while patients may select hospitals based on considerations such as proximity or practice affiliation; see, for instance, McClellan et al. (1994), Lalani et al. (2010), and Lorch et al. (2012). Some patients will either receive or be denied the treatment not for reasons unique to their own situation but simply because they live near a particular hospital.

Second, instruments are useful when there are many possible instruments that may fail in many unconnected ways, but direct comparisons always fail in the same way. There are many instruments — many possible nudges — that might make going to college a little easier or a little harder at the margin, whereas the biases that affect direct comparisons of earnings between high school and college graduates are always biased in much the same way; see Card (2001) for discussion of the varied instruments that have been used in this context. If direct comparisons always face the same biases, they may always yield the same biased answer, so repeatedly seeing similar answers in different studies does not build conviction that the answers are correct (Rosenbaum 2001). If plausible doubts surround each study with an instrument, but the doubts about the instrument in each study leave the other studies untouched, then seeing similar answers in many studies with unconnected instruments may build conviction that the estimates are actually estimating the treatment effect, not a bias; see Imbens and Rosenbaum (2004, §1).

1.2 Untestable assumptions?

The assumptions required for an instrument are often said to be untestable (e.g., Morgan and Winship 2007, p. 196). Whether this is true or not depends in part on what one means by untestable. Assumptions might be said to be untestable if they (A) are premises of a theorem that is the basis for an inference, (B) these premises are not self-evident or implied by other premises that are self-evident, (C) these premises cannot be tested against data from the observable distributions specifically mentioned in the statement of the theorem. This is an internally consistent way to use the word untestable, but it is a manner of speaking at considerable tension with typical scientific practice. Typically in science, each new claim to know something is checked for consistency with the other things we think we know. There is no reason to confine this checking for consistency to the short list of premises of a theorem. This checking may involve logical consistency, but more often the question is whether the new knowledge claim and old knowledge claims could plausibly be describing one and the same world, or whether something has to give.

At the risk of belaboring an example, consider asking: Is living near a college an instrument for going to college? This example is attractive because no specialized knowledge is needed to know many things about how kids end up in college. To be an instrument, (i) and (ii) in §1.1 must be approximately true in an appropriate sense. If one confined attention to three variables — namely college versus high school education, earnings and whether or not one grew up in a town with a college — then perhaps there is no way to test (i) and (ii), but why should anyone confine attention to these three variables? There is nothing unreasonable in checking the IV assumptions against other things we think we know. In a small college town, a college or university might be the largest private employer. (The University of Pennsylvania is the largest private employer in Philadelphia.) Some parents, perhaps more than a few, may live in a small town with a college because they wish to work at a college. The college may subsidize tuition for children of employees, and concern for financing a child's education, or perhaps simply an interest in education, may have been a consideration in selecting the college as an employer, and hence the town as a place to live. To the extent that parents choose to work in a college town so as to work at the college, growing up in a college town may fail to satisfy conditions (i) and (ii) in §1.1 for an instrument. This is all testable, but not with the three variables, and indeed it is fixable, but not with the three variables. For example, one can exclude from the study people whose parents worked at a college, or compare results in towns in which

the college is the major or one minor employer. In practical work with instruments, it is quite common to hear people announce that IV assumptions are untestable and then to see them do the sorts of checks that test IV assumptions.

Why are IV assumptions often said to be untestable when people often test them? We suspect there is a reason. A test of IV assumptions may lead neither to rejection of the assumptions nor to acceptance but rather to an aporia.

1.3 Aporia: mutually inconsistent but individually plausible claims

The Oxford American Dictionary defines the noun aporia as "an irresolvable internal contradiction ... in a text, argument or theory," with aporetic as the adjective. A collection of propositions, $\varpi_1, \ldots, \varpi_L$ is an aporia if each ϖ_ℓ is plausible on its own but they are jointly inconsistent, that is, $\varpi_1 \wedge \ldots \wedge \varpi_L$ is false or implausible; see Rescher (2009). A special case of aporia occurs in mathematical reasoning in a proof by contradiction, in which one proves $\sim \varpi_L$ by showing that $\varpi_1, \ldots, \varpi_{L-1}$ are certainly true and $\varpi_1, \ldots, \varpi_L$ is aporetic in yielding a contradiction. In contrast, in a typical aporia, in the general case, the identity of the culpable proposition or propositions is unknown. In Plato's early dialogues, Socrates would invalidate the views of his opponents by demonstrating that those views were aporetic.

To recognize that one's beliefs contain an aporia is an advance in understanding, albeit an uncomfortable one. From a false premise, one can logically deduce every conclusion, true or false (because, in elementary propositional logic, $A \Rightarrow B$ is true for all B if A is false). To believe $\varpi_1, \ldots, \varpi_L$ individually but fail to recognize them as aporetic is to risk logically deducing false propositions from beliefs one holds (because one believes $\varpi_1, \ldots, \varpi_L$, can deduce the false proposition $A = \varpi_1 \wedge \ldots \wedge \varpi_L$ from one's beliefs, and can deduce any B from A because A is false). To recognize that one's beliefs $\varpi_1, \ldots, \varpi_L$ are aporetic is to recognize that one harbors at least one false belief, to be motivated to identify that belief, and to be hesitant in deducing consequences from $\varpi_1, \ldots, \varpi_L$. To recognize an aporia is an advance in understanding, and it is certainly better than believing the component propositions without recognizing their aporetic status.

One can escape an aporia $\varpi_1, \ldots, \varpi_L$ by arbitrarily discarding propositions ϖ_ℓ until the remaining propositions are no longer inconsistent. In this process, there is nothing to ensure that one has discarded false propositions and retained true ones. Rather, one has narrowed the scope of one's beliefs to the point that one is committed to sufficiently

few beliefs that one is safe from accusations of inconsistency. For instance, one can avoid an aporia in testing the assumptions of IV by defining those assumptions so narrowly that they become untestable.

1.4 Outline: an IV study; a test of IV assumptions; two technical innovations

We are currently using an instrument in a study of the possible effects of delivery by cesarean section of extremely premature infants of 23-24 weeks gestational age. Some background is discussed in §2.1 and the IV analysis is presented in §2.2-§2.4. In §2.5, the IV assumptions are tested, resulting in an aporia that is discussed in detail. The two appendices present two technical innovations: a new simpler approach to strengthening an instrument in Appendix I, and a sensitivity analysis for an attributable effect closely related to the Wald estimator in Appendix II.

2 Does delivery by cesarean section improve survival of extremely premature neonates?

2.1 Background: Studies of cesarean section without an instrumental variable

We are currently engaged in a study of the possible effects of cesarean section on the survival of very premature babies of 23-24 gestational age. For reasons to be described shortly, we tried to find an instrument for delivery by cesarean section and to check its validity by contrast with other trusted information. Some terminology and background are needed.

The gestational age of a full-term baby is 39 weeks or 9 months. Babies born under 37 weeks gestation are considered premature, with infants born younger having more medical problems, requiring more intensive medical care to survive, and having a higher likelihood of long-term neurodevelopment and medical problems. This issue is most prominent for the infants at the limits of viability, that is, those infants born at 23 and 24 weeks gestation. Babies born between 23 and 24 weeks of gestational age are very premature and face high risks of death and life-long health problems even with special care. A fetus of 23 and 24 weeks of gestational age that is not born alive is defined as a fetal death, whereas an infant who dies after delivery is designated as a neonatal death. There are clinical indicators around a pregnancy at the limits of viability that give the physician information about the likelihood that an infant will survive first the delivery, and then the initial period of time

after delivery.

In clinical epidemiology, the phrase "confounding by indication" is often defined as the bias introduced when patients receive medical treatments based on pretreatment indications that the patient would benefit from the treatment. To the extent that such indications for treatment are incompletely recorded, thus incompletely controlled by adjustments for recorded pretreatment differences, they may lead to bias in elementary analyses that rely on adjustments for confounding factors using recorded pretreatment differences. At gestational age 23-24 weeks, delivery by cesarean section is likely to reflect clinical judgement about the clinical stability and likelihood of survival of the infant and the generally unrecorded preferences of the mother. Both of these factors are likely to be incompletely recorded in most large-scale population datasets.

A major use of instrumental variables in medicine is to break up or otherwise avoid confounding by indication, that is, to find some circumstances in which patients received a medical treatment for reasons other than that the patient was expected to benefit from treatment. In a randomized trial, patients receive treatments for no reason at all, the flip of a fair coin, and instruments are sought in observational studies to recover as best one can some aspects of the randomized situation.

Existing literature suggests that routine or optional use of cesarean delivery for babies of ≥ 30 weeks gestational age is not of benefit to the baby. For instance, Werner et al. (2013) concluded:

> In this preterm cohort, cesarean delivery was not protective against poor outcomes and in fact was associated with increased risk of respiratory distress and low Apgar score compared with vaginal delivery. (page 1195)

More than seventy percent of the preterm cohort in Werner et al. (2013) were ≥ 30 weeks gestational age, and more than half were ≥ 32 weeks, while less than 6% were less than 26 weeks. Werner et al. (2013) compared babies delivered by cesarean section and babies delivered vaginally adjusting for measured covariates using logit regression. For instance, women on Medicaid were more likely to deliver vaginally with an odds ratio of 1.43, while women with third party insurance (e.g., Blue Cross) were more likely to deliver by cesarean section with odds ratio 1.46, and additive adjustments on the logit scale were intended to correct for this. Using similar methods and focusing on premature babies of ≥ 32 weeks gestational age, Malloy (2009) reached similar findings.

In contrast, for very premature infants of 22-25 weeks gestational age, Malloy (2008) concluded: "Cesarean section does seem to provide survival advantages for the most immature infants..." (page 285). As in the other studies, the comparison was of babies delivered by one method or the other with adjustments for measured covariates by logit regression.

With varied emphasis, these studies note the problem of confounding by indication. They note that a direct comparison of babies delivered by cesarean section and babies delivered vaginally could be biased by aspects of the baby and the mother that led to the decision to deliver by one method rather than the other, and this is true even if logit regression is used to adjust for measured covariates. The decision to perform a cesarean section in one case but not in another may reflect indications that were evident to the physicians or mothers involved but not evident in measured covariates. This seems especially likely when a complex choice is made in a thoughtful, deliberate way. For a baby of gestational age 23-24 weeks, these considerations may include a medical judgement about the viability of the baby, and a mother's concern for a baby who may face severe life-long health problems. When studying a survival outcome, one is especially concerned about comparing groups of babies that may have been constructed with the viability of those babies in mind. One might prefer circumstances in which more or fewer babies were delivered by cesarean section for reasons that had nothing to do with the particular situation of the baby and mother.

The finding that cesarean sections did not benefit more mature preterm babies did not stir up much controversy, but the finding of benefit for very premature babies was more controversial and surprising. We set out to study this using an instrument for cesarean section among babies 23-24 weeks of gestational age.

2.2 An instrument: variation among hospitals in the use of cesarean section for older babies

As noted in §2.1, confounding by indication occurs when patients receive treatments for good reasons, for instance because a physician believes giving the treatment to this patient will benefit this patient. It turns out that the use of cesarean section varies substantially from one hospital to the next. A mother may deliver by cesarean section not because of anything unique to her but simply because she delivers at a hospital that makes more extensive use of cesarean section.

Our instrument is the predicted c-section rate among babies of 23-24 weeks gestational age at the hospital where the baby was delivered. The rate is predicted using logit regression with four predictors. Three predictors describe the hospital's use of c-sections for older babies, that is: (a) the rate among babies with gestational age 25-32 weeks, (b) the rate among babies with gestational age 33-36 weeks, (c) the rate among babies with gestational age 37+ weeks. The fourth predictor was (d) the malpractice insurance rate in the county in which the hospital was located. There is evidence that cesarean sections are more common in regions where the risk of malpractice litigation is greater; e.g., Dubay, Kaestner, and Waidmann (1999), Baicker, Buckles, and Chandra (2006) and Yang et al. (2009). The continuous instrument was the predicted probability from the logit regression. So the value of this instrument would have been constant within a hospital but for predictor (d) which varied from year to year, so the instrument was constant in a given hospital in a given year, and was describing the proclivity of the hospital to perform c-sections rather than anything about a particular baby or mother.

2.3 Matching to strengthen the instrument

Available pretreatment covariates described the mother (e.g., her age), her baby (e.g., birth weight), the mother's Census tract (e.g., median household income), and the hospital. Hospitals vary in their abilities to care for premature infants. In particular, neonatal intensive care units (NICUs) are graded into seven levels of care based on available technology to care for sicker newborn patients. We matched exactly for the level of the NICU; see Table 1. We also used logit regression to estimate a hospital's risk-adjusted rates of two complications, thrombosis and wound infection, and matched to balance these variables. This scores were estimated from older babies, $\geq$ 25 weeks gestational age, so the scores make no use of outcomes for the group under study, namely babies of 23-24 weeks gestational age. The literature has suggested these two factors, thrombosis and wound infection, as measures of the quality of care provided by the obstetrical hospital. In brief, the matching sought to compare similar mothers and babies from similar neighborhoods at similar hospitals.

Matched pairs were formed to be similar in terms of covariates and very different in terms of the instrument. Specifically, each of 1489 pairs contained two babies of 23-24 weeks gestational age, one at a hospital with a high frequency of use of c-sections for older babies, the other with a low frequency of use of c-sections for older babies. So

the high and low groups looked similar in measured covariates, but one group went to hospitals that often delivered by c-section for older babies and the other group went to hospitals that used c-sections sparingly. As seen in Tables 1-3 and Figure 1, the 1489 babies in the high group and the 1489 babies in the low group were similar in terms gestational weeks (23 or 24), birth weight, year of birth, mother's age, mother's education, mother's race/ethnicity, mother's health insurance, the technical level of the hospital's neonatal intensive care unit (NICU), pregnancy complications such as hypertension and oligohydramnios, number of prenatal care visits, parity, month that prenatal care started, various aspects of the mother's census tract. In Table 1, the three covariates were matched exactly. In Table 2, the five covariates had identical marginal distributions but were not exactly matched, a condition known as "fine balance." In Table 3, the difference in means for the covariates was never more than a tenth of a standard deviation, while the difference in the instrument was more than three standard deviations. This is depicted for three continuous covariates and the instrument in Figure 1.

The matching was done in a new but simple way described in the Appendix. Described informally, nonoverlapping high and low instrument groups were defined by cutting the instrument in three places, discarding the middle. High and low babies were then selectively matched to push the groups further apart on the instrument, balance the covariates, and produce close individual pairs. The match was the solution to a constrained optimization problem. The appendix describes several versions of the problem, including the one we solved, and the associated `R` software to implement each version.

2.4 Outcomes: c-section and mortality rates

The instrument is intended to manipulate one outcome, whether or not a baby is delivered by cesarean section, with possible effects on another outcome, mortality of the baby. As intended and expected, the instrument did manipulate the rate of cesarean sections; see Table 4. Table 4 counts pairs, not babies, in the manner that is commonly associated with McNemar's test; see Cox (1970). More than half the babies in both the high and low groups were delivered vaginally, but the 24.6% c-section rate in the low group was increased by more than half to 38.2% in the high group. When the two babies in a pair were delivered in different ways, the odds were $396/194 = 2.04$ to 1 that the high baby had the c-section.

Table 5 displays the main outcome, namely total in-hospital mortality. Table 5 is

examining the possible effects of delivering at a high c-section hospital rather than a low c-section hospital, not yet the effects of c-sections themselves. The point estimate of the odds ratio favoring survival at a hospital with a high c-section rate is $360/185 = 1.95$. In the high group, survival rate was 34.8% and in the low group it was 23.0%, or a difference of $360 - 185 = 175$ survivors. If one believed naively that the matching in Tables 1-3 and Figure 1 had reproduced a paired randomized experiment that assigned one baby in each pair at random to the high hospital and the other to the low hospital (i.e., if one believed (i) but perhaps not (ii) in §1.1), then, using the method in Rosenbaum (2002, §6), one would be 95% confident that $A \geq 132$ babies were caused to survive because of delivery at a high hospital. (This is a one-sided 95% confidence interval derived from the randomization distribution, but if one prefers a two-sided interval, then the one-sided 97.5% interval is $A \geq 124$ babies rather than 132. In a paired randomized experiment, A is an unobserved random variable; see Appendix II.) Moving away from the naive model for treatment assignment (i.e., moving away from (i) in §1.1), if an unobserved covariate doubled the odds of delivery at a high hospital and doubled the odds of survival, then the one-sided 95% confidence interval is $A \geq 66$ babies were caused to survive because of delivery at a high hospital. (More precisely, the 95% interval is $A \geq 66$ at $\Gamma = 1.25$ by the method in Rosenbaum (2002), and this amplifies to $(\Lambda, \Delta) = (2, 2)$ by the method in Rosenbaum and Silber (2009).) If an unobserved covariate doubled the odds of delivery at a high hospital and quadrupled the odds of survival, then the one-sided 95% confidence interval is $A \geq 23$ babies were caused to survive because of delivery at a high hospital (or technically, this the 95% interval at $\Gamma = 1.25$ which amplifies to $(\Lambda, \Delta) = (2, 4)$). The ostensible effects of delivering at a high rather than low c-section hospital are not sensitive to small departures from random assignment. So far, nothing has been said about the effects of c-sections, only about the effects of delivering at hospitals that do more of them.

In Table 4, the high c-section hospitals did $D = 396 - 194 = 202$ more c-sections than did the low c-section hospitals and 175 more babies survived. If the high-versus-low grouping were a valid instrument for delivery by c-section, then the Wald estimator would attribute the additional survivors at high c-section hospitals to the additional c-sections at those hospitals, that is, ignoring sampling variability, 175 additional survivors attributed to 202 additional c-sections. Assuming that the high-versus-low grouping is a valid instrument (that is, assuming both (i) and (ii) in §1.1), the Wald estimate of the effect of c-sections on the survival of babies who receive them because they were born at high c-section hospitals is $175/202 = 0.87$, an impressive ratio, not quite one more

survivor for one more c-section. There is substantial sampling variability and possible bias in assignment to high or low hospitals, and both must be addressed, the first using a confidence statement, the second using sensitivity analysis. An interesting quantity is A/D where A is the attributable effect in the previous paragraph and D is number of additional c-sections at high c-section hospitals. The 95% confidence intervals for A/D are $A/D \geq 132/202 = 0.65$ for randomization inference ($\Gamma = 1$), $A/D \geq 66/202 = 0.33$ for an unobserved covariate that doubled the odds of delivering at a high c-section hospital and doubled the odds of survival ($\Gamma = 1.25$), and $A/D \geq 23/202 = 0.11$ for an unobserved covariate that doubled the odds of delivering at a high c-section hospital and quadrupled the odds of survival ($\Gamma = 1.5$). (In Appendix II, it is noted that A/D is the ratio of an unobserved to an observed random variable and a confidence interval for it is discussed.)

The exclusion restriction would be false if high c-section hospitals were more aggressive in many ways in their efforts to save babies of 23-24 weeks gestational age and if some of the reduced mortality were due to other aspects of the care provided at high c-section hospitals. Is the exclusion restriction compatible with other things we think we know?

2.5 A test of the exclusion restriction

As discussed in §2.1, the literature claims that there is no benefit from cesarean section for older preterm babies, say 30-34 weeks gestational age. Presuming — that is, tentatively and uncritically assuming — that claim to be true, we tested the exclusion restriction by redoing the study for babies of 30-34 weeks gestational age. It is important to realize that the literature is based on direct comparisons of babies delivered by c-section and babies delivered vaginally, whereas we used an instrument, and there are other differences to be discussed in a moment. So we are really asking whether different methodologies concur in saying c-sections benefit babies at 23-24 weeks gestational age and not at 30-34 weeks gestational age, or whether an aporia has been produced, in which it is not reasonable to believe everyone's methodology, in the literature and our own, is producing correct conclusions about the effects of c-sections.

There were, of course, many more babies born at 30-34 weeks gestational age and the mortality rate was much lower. We matched in a manner similar to that in §2.3 and the Appendix, but because there were many more babies involved, we made more extensive use of exact matching. This produced 23631 pairs of babies of 30-34 weeks gestational age with covariate balance and instrument separation similar to that seen in Tables 1-3 and

Figure 1 for the younger babies.

As before for babies of 23-24 weeks gestational age, the instrument worked for babies of 30-34 weeks gestational age, with high babies more likely than low babies to be delivered by cesarean section. The mortality results appear in Table 6. After noting that the mortality rates are very different in Tables 5 and 6, one notes also that high babies had lower mortality rates than low babies in both tables, and the odds ratios are somewhat different in magnitude but neither is small, $360/185 = 1.95$ for 23-24 weeks and $1076/672 = 1.60$ for 30-34 weeks. We also looked for a trend, and indeed the odds ratio is larger at 30 weeks gestational age and smaller at 34 weeks. We redid the study again for babies of 25-29 weeks gestational age, finding mortality results between Tables 5 and 6.

So the claims in the literature and our results sound plausible and reasonable if taken one at a time, but they cannot all be correct inferences about the effects of cesarean section on mortality. The conclusion is an aporia, individually plausible claims that are mutually incompatible. Of course, many things could have gone wrong, either in the literature or in our study. In our study, the two assumptions required of an instrument might be false. The literature implicitly assumes that if one takes account of observed covariates, say by logit regression, then one has reproduced a randomized experiment (or formally, they implicitly assume ignorable treatment assignment), and that assumption gets people in no end of trouble in observational studies. Are there other possibilities?

Indeed, there is another possibility. The cited literature in §2.1 focused on neonatal deaths, excluding fetal deaths, whereas we looked at all deaths. If a woman was pregnant with a baby of 23-24 weeks gestational age and the pregnancy terminated at that time, then we did not distinguish a death moments before birth and a death moments after birth. Remember that a baby of 23-24 weeks gestational age will require substantial medical assistance to remain alive. To our minds, the death of a baby of 23-24 weeks gestational age is a biological event, whereas the classification of that death as before or after birth may be little more than bookkeeping, perhaps an attempt to reduce the emotional pain of an event that is typically distressing for the mother.

Because our findings differ from the literature, we separated fetal and neonatal deaths, as shown in Tables 7 and 8. Consider what Tables 7 and 8 would look like if one removed all pairs with at least one fetal death, that is, removed the first row and first column of each table. The remaining babies would be either alive or neonatal deaths, the outcomes studied in the existing literature. Indeed, the resulting tables would then agree with the existing literature, in that c-sections would look beneficial in Table 7 but not in Table 8.

By contrast, including fetal deaths, c-sections look beneficial in both tables. Arguably, a death of a fetus of 23-24 weeks gestational age is a death of an extremely premature baby, a biological event, whereas the classification of that death into a fetal death or a neonatal death is partly a style of practice and a manner of speaking. Arguably, fetal deaths should not be excluded from all deaths, as they were not excluded in Tables 5 and 6.

The available evidence is aporetic. Each part looks plausible on its own but the parts are mutually inconsistent. Something has to give, but it is less than clear what that something should be. The literature finds a benefit from c-sections at 23-24 weeks gestational age but not at 30-34 weeks gestational age. The literature makes no effort to address unmeasured biases in the selection of individual babies for delivery by cesarean section, though biases at the individual level are at least plausible, perhaps more plausible than not. In contrast, our analysis uses an instrument to avoid selection biases operating at the level of individual babies, using the frequency of c-sections among older babies at a hospital as an instrument for c-sections among babies of 23-23 weeks gestational age. Hospitals with higher frequencies of c-sections have somewhat lower mortality, and this difference is not sensitive to small biases of selection into high or low c-section hospitals. By virtue of assuming the exclusion restriction, the Wald estimator attributes higher survival to higher rates of c-sections, producing a point estimate of 87%, and that seems implausibly large — that is, 87% of c-sections save babies who would otherwise have died — however, confidence intervals include substantially smaller effects. The exclusion restriction could easily be false here if hospitals that do more c-sections also are more aggressive in other ways in their treatment of extremely premature infants — the exclusion restriction would wrongly attribute the effects of those other efforts to c-sections. Our results would look much more like the existing literature if we followed the literature in ignoring fetal deaths at 23-24 weeks gestational age, counting only neonatal deaths at 23-24 weeks gestational age, but we worry that in many cases the distinction between a fetal death and a neonatal death at 23-24 weeks gestational age is a distinction without much of a difference. The element that seems least ambiguous in all this is that hospitals that do more c-sections have lower total mortality at 23-24 weeks gestational age, a difference that is not easily attributed to small biases in selection of mothers into hospitals, although it could conceivably be explained by moderately large biases. Whether this difference is caused by c-sections or by something else these hospitals are doing is not as clear.

3 Summary

We have suggested that the assumptions of the instrumental variable argument are often testable providing an aporia is seen as an acceptable conclusion. An aporia is a collection of individually plausible but mutually incompatible propositions. An aporia is an advance in understanding, albeit an uncomfortable one. In the example, the result of testing the exclusion restriction is a heightened concern that the exclusion restriction may be false, and the IV analysis may be wrong, but also a heightened concern that some of the things we think we know from the literature, some of the things we assumed in testing the exclusion restriction, may themselves be false.

Appendix I: A new bipartite matching algorithm for strengthening an instrumental variable

Following Baiocchi et al. (2010) and Zubizarreta et al. (2013), we used matching to strengthen the instrumental variable while balancing observed covariates. However, we changed, simplified, and in some contexts improved, a key element. These two papers both took a single population, discarded part of the population, split the remainder into pairs, where the pairs balance covariates while being far apart on the instrument. Discarding a middle portion, an ambiguous portion, of the population makes the instrument stronger, improving its design sensitivity, making the study less sensitive to bias from nonrandom assignment of encouragement; see Small and Rosenbaum (2008). Traditionally, splitting a single population into pairs is called by the awkward name "nonbipartite matching" which means "not two parts." The history of the awkward name involves the fact that optimal two-part matching (e.g., treatment versus control matching), so called optimal bipartite matching, was studied and solved first; see Korte and Vygen (2008) for a textbook discussion of both problems with comprehensive references. Baiocchi et al. (2010) used an algorithm and Fortran code for optimal nonbipartite matching created by Derigs (1988), as implemented in Lu et al.'s (2011) R package `nbpmatching`; it minimizes the total distance within pairs formed from a single population, discarding a portion of the population using a technical trick called "sinks". Zubizarreta et al. (2013) used integer programming, specifically Zubizarreta's (2012) `mipmatch` package, to impose additional linear constraints on the nonbipartite match, such as requiring nominal covariates to be perfectly balanced or requiring means of continuous covariates to be close. Also, Zubizarreta et al. (2013)

changed the optimized objective function along the lines suggested in Rosenbaum (2012), so as to optimize the number of individuals discarded. A feature of the nonbipartite approach is that individual pairs are far apart on the instrument, but the high baby in one pair may be lower on the instrument than the low baby in some other pair. Depending upon the nature of the instrument and the covariates, that feature may or may not be reasonable. It might be reasonable if the meaning of the instrument changed with the levels of the covariate. In the current study, with an instrument defined in terms of a hospital's rate of use of c-sections in older babies, this feature did not seem reasonable.

We wanted each and every baby in the high group to have a higher value of the instrument than each and every baby in the low group. This change was implemented in a simple way using bipartite matching. We cut the population into three groups based on the value of the instrument, V, where the middle group, $0.29 \leq V \leq 0.31$ contained 10% of the population and was discarded. Write $\{\alpha_1, \ldots, \alpha_h, \ldots, \alpha_H\}$ for the H remaining babies in the high group and $\{\beta_1, \ldots, \beta_\ell, \ldots, \beta_L\}$ for the L remaining babies in the low group, noting that $V_{\alpha_h} > V_{\beta_\ell}$ for every h, ℓ. We then matched babies in the high group to babies in the low group to be close in terms of a covariate distance, $\delta_{h\ell}$, measuring how similar baby α_h and baby β_ℓ were in terms of covariates, and far apart on the instrument, with $\delta_{h\ell} = \infty$ if $V_{\alpha_h} - V_{\beta_\ell} < \omega$ for an $\omega > 0$. The covariate distance combined a robust Mahalanobis distance for covariates with $\delta_{h\ell} = \infty$ for mismatches on the variables in Table 1. Write $a_{h\ell} = 1$ if baby α_h in the high group is paired with baby β_ℓ in the low group, $a_{h\ell} = 0$ otherwise, so that we require $a_{h\ell} \in \{0, 1\}$, $\sum_{i=1}^{H} a_{i\ell} \leq 1$, $\sum_{j=1}^{L} a_{hj} \leq 1$, for each h, ℓ. In principle, one could simply minimize the total distance within matched pairs, $\sum_{h=1}^{H} \sum_{\ell=1}^{L} a_{h\ell}\, \delta_{h\ell}$, subject to $\sum_{h=1}^{H} \sum_{\ell=1}^{L} a_{h\ell} = \min(H, L)$, and this could be done using the optimal assignment algorithm — e.g., Bertsekas' (1981) auction algorithm as made available in the `pairmatch` function of Hansen's (2007) `optmatch` package in `R`. Alternatively, one could make ω larger, as we did, to further strengthen the instrument, discarding some babies to achieve this more stringent objective. This can be done using the same software for the assignment algorithm without constraining $\sum_{h=1}^{H} \sum_{\ell=1}^{L} a_{h\ell}$ and instead minimizing $\sum_{h=1}^{H} \sum_{\ell=1}^{L} a_{h\ell}\, \delta_{h\ell} - \lambda \sum_{h=1}^{H} \sum_{\ell=1}^{L} a_{h\ell}$ for specified $\lambda > 0$, and this determines an optimal number of babies to discard; see Rosenbaum (2012) for extensive specifics.

As in Zubizarreta (2012) and Zubizarreta et al. (2013), we used integer programming, not the optimal assignment algorithm, to minimize $\sum_{h=1}^{H} \sum_{\ell=1}^{L} a_{h\ell}\, \delta_{h\ell} - \lambda \sum_{h=1}^{H} \sum_{\ell=1}^{L} a_{h\ell}$ but with additional linear constraints. As in these references, these added constraints forced the fine balance in Table 2 and the close mean match seen in Table 3. Moreover,

we added a new constraint to further strengthen the instrument. Setting $\delta_{h\ell} = \infty$ if $V_{\alpha_h} - V_{\beta_\ell} < \omega$ forces each matched pair to differ by $\geq \omega$ in terms of the instrument. The new additional constraint forced the mean difference in the instrument V to differ by a larger number, $\Omega > \omega$, so every pair meets the minimum requirement of ω, but on average a larger difference of Ω is achieved. The new constraint was $\sum_{h=1}^{H} \sum_{\ell=1}^{L} a_{h\ell} \left(V_{\alpha_h} - V_{\beta_\ell} - \Omega \right) > 0$.

Appendix II: Confidence intervals and sensitivity analyses for A/D

Section 2.4 reported confidence intervals for ratios of survival effects to differences in the frequency of use of c-sections. These intervals are new but are a direct extension of an existing method. This appendix describes the new method and briefly indicates its justification. There are I matched pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one encouraged, $Z_{ij} = 1$, the other not, $Z_{ij} = 0$, so $Z_{i1} + Z_{i2} = 1$ for each i. In §2.3, there are $I = 1489$ pairs of two babies, one at a high c-section hospital, $Z_{ij} = 1$, the other at a low c-section hospital, $Z_{ij} = 0$. Pairs were matched for observed covariates $\mathbf{x}_{ij}$, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for each i, but the matching may have failed to control an unobserved covariate u_{ij}, so possibly $u_{i1} \neq u_{i2}$ for many or all i. Baby ij has two potential binary responses (r_{Tij}, r_{Cij}), one r_{Tij} if encouraged with $Z_{ij} = 1$, the other r_{Cij} if unencouraged with $Z_{ij} = 0$. In §2.4, $r_{Tij} = 1$ if baby ij would survive at the high c-section hospital in the i^{th} pair, $r_{Tij} = 0$ otherwise, and $r_{Cij} = 1$ if baby ij would survive at the low c-section hospital in the i^{th} pair, $r_{Cij} = 0$, otherwise. Fisher's (1935) sharp null hypothesis of no treatment effect asserts $H_0 : r_{Tij} = r_{Cij}$ for all babies ij — in words, switching from a low c-section hospital to a high c-section hospital does not change any baby's survival. In a randomized paired experiment with binary response, McNemar's test is the randomization test of Fisher's H_0. Each baby is observed under one treatment, so the effect of the treatment, $r_{Tij} - r_{Cij}$, is not observed for any baby; see Neyman (1923), Welch (1937) and Rubin (1974). In constructing one-sided tests and confidence intervals, we assume $r_{Tij} \geq r_{Cij}$; however, two sided inferences are straightforwardly obtained by combining two one-sided inferences in opposing directions. An important unobservable quantity in §2.4 is the attributable effect $A = \sum_{i=1}^{I} \sum_{j=1}^{2} Z_{ij} \left(r_{Tij} - r_{Cij} \right)$; it is the unobservable number of babies caused to survive by virtue of delivering at the high c-section hospital. For inference about A we follow Angrist, Imbens and Rubin (1996) in additionally assuming $r_{Tij} \geq r_{Cij}$, so a 23-24 week baby who would survive with the stress of a vaginal delivery, $r_{Cij} = 1$, would also survive with the reduced stress of a cesarean delivery, $r_{Tij} = 1$. Under Fisher's null

hypothesis of no effect, every $r_{Tij} - r_{Cij} = 0$, so $A = 0$ no matter how treatments Z_{ij} are assigned.

Similarly, (d_{Tij}, d_{Cij}) is the binary indicator of delivery by cesarean section or vaginal delivery (1 for c-section, 0 for vaginal delivery) at the high and low c-section hospital. Baby ij is said to be a complier if encouragement shifts the baby's delivery in the encouraged direction, that is, if $1 = d_{Tij} > d_{Cij} = 0$, so this baby would be delivered by c-section at the high c-section hospital in pair i and would be delivered vaginally at the low c-section hospital in pair i. Baby ij is said to be an always taker if $d_{Tij} = d_{Cij} = 1$, a never taker if $d_{Tij} = d_{Cij} = 0$, and a defier if $0 = d_{Tij} < d_{Cij} = 1$, and we follow the usual practice of assuming there are no defiers, $d_{Tij} \geq d_{Cij}$, so a baby who would be delivered by c-section at a low c-section hospital would also be delivered by c-section at a high c-section hospital; see Angrist, et al. (1996) for discussion of this terminology. Write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, d_{Tij}, d_{Cij}, \mathbf{x}_{ij}, u_{ij}), i = 1, \ldots, I, j = 1, 2\}$ and $\mathcal{Z}$ for the event that $Z_{i1} + Z_{i2} = 1$ for each i. In a randomized paired encouragement design, encouragement Z_{ij} is assigned by $\Pr(Z_{i1} = 1 \,|\, \mathcal{F}, \mathcal{Z}) = 1/2$, $Z_{i2} = 1 - Z_{i1}$, and assignments in distinct pairs are independent. A simple model for sensitivity analysis in observational studies has $1/(1+\Gamma) \leq \Pr(Z_{i1} = 1 \,|\, \mathcal{F}, \mathcal{Z}) \leq \Gamma/(1+\Gamma)$ for specified $\Gamma \geq 1$, $Z_{i2} = 1 - Z_{i1}$, with independent assignments in distinct pairs, so randomization inference corresponds with $\Gamma = 1$; see Rosenbaum (1987; 2002, §4) for discussion of this method of sensitivity analysis, and for other methods, see Cornfield et al. (1959), Rosenbaum and Rubin (1983), Gastwirth (1992), Marcus (1997), Small (2007), Yu and Gastwirth (2005), Hosman et al. (2010), and Schwartz et al. (2012). Write R_{ij} for the baby ij's observed survival response, $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\, r_{Cij}$, and D_{ij} for the observed delivery, $D_{ij} = Z_{ij}\, d_{Tij} + (1 - Z_{ij})\, d_{Cij}$. Table 9 renumbers the two babies in a pair so $Z_{i1} = 1$, $Z_{i2} = 0$, and then records the joint distribution of $(R_{i1}, D_{i1}, R_{i2}, D_{i2}) = (r_{Ti1}, d_{Ti1}, r_{Ci2}, d_{Ci2})$.

As $I \rightarrow \infty$ in a randomized encouragement design, for fixed α, $0 < \alpha < 1$, conventionally $\alpha = 0.05$, it is possible to find an observed random variable $\widetilde{A}$ such that $\Pr\left(A \geq \widetilde{A} \,\middle|\, \mathcal{F}, \mathcal{Z}\right)$ tends to a probability $\geq 1 - \alpha$, so that $A \geq \widetilde{A}$ holds with 95% confidence, that is, the unobserved attributable effect A is at least equal to $\widetilde{A}$ except in at most $100\alpha\%$ of experiments; see Rosenbaum (2002) for specifics and Weiss (1955) for general discussion of confidence sets for unobserved random variables in terms of observed random variables. Moreover, in a sensitivity analysis in an observational study, if the bias in treatment assignment is at most $\Gamma \geq 1$, then there is an observed random variable $\widetilde{A}_\Gamma$ such that $\Pr\left(A \geq \widetilde{A}_\Gamma \,\middle|\, \mathcal{F}, \mathcal{Z}\right)$

tends to a probability $\geq 1-\alpha$ as $I \to \infty$; again, see Rosenbaum (2002).

The exclusion restriction says that encouragement that does not change the delivery (d_{Tij}, d_{Cij}) does not change the response (r_{Tij}, r_{Cij}), that is, $r_{Tij} = r_{Cij}$ whenever $d_{Tij} = d_{Cij}$. Stated informally, the exclusion restriction says that if high c-section hospitals sometimes save the lives of babies, then they do it by performing c-sections not by doing something else. The exclusion restriction could easily be false: high c-section hospital could be more aggressive in many ways in trying to save the lives of babies of 23-24 weeks gestational age, and c-sections may produce only a part or even none of the survival effect of generally more aggressive treatment. The exclusion restriction places a series of constraints on the relationship between the observed Table 9 and the unobservable table recording $(r_{Ci1}, d_{Ci1}, r_{Ci2}, d_{Ci2})$. The unobserved table is called the pivot table. Consider, for example, the 44 pairs in the first row and first column of the observed Table 9. Because the exclusion restriction says $r_{Tij} = r_{Cij}$ whenever $d_{Tij} = d_{Cij}$, those 44 pairs could be in the same place in the pivot table or some could move to the third and fourth row of the first column, but none could move to the second row. In fact, the only differences that can exist between the observed and pivot tables are movements from the first row to the third or fourth row in the same column. Under the exclusion restriction, A is the total number of pairs that are in the first row of the observed table and in the fourth row of the pivot table.

Let $b_{ij} = 1$ if $r_{Tij} > r_{Cij}$ and $d_{Tij} > d_{Cij}$, and $b_i = 0$ otherwise. If $b_{ij} = 1$, then baby ij would survive receiving a c-section at the high c-section hospital in pair i and would die without a c-section at the low c-section hospital in pair i. Using the exclusion restriction, $r_{Tij} - b_{ij}\left(d_{Tij} - d_{Cij}\right) = r_{Cij}$, and the attributable effect is $A = \sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}\left(r_{Tij} - r_{Cij}\right) = \sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij} b_{ij}\left(d_{Tij} - d_{Cij}\right)$. The mean difference in survival is:

$$
\begin{aligned}
T_r &= \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}R_{ij} - \left(1 - Z_{ij}\right)R_{ij} = \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}\left\{r_{Cij} + b_{ij}\left(d_{Tij} - d_{Cij}\right)\right\} - \left(1 - Z_{ij}\right)r_{Cij} \\
&= \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2}\left(2Z_{ij} - 1\right)r_{Cij} + \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij} b_{ij}\left(d_{Tij} - d_{Cij}\right) \\
&= \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2}\left(2Z_{ij} - 1\right)r_{Cij} + \frac{A}{I}.
\end{aligned}
$$

In a randomized paired encouragement experiment, $\mathrm{E}\left(Z_{ij}=1 \,|\, \mathcal{F},\, \mathcal{Z}\right)=1/2$ so that

$$\mathrm{E}\left\{\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2}\left(2Z_{ij}-1\right)r_{Cij}\,\middle|\,\mathcal{F},\,\mathcal{Z}\right\}=0,\text{ and }\mathrm{E}\left(\frac{A}{I}\,\middle|\,\mathcal{F},\,\mathcal{Z}\right)=\frac{1}{2I}\sum_{i=1}^{I}\sum_{j=1}^{2}\left(r_{Tij}-r_{Cij}\right)=\tau,\text{ say,}$$

so that T_r and A/I are both unbiased for the average effect of encouragement, τ; however, departures from random assignment (i.e., failures of (i) in §1.1) can introduce bias. The observable random variable $T_d=\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2}Z_{ij}D_{ij}-\left(1-Z_{ij}\right)D_{ij}$ is the difference between the number of c-sections performed by the high and low c-section hospitals; in Table 4 it is $T_d=396-194=202$. It is a descriptive, not a causal quantity: it describes what happened, not what would happen. The Wald estimator is T_r/T_d. For the Wald estimate to work, encouragement must increase the frequency of what is encouraged so that T_d converges in probability to a strictly positive quantity $\delta>0$ as $I\to\infty$, and that is assumed here; therefore, with high probability, high c-section hospitals have done more c-sections among the I pairs than low c-section hospitals for sufficiently large I, and $\Pr\left(T_d\le 0\,|\,\mathcal{F},\,\mathcal{Z}\right)$ is negligible for large I. The quantity

$$W=\frac{\sum_{i=1}^{I}\sum_{j=1}^{2}Z_{ij}b_{ij}\left(d_{Tij}-d_{Cij}\right)}{\sum_{i=1}^{I}\sum_{j=1}^{2}Z_{ij}D_{ij}-\left(1-Z_{ij}\right)D_{ij}}=\frac{A}{\sum_{i=1}^{I}\sum_{j=1}^{2}Z_{ij}D_{ij}-\left(1-Z_{ij}\right)D_{ij}}=\frac{A/I}{T_d}$$

is the number of babies caused to survive by a c-section in a high c-section hospital as a fraction of the number of additional c-sections performed by high c-section hospitals. Now, W is the ratio of an unobservable random variable A/I, a causal quantity, and an observed random variable T_d, a descriptive quantity, so W is unobservable. The quantity W is directly interpretable on its own; however, it might reasonably be regarded as the intended finite sample estimand of the Wald estimator, in the sense that T_r/T_d and W both converge in probability as $I\to\infty$ to the average effect of c-sections on compliers if encouragement is randomized within pairs; see Angirst et al. (1996) for discussion of this estimand. Given the large sample confidence interval, $A\ge\widetilde{A}_\Gamma$ with $\Pr\left\{A\ge\widetilde{A}_\Gamma\,\middle|\,\mathcal{F},\,\mathcal{Z}\right\}\ge 1-\alpha$ for sufficiently large I, and continuing to regard $\Pr\left(T_d\le 0\,|\,\mathcal{F},\,\mathcal{Z}\right)$ is negligible for large I, we have $\Pr\left\{A/T_d\ge\widetilde{A}_\Gamma/T_d\,\middle|\,\mathcal{F},\,\mathcal{Z}\right\}=\Pr\left\{W\ge\widetilde{A}_\Gamma/T_d\,\middle|\,\mathcal{F},\,\mathcal{Z}\right\}\ge 1-\alpha$ for sufficiently large I. The confidence interval $W\ge\widetilde{A}_\Gamma/T_d$ was reported in §2.4.

References

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.

Baiocchi, M., Small, D. S., Lorch, S. and 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, 1285-1296.

Bertsekas, D. P. (1981), "A new algorithm for the assignment problem," *Mathematical Programming*, 21, 152-171.

Baicker, K., Subkles, K. S., and Chandra, A. (2006), "Geographic variation in the appropriate use of cesarean delivery," *Health Affairs*, 25, w355-w367.

Card, D. (1995), "Using geographic variation in college proximity to estimate the return to schooling," Chapter 7 in *Aspects of Labour Market Behavior* (eds L. N. Christofides, E. K. Grant and R. Swidinsky), Toronto: University of Toronto Press.

Card, D. (2001), "The causal effect of education," in *Handbook of Labor Economics* (eds O. Ashenfelter and D. Card), New York: North-Holland.

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

Cox, D. R. (1970), *Analysis of Binary Data*, London: Metheun.

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

Dubay, L., Kaestner, R., and Waidmann, T. (1999), "The impact of malpractice fears on cesarean section rates," *Journal of Health Economics*, 18, 491-522.

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

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

Hansen, B. B. (2007), "Optmatch," *R News*, 7, 18-24. `R` package `optmatch`.

Holland, P. W. H. (1988), "Causal inference, path analysis, and recursive structural equations models," *Sociological Methodology*, 18, 449-484.

Hosman, C. A., Hansen, B. B., and Holland, P. W. H. (2010), "The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder," *Annals of Applied Statistics*, 4, 849-870.

Imbens, G. W. and Rosenbaum, P. R. (2004), "Robust, accurate confidence intervals with

a weak instrument," *Journal of the Royal Statistical Society*, A 168, 109-126.

Korte, B. and Vygen, J. (2008), *Combinatorial Optimization: Theory and Algorithms*, New York: Springer.

Lalani, T., Cabell, C. H., Benjamin, D. K. et al. (2010), "Analysis of the impact of early surgery on in-hospital mortality of native valve dndocarditis: use of propensity score and instrumental variable methods to adjust for treatment-selection bias," *Circulation*, 121, 1005-1013

Lorch, S., Baiocchi, M., Ahlberg, C. and Small, D. (2012), "The differential impact of delivery hospital on the outcomes of premature infants," *Pediatrics*, 130, 270-278.

Lu, B., Greevy, R., Xu, X., and Beck, C. (2011), "Optimal nonbipartite matching and its statistical applications," *American Statistician*, 65, 21-30. `R` package `nbpmatching`.

Malloy, M. H. (2008), "Impact of cesarean section on neonatal mortality rates among very preterm infants in the United States, 2000-2003," *Pediatrics*, 122, 285-292.

Malloy, M. H. (2009), "Impact of cesarean section on intermediate and late preterm births: United States 2000-2003," *Birth*, 36, 26-33

Marcus, S. M. (1997), "Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect," *Journal of Educational Statistics*, 22, 193-201.

McClellan, M., McNeil, B. J., Newhouse, J. P. (1994), "Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables," *Journal of the American Medical Association*, 272, 859–66.

Morgan, S. L. and Winship, C. (2007), *Counterfactuals and Causal Inference*, New York: Cambridge University Press.

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

Rescher, N. (2009), *Aporetics: Rational Deliberation in the Face of Inconsistency*, Pittsburgh: University of Pittsburgh Press.

Rosenbaum, P. and Rubin, D. (1983), "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome," *Journal of the Royal Statistical Society* B, 45, 212-218.

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

Rosenbaum, P. R. (2001), "Replicated effects and biases," *American Statistician*, 55, 223-227.

Rosenbaum, P. R. (2002), "Attributing effects to treatment in matched observational studies," *Journal of the American Statistical Association*, 97, 183-192.

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

Rosenbaum, P. R. (2012), "Optimal matching of an optimally chosen subset in observational studies," *Journal of Computational and Graphical Statistics*, 21, 57-71.

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

Schwartz, S., Li, F. and Reiter, J. (2012), "Sensitivity analysis for unmeasured confounding in principal stratification settings with binary variables," *Statistics in Medicine*, 31, 949-962.

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

Small, D. S. 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.

Weiss, L. (1955), "A note on confidence sets for random variables," *Annals of Mathematical Statistics*, 26, 142-144.

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

Werner, E. F., Han, C. S., Savitz, D. A., Goldshore, M., Lipkind, H. S. (2013), "Health outcomes for vaginal compared with Cesarean delivery of appropriately grown preterm neonates," *Obstetrics and Gynecology*, 121, 1195-1200.

Yang, Y. T., Mello, M. M., Subramanian, S. V. and Studdert, D. M. (2009), "Relationship between malpractice litigation pressure and rates of cesarean section and vaginal birth after cesarean section," *Medical Care*, 47, 234-242.

Yu, B. B. and Gastwirth, J. L. (2005), "Sensitivity analysis for trend tests: application to the risk of radiation exposure," *Biostatistics*, 6, 201-209.

Zubizarreta, J. R. (2012), "Using mixed integer programming for matching in an observational study of kidney failure after surgery," *Journal of the American Statistical Association*, 107, 1360-1371. (R software `mipmatch` at http://www-stat.wharton.upenn.edu/~josezubi/)

Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S., and Rosenbaum, P. R. (2013), "Stronger instruments via integer programming in an observational study of late preterm birth outcomes," *Annals of Applied Statistics*, 7, 25-50.

Table 1: Three variables were exactly matched in forming 1489 pairs of two babies with gestational ages 23-24 weeks, namely gestational age (23 or 24 weeks), the capability or level of the neonatal intensive care unit (NICU), and the year of birth (1993-2005). The table gives counts of babies, and these are identical in the high and low instrument group defined by the estimated probability of a c-section at a given hospital.

	Instrument Group	
	High	Low
Gestational age in weeks		
23 weeks	726	726
24 weeks	763	763
NICU Level		
1	333	333
2	56	56
3A	126	126
3B	480	480
3C	438	438
3D	15	15
FC	41	41
Year of birth		
1993	30	30
1994	47	47
1995	90	90
1996	89	89
1997	104	104
1998	124	124
1999	133	133
2000	132	132
2001	129	129
2002	166	166
2003	188	188
2004	132	132
2005	125	125

Table 2: Five variables were finely balanced in forming 1489 pairs of two babies with gestational ages 23-24 weeks, meaning that these variables had the same marginal distributions in the high and low instrument groups, so the counts are identical. The table gives counts of babies, and these are identical in the high and low instrument group defined by the estimated probability of a c-section at a given hospital.

	Instrument Group	
	High	Low
Mother had hypertension during pregnancy		
Yes	75	75
No	1437	1437
Oligohydramnios		
Yes	52	52
No	1308	1308
Mother's race/ethnicity		
Non-Hispanic White	551	551
Non-Hispanic Black	305	305
Hispanic	478	478
Non-Hispanic Asian/P. Islander	87	87
Other	36	36
Missing	32	32
Mother's education		
8th grade or less	128	128
Some high school	249	249
High school graduate	473	473
Some college	303	303
College graduate	164	164
More than college (MS, PhD)	108	108
Missing	64	64
Mother's health insurance		
Fee for service	116	116
HMO	647	647
Federal/State	662	662
Other	20	20
Uninsured	42	42
Missing	2	2

Table 3: Covariates balanced in mean only and forced imbalance in mean in the instrument in forming 1489 pairs of two babies with gestational ages 23-24 weeks. The table gives the mean of each covariate or instrument before and after matching, together with the difference in means divided by the standard deviation before matching (S-Dif). For Yes/No = Y/N variables, 1=Yes, 0=No. RAHR = risk adjusted hospital rate.

	Before matching			After matching		
	Mean		S-Dif	Mean		S-Dif
	High	Low		High	Low	
Birth weight (grams)	591.12	577.25	0.16	587.08	580.31	0.08
Hypertension (Y/N)	0.07	0.04	0.12	0.05	0.05	0.00
Chorioamnionitis (Y/N)	0.28	0.26	0.04	0.27	0.26	0.02
RAHR of thrombosis	0.00	0.00	0.31	0.00	0.00	0.09
RAHR of wound infection	0.00	0.00	0.18	0.00	0.00	-0.05
Mother's age (years)	28.15	26.89	0.19	27.69	27.61	0.01
Prenatal care visits (#)	7.04	5.89	0.27	6.52	6.32	0.05
Prenatal care missing (Y/N)	0.09	0.05	0.14	0.07	0.05	0.07
Parity	1.90	1.90	-0.00	1.91	1.77	0.09
Parity missing (Y/N)	0.01	0.01	0.02	0.01	0.01	0.03
Month prenatal care started	2.00	2.16	-0.14	2.00	2.12	-0.10
Month care started missing (Y/N)	0.08	0.04	0.20	0.06	0.04	0.09
Multiple delivery	1.27	1.19	0.15	1.22	1.18	0.09
Congenital (Y/N)	0.15	0.14	0.04	0.15	0.14	0.03
Placentation (Y/N)	0.23	0.20	0.07	0.22	0.20	0.04
Diabetes (Y/N)	0.03	0.03	0.00	0.03	0.03	-0.03
Pre-term labor (Y/N)	0.81	0.74	0.17	0.80	0.76	0.09
PROMM (Y/N)	0.35	0.28	0.15	0.33	0.30	0.08
Household median income ($)	45024	41435	0.21	44730	44848	-0.01
Income missing (Y/N)	0.00	0.00	0.03	0.00	0.00	0.00
Poor (????)	0.11	0.16	-0.10	0.15	0.16	-0.03
Hospital delivery volume (#)	2850	2903	-0.03	2568	2722	-0.09
Small for gestation age (Y/N)	0.09	0.12	-0.11	0.09	0.12	-0.09
C-sec. predicted prob.	0.38	0.23	2.56	0.40	0.22	3.12

Table 4: C-sections in 1489 matched pairs of babies of 23-24 weeks gestational age. The table counts pairs, not babies. As expected, c-section rates are higher in the high c-section group.

	Low Baby			
High Baby	C-section	Other	Total	High Baby Rate
C-section	173	396	569	38.2%
Other	194	726	920	61.8%
Total	367	1122	1489	
Low Baby Rate	24.6%	75.4%		100.0%

Table 5: Mortality in 1489 matched pairs of babies of 23-24 weeks gestational age. The table counts pairs, not babies. Mortality rates are higher in the low c-section group.

	Low Baby			
High Baby	Dead	Alive	Total	High Baby Rate
Dead	786	185	971	65.2%
Alive	360	158	518	34.8%
Total	1146	343	1489	
Low Baby Rate	77.0%	23.0%		100.0%

Table 6: Mortality in 23631 matched pairs of babies of 30-34 weeks gestational age. The table counts pairs, not babies. Mortality rates are higher in the low c-section group.

	Low Baby			
High Baby	Dead	Alive	Total	High Baby Rate
Dead	108	672	780	3.3%
Alive	1076	21775	22851	96.7%
Total	1184	22447	23631	
Low Baby Rate	5.0%	95.0%		100.0%

Table 7: Mortality by type of death in 1489 matched pairs of babies of 23-24 weeks gestational age. The table counts pairs, not babies. Mortality rates are higher in the low c-section group.

	Low Baby				
High Baby	Fetal Death	Neonatal Death	Alive	Total	High Baby Rate
Fetal Death	111	99	47	257	17.2%
Neonatal Death	220	356	138	714	48.0%
Alive	141	219	158	518	34.8%
Total	472	674	343	1489	
Low Baby Rate	31.7%	45.3%	23.0%		100.0%

Table 8: Mortality by type of death in 23631 matched pairs of babies of 30-34 weeks gestational age. The table counts pairs, not babies. Mortality rates are higher in the low c-section group.

	Low Baby				
High Baby	Fetal Death	Neonatal Death	Alive	Total	High Baby Rate
Fetal Death	64	6	298	368	1.6%
Neonatal Death	26	12	374	412	1.7%
Alive	692	384	21775	22851	96.7%
Total	782	402	22447	23631	
Low Baby Rate	3.3%	1.7%	95.0%		100.0%

Table 9: Mortality R_{ij} and mode of delivery D_{ij} (C = C-section, V = vaginal) in 1489 matched pairs of babies of 23-24 weeks gestational age. For the high baby with $Z_{ij} = 1$, mortality is $R_{ij} = r_{Tij}$ and delivery is $D_{ij} = d_{Tij}$, whereas for the low baby with $Z_{ij} = 0$, mortality is $R_{ij} = r_{Cij}$ and delivery is $D_{ij} = d_{Cij}$. To avoid notational ambiguity, in this table j is changed so the first baby, $j = 1$, is the high baby. The table counts pairs, not babies.

	Low Baby, $Z_{i2} = 0$			
	C-Alive	C-Dead	V-Alive	V-Dead
	$r_{Ci2} = 0$	$r_{Ci2} = 1$	$r_{Ci2} = 0$	$r_{Ci2} = 1$
High Baby, $Z_{i1} = 1$	$d_{Ci2} = 1$	$d_{Ci2} = 1$	$d_{Ci2} = 0$	$d_{Cij} = 0$
C-Alive, $r_{Ti1} = 0$, $d_{Ti1} = 1$	44	54	37	144
C-Dead, $r_{Ti1} = 1$, $d_{Ti1} = 1$	37	38	36	179
V-Alive, $r_{Ti1} = 0$, $d_{Ti1} = 0$	31	35	46	127
V-Dead $r_{Ti1} = 1$, $d_{Ti1} = 0$	47	81	65	488

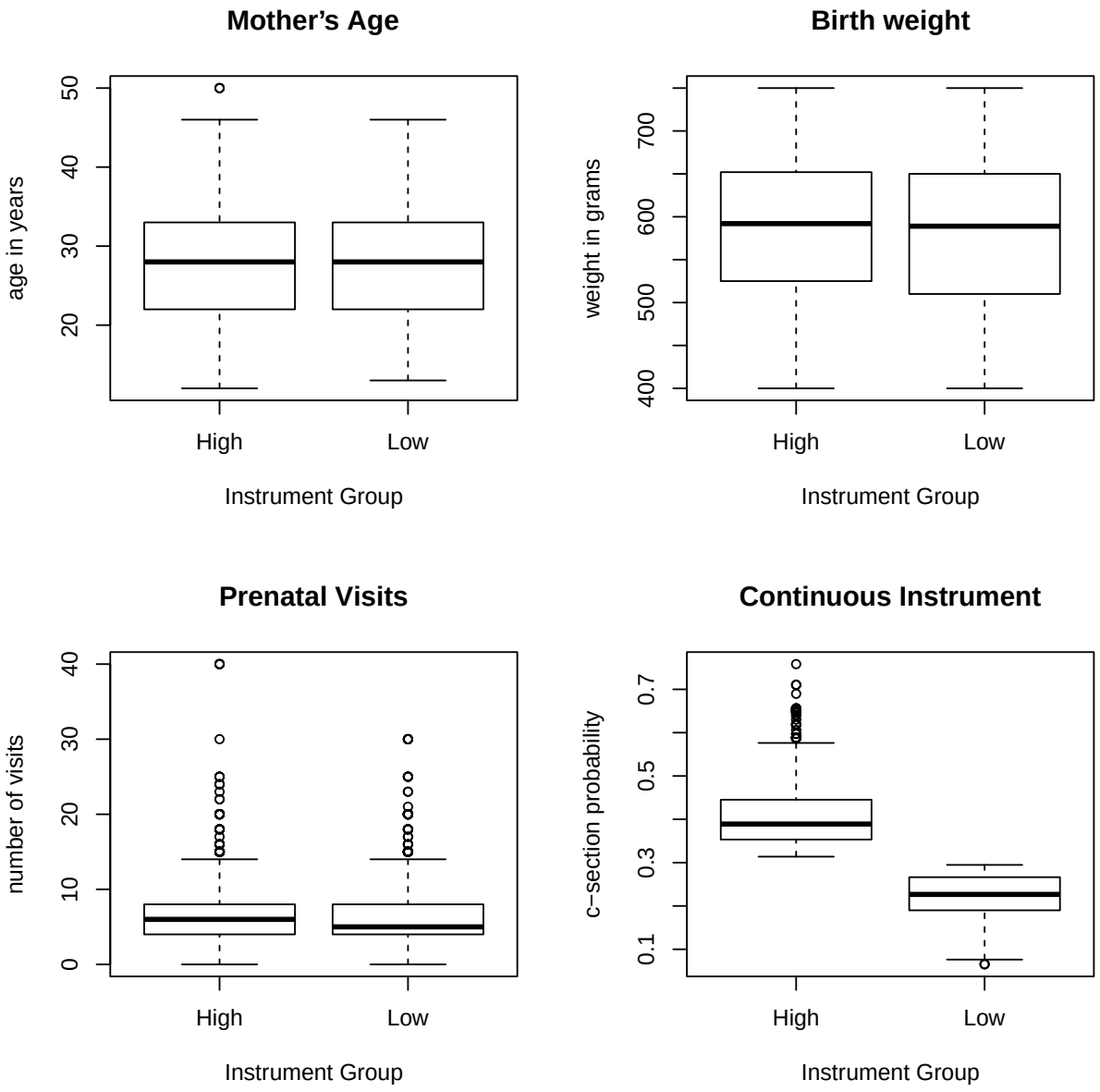

Figure 1: The match was intended to balance covariates and imbalance the instrument, and the boxplots depict this for three continuous covariates – mother's age, birth weight, and number of prenatal visits – and for the continuous instrument – the estimated probability of a c-section at the hospital predicted from c-section rates for older babies.