Dissonant 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 Chicago, 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. Instruments are increasingly used in health outcomes research. An instrument is a haphazard push 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: (R) the push is haphazard or essentially random once adjustments have been made for observed covariates, (E) 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 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; health outcomes research; instrumental variable; observational study.

1 Testing untestable assumptions in causal inference with instrumental variables

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

An instrument is a haphazard push to accept a treatment where the push 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

[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. Fan Yang is assistant professor, Department of Health Studies, University of Chicago, Chicago, IL 60637. José R. Zubizarreta is an assistant professor at the Columbia University School of Business. Dylan S. Small and Paul R. Rosenbaum are professors in the Department of Statistics at the Wharton School of the University of Pennsylvania. Scott Lorch is an associate professor of pediatrics at the University of Pennsylvania School of Medicine. E-mail: dsmall@wharton.upenn.edu. 1 August 2014.

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 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. In Holland's encouragement experiment, the treatment is encouragement or no encouragement, but the active ingredient in encouragement is a change in behavior that may or may not be forthcoming. 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 comparing groups formed by randomization and 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:

(R) encouragement is randomized within pairs or strata defined by observed covariates,

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

In the encouragement design, (R) is ensured by the use of randomization, and (E) 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, sometimes much less compelling, because (R) is

not ensured by actual random assignment, and (E) is less firmly grounded in other things we think we know. In particular, an effort is made to create conditions similar to (R) by carefully and adequately pairing or stratifying to control for important measured covariates, but of course this strategy may fail because an important covariate 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, (E), 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, (R) and (E), that define an instrument.

Instruments are increasingly used in the study of health outcomes; see, for example, McClellan et al. (1994), Lalani et al. (2010) and Lorch et al. (2012). Outside of randomized clinical trials, the treatment a patient receives may reflect a physician's judgment about the best treatment for this patient or else a patient's preference for a particular treatment. Instruments are used in health outcomes research in the hope of finding circumstances in which attributes of the patient do not decide treatment assignment, and instead something haphazard and irrelevant decides the treatment. For instance, if a patient has a heart attack and lives far from a hospital capable of performing coronary bypass surgery, then the heart attack may be treated without bypass surgery just because of where the patient happens to live. Obviously, geography might appear to be haphazard and irrelevant, might appear to satisfy conditions (R) and (E), yet these appearances may be incorrect; so, testing (R) and (E) is important. For several recent discussions of instrumental variables in health outcomes, see Baiocchi et al. (2010), Brookhart and Schneeweiss (2007), Cheng et al. (2011), Swanson and Hernán (2013) and Tan (2006).

1.2 Untestable assumptions or dissonant conclusions

The assumptions required for an instrument are often said to be untestable (e.g., Stock 2002, §4.1). Is this true? Suppose that, if two assumptions, α_1 and α_2, were both true, then it would be possible to prove a theorem deriving some useful consequence from α_1 and α_2. For instance, assumptions (R) and (E) in §1.1 are essentially what is needed to justify a particular confidence interval for an average effect of the treatment on people who change their behavior in response to encouragement; see Imbens and Rosenbaum (2005) and Baiocchi et al. (2010, §3.3). Taken in isolation, the conjunction of α_1 and α_2

may be without testable consequences. We use conventional logical notation to denote conjunction, so $\alpha_1 \wedge \alpha_2$ is true if α_1 is true and α_2 is also true; otherwise, $\alpha_1 \wedge \alpha_2$ is false, that is, $\alpha_1 \wedge \alpha_2$ is false if either α_1 is false or α_2 is false or both are false. In stating a theorem, we often prefer to assume as little as is absolutely needed to secure the proof, and assuming as little as possible may result in assumptions, say $\alpha_1 \wedge \alpha_2$, that are without testable consequences. Now ask: If $\alpha_1 \wedge \alpha_2$ is untestable in isolation, are we, in scientific investigations, therefore absolved from testing $\alpha_1 \wedge \alpha_2$?

In practice, we invariably begin an inquiry with various beliefs, say β_1, ..., β_B. As scientists, our current beliefs are based on empirical evidence, so they do not budge easily, but we recognize that our beliefs may harbor errors, may require revision in light of further evidence. Now it may be that $\alpha_1 \wedge \alpha_2 \wedge \beta_1 \wedge \cdots \wedge \beta_B$ is testable, even though $\alpha_1 \wedge \alpha_2$ is not testable in isolation. What are we to think and say if we test $\alpha_1 \wedge \alpha_2 \wedge \beta_1 \wedge \cdots \wedge \beta_B$ and reject it?

By ancient tradition, the name "aporia" is given to a collection $\mathcal{C}$ of propositions, say $\mathcal{C} = \{\alpha_1, \alpha_2, \beta_1, \ldots, \beta_B\}$, whose members are individually plausible but whose conjunction $\alpha_1 \wedge \alpha_2 \wedge \beta_1 \wedge \cdots \wedge \beta_B$ is false or implausible; see Rescher (2009). In Plato's early dialogues, Socrates would invalidate the views of his opponents by demonstrating that those views constituted an aporia; see Vlastos (1994, p. 58).

We will describe $\mathcal{C}$ using the English phrase "dissonant collection" in place of the Greek word "aporia," but we intend the meaning to be exactly the same. Here, dissonant refers to an absence of harmony. A special case of a dissonant collection $\mathcal{C}$ occurs in mathematics in a proof by contradiction, in which α_1, α_2, β_1, ..., β_{B-1} are known to be true and β_B is proved to be false by showing that $\alpha_1 \wedge \alpha_2 \wedge \beta_1 \wedge \cdots \wedge \beta_B$ is false or equivalently that $\mathcal{C}$ is dissonant. In a proof by contradiction, we know α_1, α_2, β_1, ..., β_{B-1} are true, so if conjoining β_B yields a contradiction, we know not only that $\mathcal{C}$ is dissonant, but also why it is dissonant: it is dissonant because of β_B, and we deduce that β_B is false. A randomization test of the null hypothesis of no effect in a randomized experiment follows a similar logic, where α_1, α_2, β_1, ..., and β_{B-1} describe the randomized experimental design and β_B is the null hypothesis. In sharp contrast, in much scientific work, we may recognize $\mathcal{C}$ as dissonant but not know which elements of $\mathcal{C}$ are culpable for the contradiction.

Our main claim is that the IV assumptions, (R) and (E), are often testable when conjoined with other things we think we know — i.e., that $\mathcal{C}$ is often testable — but this test may result in dissonance, in evidence that $\mathcal{C}$ contains at least one false claim, without clear indication as to which claims in $\mathcal{C}$ are responsible for the contradiction or

the implausibility of $\mathcal{C}$ as a whole. We discuss an example in detail in §2.

1.3 Why does dissonance matter? The logic of dissonance

The current section, which may be skipped, makes use of elementary notions from propositional logic to clarify the concept of a dissonant collection $\mathcal{C}$ of propositions. To recognize that one's beliefs $\mathcal{C}$ are dissonant is an advance in understanding, albeit an uncomfortable one. In elementary logic, from a false premise, one can deduce every conclusion, true or false (because, in elementary propositional logic, $\delta \Rightarrow \kappa$ is true for all κ if δ is false). To fail to recognize $\mathcal{C}$ as dissonant is to risk logically deducing false propositions from beliefs one holds (because one believes α_1, α_2, β_1, ..., β_B, can deduce the false proposition $\delta = \alpha_1 \wedge \alpha_2 \wedge \beta_1 \wedge \cdots \wedge \beta_B$ from one's beliefs, and can deduce any κ from δ because δ is false). To recognize $\mathcal{C}$ as dissonant 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 $\mathcal{C}$. To recognize $\mathcal{C}$ as dissonant is an advance in understanding, and it is certainly better than unqualified belief in $\mathcal{C}$.

The components of $\mathcal{C} = \{\alpha_1, \alpha_2, \beta_1, \ldots, \beta_B\}$ may not be testable one at a time, in isolation from each other, whereas δ may be testable. In this sense, the assumptions of IV, namely R and E in §1.1 (which we now let equal α_1 and α_2 respectively), differ from the assumptions of a one-way analysis of variance, where Normality of errors or equality variances each can be tested in isolation from the null hypothesis of equality of group means.

In a logical argument, one could eliminate the dissonance of $\mathcal{C} = \{\alpha_1, \alpha_2, \beta_1, \ldots, \beta_B\}$ by ejecting elements of $\mathcal{C}$ to yield $\mathcal{C}' \subset \mathcal{C}$ such that $\mathcal{C}'$ is no longer logically contradictory. In this process of ejection, there is nothing to ensure that one has ejected false beliefs 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 in $\mathcal{C}'$ that belief in $\mathcal{C}'$ is safe from accusations of inconsistency in $\mathcal{C}'$. In scientific work with instrumental variables, two types of arbitrary ejection are common, perhaps typical: ejection of $\beta_1, \ldots, \beta_B$ because they are based on other prior investigations done by someone else, or ejection of the IV assumptions α_1, α_2 because they imply conclusions inconsistent with prior investigations. Both of these forms of arbitrary ejection attain consistency of $\mathcal{C}' \subset \mathcal{C}$ without providing a reason to believe the propositions retained in $\mathcal{C}'$ are true. Perhaps the fault lies with the current investigation, perhaps with a prior investigation, but simply rejecting δ cannot settle that.

Suffice it to say, we do not believe arbitrary ejection of elements of $\mathcal{C}$ is a sound approach to eliminating dissonance. Arbitrary ejection eliminates contradiction but cannot be expected to secure truth. Rejecting the IV assumptions α_1, α_2 and keeping β_1, ..., β_B risks failing to recognize an error in our background beliefs β_1, ..., β_B. Declaring the IV assumptions α_1, α_2 as untestable by excluding consideration of β_1, ..., β_B means failing to look at evidence that α_1, α_2 may be false. To recognize $\mathcal{C}$ as dissonant is an uncomfortable advance in understanding, but it is nonetheless an advance, and it is better than preserving error and persevering on the basis of error. The situation is well described by John Stuart Mill (1859, §2):

> It is the fashion of the present time to disparage negative logic—that which points out weaknesses in theory or errors in practice, without establishing positive truths. Such negative criticism would indeed be poor enough as an ultimate result; but as a means to attaining any positive knowledge or conviction worthy the name, it cannot be valued too highly.

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. Background is discussed in §2.1 and the IV analysis is presented in §2.2-§2.4. In §2.6, the IV assumptions are tested, resulting in dissonance that is discussed in detail. The two appendices in an on-line supplement present two technical innovations: a new simpler approach to strengthening an instrument in on-line Appendix I, and a sensitivity analysis for an attributable effect closely related to the Wald estimator in on-line Appendix II. A reader who wishes to reproduce the analysis will need to consult these appendices. We placed this material in appendices because we wanted to emphasize the conceptual discussion of dissonant conclusions when testing the assumptions underlying IV.

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 versus vaginal delivery on the survival of very premature babies of 23-24 weeks gestational age. For rea-

sons 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 40 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). A survival advantage may occur because of lower rates of severe complications of premature birth in infants born via cesarean section, such as intraventricular hemorrhage; see Dani (2010) and Deulofeut et al. (2005). 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 but do not address the problem of confounding by indication. The analyses implicitly assume that the logit models include all covariates simultaneously relevant to survival and mode of delivery, that there is no unmeasured confounding. In contrast to their analyses, their discussions 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 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. Delivering at a hospital that makes extensive use of c-sections is intended to be analogous to the "encouragement treatment" in Holland's encouragement design in §1.1, but the active ingredient in encouragement is a change in method of delivery, c-section or vaginal.

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. 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. A 23-24 week baby with a high value of the instrument is in an hospital that often uses c-sections for older babies ($\geq$ 25

weeks) and is in a county with a high malpractice insurance rate, whereas a baby with a low value of the instrument is in a hospital that rarely uses c-sections for older babies and is in a county with a lower malpractice insurance rate. The value of the instrument is a function of the hospital's c-section rate for three groups of older babies, (a)-(c), and the malpractice rate (d). The instrument does not vary with the hospital's own c-section rate for 23-24 week babies, except in the very oblique sense that this rate affects the four estimated logit regression coefficients in an analysis using all the hospitals. The situation is similar to that in two-stage least squares, one common technique for estimation with instrumental variables. Again, the idea is that the value of the instrument tells you little or nothing about the health and prospects of an individual 23-24 week baby, but it does tell you something about the chance the baby will be delivered by c-section.

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. These scores were estimated from older babies, ≥ 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. There were 27 covariates plus several indicators for missing values for these covariates. 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. Low and high were defined by the 45th and 55th quantiles of the instrument, ≤ 0.29 and ≥ 0.31, with the middle 10% discarded; see the on-line Appendix I. As discussed by Baiocchi et al.

(2010), forcing separation on the instrument tends to strengthen it, eliminating some of the problems with weak instruments; see on-line Appendix I. 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 of 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, and 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 other covariates that were continuous or integer-valued were hospital volume, risk adjusted rates of thrombosis and wound infection, parity, month prenatal care started, multiple delivery (1, 2, 3, 4, with 2 = twins), Census tract median household income and percent below poverty; moreover, boxplots of these covariates exhibit balance comparable to Figure 1.

Tables 1-3 and Figure 1 demonstrate that the matching was reasonably successful at balancing the observed covariates, but differences in unmeasured covariates remain as a possibility. We speak to this possibility in §2.4 by examining sensitivity to bias from unmeasured covariates. Before matching, the high instrument group had several advantages that were removed by matching: more prenatal care visits, higher neighborhood median income and lower neighborhood poverty level. Although these specific differences were removed, it is not inconceivable that unmeasured covariates would also suggest the high instrument group is wealthier and receiving better health care.

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. PROM = premature rupture of membrane.

	Before matching			After matching		
	Mean		S-Dif	Mean		S-Dif
	High	Low		High	Low	
	Hospital Covariates					
Hospital delivery volume (#)	2850	2903	-0.03	2568	2722	-0.09
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/Baby Covariates					
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
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
PROM (Y/N)	0.35	0.28	0.15	0.33	0.30	0.08
Small for gestation age (Y/N)	0.09	0.12	-0.11	0.09	0.12	-0.09
	Neighborhood Covariates from the Census					
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
Below Poverty Level (fraction)	0.11	0.16	-0.10	0.15	0.16	-0.03
	Instrumental variable					
C-sec. predicted prob.	0.38	0.23	2.56	0.40	0.22	3.12

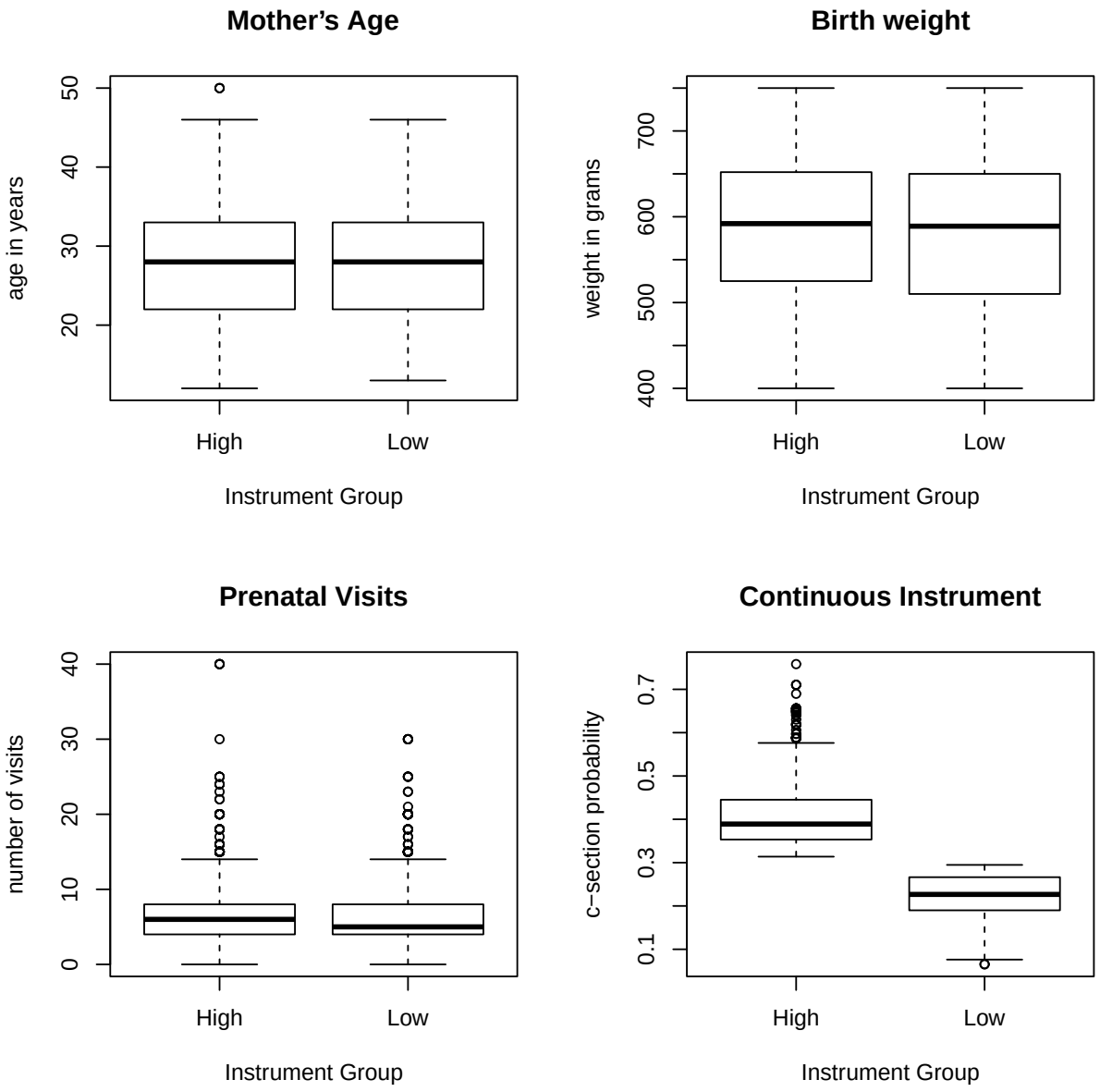

Figure 1: The match was intended to balance covariates and imbalance the instrument, and the boxplots depict this for three continuous covariates – mother's age, birth weight, and number of prenatal visits – and for the continuous instrument – the estimated probability of a c-section at the hospital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

The matching was done in a new but simple way described in on-line Appendix I. The approach taken here is a small extension and slight simplification of the approach taken in Zubizarreta et al. (2013). 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.

2.4 Outcomes: c-section and mortality rates

In the current section, the effect of c-sections on mortality are estimated assuming (R) and (E) from §1.1; then, §2.6 tests these assumptions. The analysis that follows summarizes in words an analysis that is described in formal language in on-line Appendix II. A reader who wants more precision or technical detail should turn to Appendix II.

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 (1952, 1970). More than half the babies in both the high and low groups were delivered vaginally; however, 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. So in comparing the high and low instrument groups, we are comparing groups with quite different rates of c-sections. These differing rates of c-sections are attributes of the hospitals at which the babies were born, not attributes of the individual babies in the comparison.

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 matched-pair point estimate of the odds ratio (Cox 1952, §2; 1970, §5.2) favoring survival at a hospital with a high c-section rate is $360/185 = 1.95$. In the high group, the survival rate was 34.8% and in the low group it was 23.0%, or a difference of $360 - 185 = 175$ survivors. Some of these additional survivors may reflect effects caused by c-sections. A certain unknown number, A, of babies born at high c-section hospitals survived by virtue of being born at a high

c-section hospital rather than a low c-section hospital, where this causal effect A is called the "attributable effect"; see Rosenbaum (2002) or on-line Appendix II. 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 (R) but perhaps not (E) in §1.1), then, using the method in Rosenbaum (2002, §6), one would be 95% confident that $A \geq 133$ babies were caused to survive because of delivery at a high hospital. Moving away from the naive model for treatment assignment — i.e., moving away from (R) in §1.1 — if the study were biased by an unobserved covariate that at most doubled the odds of delivery at a high hospital and at most doubled the odds of survival, then the one-sided 95% confidence interval is $A \geq 67$ babies caused to survive because of delivery at a high hospital; see Rosenbaum (2002, §6) and 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; again, see Appendix II for precise statements. 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.

With a binary instrument, high-versus-low, the Wald estimate is the high-versus-low difference in mean outcomes, here mortality, divided by the high-versus-low difference in mean treatment, here c-sections; see, again, §1.1 for informal motivation for the Wald estimate, and Angrist et al. (1996) for a formal discussion. 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, ignoring sampling variability, would attribute the additional 175 survivors at high c-section hospitals to the 202 additional c-sections at those hospitals. If it were assumed that the high-versus-low grouping is a valid instrument — that is, assuming both (R) and (E) in §1.1 — then 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 would be $175/202 = 0.87$. This is 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

the number of additional c-sections at high c-section hospitals. In on-line Appendix II, it is noted that A/D is the ratio of an unobserved to an observed random variable and a new confidence interval for it is discussed. The 95% confidence intervals for A/D are $A/D \geq 133/202 = 0.66$ for randomization inference, $A/D \geq 67/202 = 0.33$ allowing for an unobserved covariate that doubled the odds of delivering at a high c-section hospital and doubled the odds of survival, 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.

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%

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 Assumptions and beliefs: the basis for a test of the IV assumptions

We wish to test the conjunction of the two IV assumptions, (R) and (E), by deducing consequences of these assumptions and other background beliefs, and checking the consequences against the data. Here, there are the two IV assumptions and one plausible claim from the literature, or $\mathcal{C} = \{\alpha_1, \alpha_2, \beta_1\}$ in the notation of §1.2, as follows.

α_1: After adjusting for the observed covariates in Tables 1-3 and Figure 1, the high or low level of the instrument is, in effect, randomized; that is, the level of the instrument is conditionally independent of the potential outcomes of the baby conditionally given the observed covariates. This is (R) in §1.1. See Imbens and Rosenbaum (2005) for a precise statement.

α_2: Switching a baby from the low level of the instrument to the high level can only change the baby's survival if it changes whether the baby was delivered vaginally or by c-section. This is (E) in §1.1.

β_1: Based on the literature cited in §2.1, available evidence suggests increased use of c-sections may or may not benefit extremely premature babies of 23-24 weeks gestational age, but it is without survival benefit in older premature babies of 30-34 weeks gestational age.

If $\alpha_1 \wedge \alpha_2$ were true, then IV would produce valid permutational confidence intervals for the effect of c-sections on survival; see Baiocchi et al. (2010, §3.3). If $\alpha_1 \wedge \alpha_2 \wedge \beta_1$ were true, then a permutational 95% confidence interval from IV for the effect of c-sections on 30-34 week babies will exclude "no effect" with probability at most 0.05. In §2.4, (i) we estimate an effect for extremely premature infants of 23-24 weeks gestational age where the effect is not thought to be known, (ii) but in §2.6 we test the IV assumptions by also estimating the effect for older premature infants of 30-34 weeks gestation age where the effect is thought to be zero.

As it turns out, the test in §2.6 rejects $\alpha_1 \wedge \alpha_2 \wedge \beta_1$, so we find $\mathcal{C} = \{\alpha_1, \alpha_2, \beta_1\}$ is dissonant, that at least one element of $\mathcal{C}$ should be rejected, but we do not know whether $\alpha_1 \wedge \alpha_2$ is false or β_1 is false. To our minds, the IV assumptions, $\alpha_1 \wedge \alpha_2$ are plausible but far from certain in this example. The claim β_1 is plausible on its face, but the empirical evidence in favor of β_1 is essentially based on various logit models in the literature that predict survival using covariates and type of delivery, c-section or vaginal. So unlike IV

estimates, the evidence in favor of β_1 could be biased if c-sections are used selectively based on unmeasured attributes of individual babies or their mothers. Rejecting $\alpha_1 \wedge \alpha_2 \wedge \beta_1$ would make $\mathcal{C} = \{\alpha_1, \alpha_2, \beta_1\}$ dissonant — you then could not reasonably believe both β_1 and the IV estimates for 23-24 week babies — but it leaves open the cause of the dissonance.

2.6 A test of the exclusion restriction

A complete description of the test of IV assumptions requires some attention to detail, but before presenting that detail, we state the finding concisely. If $\alpha_1 \wedge \alpha_2 \wedge \beta_1$ were true then you get a contradiction, or as close as statistical evidence can get to a logical contradiction, in that $\alpha_1 \wedge \alpha_2$ justify estimates that contradict β_1. Specifically, $\alpha_1 \wedge \alpha_2$ justify IV estimates that find benefits for babies of 23-24 weeks gestational age, but also for babies of 30-34 weeks gestational age, the latter contradicting the literature's consensus that there are no benefits for babies of 30-34 weeks gestational age. It is possible that the literature and the professional consensus are correct and the IV estimates are wrong. The literature, however, is based on assuming that logit regression removes all selection biases, and that assumption is no more plausible, perhaps substantially less plausible, than the IV assumptions; so, possibly β_1 is wrong. If $\mathcal{C} = \{\alpha_1, \alpha_2, \beta_1\}$ is dissonant, then available evidence is not adequate for evidence-based medical practice, contrary to what is claimed in the literature.

As discussed in §2.1 and §2.5, 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 dissonance 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, 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 from §2.4 and $1076/672 = 1.60$ for 30-34 weeks in Table 6. 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.

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%

So the claims in the literature, that is β_1, and our results, based on α_1 and α_2, 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 IV estimates based on α_1 and α_2 find a benefit for older newborns that β_1 denies. The conclusion is dissonant, 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 several cited papers on c-sections implicitly assume 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), but that is not a safe assumption; rather it is an assumption that gets people in no end of trouble in observational studies. Are there other possibilities?

2.7 A digression about an issue unrelated to testing IV assumptions

Indeed, there is another possibility. The cited literature in §2.1 focused on neonatal deaths, excluding fetal deaths, whereas we looked at all deaths; see §2.1 for definitions of these terms. 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.

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%

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. So, this distinction between fetal and prenatal deaths may explain an aspect of dissonance among published findings. However, neonatologists doubt that c-sections benefit more mature newborns, those of 30-34 weeks gestational age, whereas the comparison in Table 8 suggests a nontrivial survival benefit, and this continues to raise concerns about the instrumental variable assumptions (R) and (E) in §1.1 that underlie Table 8.

2.8 Dissonant evidence should spur further investigation

The available evidence is dissonant. 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 claims a benefit from c-sections at 23-24 weeks gestational age but not at 30-34 weeks gestational age, and the latter absence of benefit is consistent with what neonatologists expected. 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 23-24 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. It is not inconceivable that the IV assumptions, (R) and (E) in §1.1, are true for one group of babies defined by gestational age and false for the other, but there is no evidence to support this. 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 dissonance is seen as an acceptable conclusion. Dissonance refers to a collection of individually plausible but mutually incompatible propositions. Dissonance 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.

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.

Baicker, K., Buckles, K.S. and Chandra, A. (2006), "Geographic variation in the appropriate use of cesarean delivery " *Health Affairs*, 24, w355-w467.

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.

Brookhart, M. and Schneeweiss, S. (2007), "Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results," *International Journal of Biostatistics*, 3, 14.

Cheng, J., Qin, J. and Zhang, B. (2011), "Semiparametric estimation and inference for distributional and general treatment effects," *Journal of the Royal Statistical Society*, B, 71, 881-904.

Cox, D. R. (1952), "Two further applications of a model for binary regression," *Biometrika*, 45, 562-565.

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

Dani, C., Poggi, C., Bertini, G., Pratesi, S., Di Tommaso, M., Scarselli, G. and Rubaltelli, F. F. (2010), "Method of delivery and intraventricular haemorrhage in extremely preterm infants," *Journal of Maternal, Fetal and Neonatal Medicine*, 23, 1419-1423.

Deulofeut, R., Sola, A., Lee, B., Buchter, S., Rahman, M. and Rogido, M. (2005), "The impact of vaginal delivery in premature infants weighing less than 1,251 grams," *Obstetrics and Gynecology*, 105, 525-531.

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

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

Imbens, G. W. and Rosenbaum, P. R. (2005), "Robust, accurate confidence intervals with a weak instrument," *Journal of the Royal Statistical Society* A 168, 109-126.

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 endocarditis: 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.

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

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.

Mill, J. S. (1859), *On Liberty*, London: Parker.

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

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. (`amplify` function in `R` package `sensitivitymv`)

Swanson, S. A. and Hernan, M. A. (2013), "How to report instrumental variable analyses," *Epidemiology*, 24, 924-933.

Stock, J. (2002), "Instrumental Variables in Economics and Statistics," in: *International Encyclopedia of the Social Sciences*, Amsterdam: Elsevier, 7577-7582.

Tan, Z. (2006), "Regression and weighting methods for causal inference using instrumental variables," *Journal of the American Statistical Association*, 101, 1607–1618.

Vlastos, G. (1994), *Socratic Studies*, New York: Cambridge University Press.

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.

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.