Estimating marginal returns to medical care: Evidence from at-risk newborns*

Douglas Almond†
Joseph J. Doyle, Jr.‡
Amanda E. Kowalski§
Heidi Williams¶

November 21, 2008

Abstract

We estimate marginal returns to medical care for at-risk newborns by comparing health outcomes and medical treatment provision on either side of common risk classifications, most notably the "very low birth weight" threshold at 1500 grams. First, using data on the census of US births in available years from 1983-2002, we find evidence that newborns with birth weights just below 1500 grams have *lower* one-year mortality rates than do newborns with birth weights just above this cutoff, even though mortality risk tends to decrease with birth weight. One-year mortality falls by approximately one percentage point as birth weight crosses 1500 grams from above, which is large relative to mean one-year mortality of 5.5% just above 1500 grams. Second, using hospital discharge records for births in five states in available years from 1991-2006, we find evidence that newborns with birth weights just below 1500 grams have discontinuously higher costs and frequencies of specific medical inputs. We estimate a $4,000 increase in hospital costs as birth weight approaches 1500 grams from above, relative to mean hospital costs of $40,000 just above 1500 grams. Taken together, these estimates suggest that the cost of saving a statistical life of a newborn with birth weight near 1500 grams is on the order of $550,000 in 2006 dollars.

*We are very grateful to Christine Pal and Jean Roth for assistance with the data, to Christopher Afendulis and Ciaran Phibbs for sharing the California neonatal intensive care unit data, and to Drs. Chris Almond, Burak Alsan, Munish Gupta, Chafen Hart, and Katherine Metcalf for helpful discussions regarding neonatology. David Autor, Amitabh Chandra, David Cutler, Dan Fetter, Amy Finkelstein, Edward Glaeser, Michael Greenstone, Jonathan Gruber, Jerry Hausman, Guido Imbens, Lawrence Katz, Michael Kremer, Ellen Meara, Derek Neal, Joseph Newhouse, James Poterba, Gary Solon, Tavneet Suri, and participants in seminars at Harvard, the Harvard School of Public Health, MIT, and the fall 2008 NBER Labor Studies meeting provided helpful comments and feedback. We use discharge data from the Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality, provided by the Arizona Department of Health Services, Maryland Health Services Cost Review Commission, New Jersey Department of Health and Senior Services, and the New York State Department of Health. Funding from the National Institute on Aging, Grant Number T32-AG000186 to the National Bureau of Economic Research, is gratefully acknowledged (Doyle, Kowalski, Williams).

†Columbia University and NBER: da2152@columbia.edu
‡MIT and NBER: jjdoyle@mit.edu
§NBER: kowalski@nber.org
¶Harvard University: hlwill@fas.harvard.edu

I. Introduction

Medical expenditures in the United States are high and increasing. A key policy question is whether the benefits of additional medical expenditures exceed their costs, yet several challenges complicate the empirical measurement of returns to medical spending. One fundamental challenge is that we expect less healthy patients to receive more medical inputs, likely leading to a downward bias in estimates of returns. To overcome this challenge, a variety of studies have used cross-sectional, time-series, and panel data techniques to identify patients who are arguably similar in terms of underlying health status but who for some reason receive different levels of medical spending. The results of such studies are mixed. On one hand, time-series and panel data studies that compare increases in spending and improvements in health outcomes over time have argued that increases in costs have been less than the value of the associated benefits, at least for some technologies.[1] On the other hand, cross-sectional studies that compare "high-spending" and "low-spending" geographic areas tend to find large differences in spending yet remarkably similar health outcomes.[2]

The lack of consensus in these studies may not be surprising since these studies often estimate returns to different types of medical spending. The return to a dollar of medical spending likely differs across medical technologies and across populations, and the return to the first dollar of medical spending in any given context may differ from the return to the last dollar of medical spending. The time-series studies often estimate returns to large changes in treatments that occur over long periods of time. The cross-sectional studies, on the other hand, estimate returns to additional, incremental spending that occurs in some areas but not others.[3] While estimates of returns to large changes in medical spending are useful summaries of changes over time, estimates of marginal returns are needed to inform policy decisions over whether to increase or decrease the level of care in a given context.

The main innovation in this paper is a novel research design for more direct estimation of the marginal costs and marginal benefits of medical spending, which we can combine to estimate

[1] See, for example, Cutler & McClellan (2001), Cutler *et al.* (1998), Cutler *et al.* (2006), Luce *et al.* (2006), McClellan (1997), Murphy & Topel (2003), and Nordhaus (2002).

[2] See, for example, Baicker & Chandra (2004), Fisher *et al.* (1994), Fuchs (2004), Kessler & McClellan (1996), O'Connor *et al.* (1999), Pilote *et al.* (1995), Stukel *et al.* (2005), and Tu *et al.* (1997).

[3] Some have interpreted the cross-sectional results as evidence that marginal returns to spending may be low in many contexts.

marginal returns. Our research design is applicable in settings with an observable, continuous measure of health risk and a diagnostic threshold (based on this risk variable) that generates a discontinuous probability of receiving additional treatment. In such empirical settings, we can use a regression discontinuity framework: as long as other factors are smooth across the threshold (an assumption we investigate in several empirical tests), individuals within a small bandwidth on either side of the threshold should differ only in their probability of receiving additional health-related inputs and not in their underlying health. This research design allows us to estimate marginal returns to medical spending for patients near such thresholds in the following sense: conditional on estimating that, on average, patients on one side of the threshold incur additional medical costs, we can estimate the associated benefits by examining average differences in health outcomes across the threshold. Combining these cost and benefit estimates then allows us to calculate the return to this increment of additional spending, or "average marginal returns."

We apply our research design to study "at risk" newborns, a population that is of interest for several reasons. First, even relatively small reductions in mortality for newborns can be magnified in terms of the total number of years of life saved. Second, technologies for treating at-risk newborns have expanded tremendously in recent years, at very high cost. For example, a 2005 Agency for Healthcare Research and Quality study finds that the top two most expensive diagnoses (regardless of age) are "infant repiratory distress syndrome" (average charges of $91,000) and "premature birth and low birthweight" (average charges of $79,300). Third, although existing estimates suggest that the benefits associated with large changes in spending on at risk newborns over time have been greater than their costs (Cutler & Meara, 2000), there is a dearth of evidence on the returns to incremental spending in this context. Fourth, studying newborns allows us to focus on a large portion of the health care system, as child birth is one of the most common reasons for hospital admission in the US. This patient population also provides samples large enough to detect effects of additional treatment on infant mortality.

This paper focuses on the "very low birth weight" (VLBW) classification at 1500 grams (just under 3 pounds, 5 ounces) - a threshold commonly cited in the medical literature. We also consider other classifications based on birth weight and alternative measures of newborn health. From an empirical perspective, birth weight-based thresholds provide an attractive basis for a

regression discontinuity design for several reasons. First, they are unlikely to represent breaks in underlying health risk. A 1985 *Institute of Medicine* report, for example, notes: "...designation of very low birth weight infants as those weighing 1,500 grams or less reflected convention rather than biologic criteria." Second, it is generally agreed that birth weight cannot be predicted in advance of delivery with the accuracy needed to change (via birth timing) the classification of a newborn from being just above 1500 grams to being just below 1500 grams. Thus, although we empirically investigate our assumption that the position of a newborn just above 1500 grams relative to just below 1500 grams is "as good as random," the medical literature suggests this assumption is also intuitively plausible.

To preview our main results, using data on the census of US births in available years from 1983-2002, we find that one-year mortality increases by approximately one percentage point as birth weight crosses the VLBW threshold from above, which is large relative to mean one-year mortality of 5.5% just above 1500 grams. This increase in mortality is in stark contrast to the overall decline in mortality as birth weight increases, and to the extent that heavier newborns are healthier in unobservable ways, the increase in mortality that we observe is all the more striking. Second, using hospital discharge records for births in five states in available years from 1991-2006, we estimate a $4,000 increase in hospital costs for infants just below the 1500 gram threshold, relative to mean hospital costs of $40,000 just above 1500 grams. Taken together, our estimates suggest that the cost of saving a statistical life for newborns near 1500 grams is approximately $550,000 in 2006 dollars. The upper bound of the 95% confidence interval on this estimate is approximately $1.2 million. As we discuss in Section IX, our estimate can be compared to a variety of cost-effectiveness benchmarks. Compared to a time-series estimate in the spirit of Cutler & Meara (2000), which compares changes in spending and mortality over time for newborns within our bandwidth (described in more detail in Section IX), our estimate of marginal returns is similar or slightly more cost-effective.

The remainder of the paper is organized as follows. Section II discusses the available evidence on the costs and benefits of medical care for at-risk newborns, gives a brief background on the origins of the at-risk newborn classifications we study, and discusses why these classifications might affect the provision of medical treatment. Section III describes our data and analysis sample, and Section IV outlines our empirical framework, estimation strategy, and bandwidth

selection. Section V presents our main results, and Section VI presents several robustness and specification checks. Section VII examines variation in our estimated treatment effects over time, across hospitals, and across newborn subgroups. Section VIII discusses evidence from other thresholds. In Section IX we combine our main estimates to calculate two-sample estimates of marginal returns, and we conclude in Section X.

II. Background

A. Costs and benefits of medical care for at-risk newborns

Technologies for treating at-risk newborns have expanded tremendously in recent years - including the development of improved ventilators and an expansion in the number of neonatal intensive care units. Many of these medical advances have been very expensive. For example, in 2005 the US Agency for Healthcare Research and Quality estimated that the two most expensive hospital diagnoses (regardless of age) were "infant respiratory distress syndrome" and "premature birth and low birth weight."[4] Russell *et al.* (2007) estimated that in the US in 2001, preterm and low-birth weight diagnoses accounted for 8% of newborn admissions, but 47% of the costs for all infant hospitalizations (at $15,100 on average).[5] Despite their high and highly-concentrated costs, use of new neonatal technologies has continued to expand. An example related to our threshold of interest is provided by the Oxford Health Network's 362 hospitals, where the use of high-frequency ventilation among VLBW infants tripled between 1991 and 1999 (Horbar *et al.*, 2002).

These high costs naturally motivate the question of what these medical advances have been "worth" in terms of improved health outcomes. Anspach (1993) and others discuss the paucity of randomized controlled trials which measure the effectiveness of neonatal intensive care. In the absence of such evidence, some have questioned the effectiveness of these increasingly intensive treatment patterns (Enthoven, 1980; Grumbach, 2002; Goodman *et al.*, 2002).[6] On the other

[4] See http://www.ahrq.gov/data/hcup/factbk6/factbk6.pdf (accessed 29 October 2008).

[5] It is worth noting that these costs may disproportionately fall on the public sector. Russell *et al.* (2007) note that of all preterm/low birth weight infant stays, 42% designated Medicaid as the expected payer, relative to 37.5% for "uncomplicated" newborns.

[6] Grumbach (2002), for example, argues: "Neonatal intensive care units are profit-making centers for hospitals, commanding high payments from private and public insurance plans," and that "uncontrolled growth" in neonatology "has less to do with the true need of communities for effective clinical services than with the financial incentives." Going a step further, Goodman *et al.* (2002) ask whether "infants might be harmed by the availability

hand, Cutler & Meara (2000) examine time-series variation in birth weight-specific treatment costs and mortality outcomes and argue that medical advances for newborns have had large returns.[7]

B. "At risk" newborn classifications

We focus on a commonly-used newborn risk classification: the "very low birth weight" (VLBW) classification at 1500 grams (just under 3 pounds, 5 ounces). We also examine other birth weight classifications, including the "extremely low birth weight" (ELBW) classification at 1000 grams (just over 2 pounds, 3 ounces) and the "low birth weight" (LBW) classification at 2500 grams (just over 5 pounds, 8 ounces), as well as gestational-age based measures such as the "prematurity" classification at 37 gestational weeks - where gestational weeks are defined as the number of weeks since the first day of the mother's last menstrual period. This subsection provides some background on these three classifications.[8]

Physicians had begun to recognize and assess the relationships among inadequate growth (low birth weight), shortened gestation (prematurity), and mortality at least by the early 1900s. The 2500 gram low birth weight classification, for example, has existed since at least 1930, when a Finnish pediatrician advocated 2500 grams as the birth weight below which infants were at high risk of adverse neonatal outcomes. This recommendation was formally adopted by the World Health Organization (WHO) on two separate occasions, in 1948 and 1950. A 1985 Institute of Medicine (IOM) report notes that both recommendations indicated that the use of a birth weight marker served as a shorthand notation for a variety of interrelated physiologic processes affecting fetal growth and the duration of gestation. The 1985 IOM report also notes that over time, interest increased in the outcomes of subgroups of low birth weight infants, in particular the fate of the smallest infants, and that conventionally "very low birth weight" infants were defined as those born weighing 1500 grams or less.

of higher levels of resources."

[7]Cutler and Meara's empirical approach assumes that all within-birth weight changes in survival have been due to improvements in medical technologies. This approach is motivated by the argument that conditional on birth weight, the overwhelming factor influencing survival for low birth weight newborns is medical care in the immediate postnatal period (Paneth, 1995; Williams & Chen, 1982). However, others have noted that it is possible that underlying changes in the health status of infants within each weight group (due to, for example, improved maternal nutrition) are responsible for neonatal mortality independent of newborn medical care (United States Congress, Office of Technology Assessment, 1981).

[8]The discussion in this section draws heavily from United States Institute of Medicine (1985).

The key to our empirical strategy is that these cutoffs appear to truly reflect convention rather than strict biologic criteria. For example, the 1985 IOM report notes:

> *"Birth weight is a continuous variable and the limit at 2,500 grams does not represent a biologic category, but simply a point on a continuous curve. The infant born at 2,499 grams does not differ significantly from one born at 2,501 grams on the basis of birth weight alone...As with the 2,500 gram limit, designation of very low birth weight infants as those weighing 1,500 grams or less reflected convention rather than biologic criteria."*

The 1961 WHO Expert Committee on Maternal and Child Health issued further recommendations emphasizing the importance of the 2500 gram and 37 gestational week classifications. While gestational age is a natural consideration when determining treatment for newborns with low birth weights, a comparison of treatments by gestational age has the added complication that gestational age is a choice variable that is likely affected by the prematurity thresholds. Gestational age is known to women in advance of giving birth, and women can choose to time their birth (for example, through an induced vaginal birth or through a C-section) based on gestational age. Thus, *a priori* we would expect that mothers who give birth prior to 37 gestational weeks may be different from mothers who give birth after 37 gestational weeks on the basis of factors other than gestational age. It is thought that birth weight, on the other hand, cannot be predicted in advance of birth with the accuracy needed to change (via birth timing) the classification of a newborn from being just above 1500 grams to being just below 1500 grams; this assertion has been confirmed from conversations with physicians,[9] as well from studies such as Pressman *et al.* (2000). Gestational age is also thought to be considerably more difficult to measure than is birth weight (see, for example, the 1985 IOM report); however, to the extent that we as researchers observe the same measure of gestational age that women and doctors

[9]We use the phrase "conversations with physicians" somewhat loosely throughout the text of the paper to reference discussions with several physicians as well as readings of the relevant medical literature and references such as the *Manual of Neonatal Care* for the Joint Program in Neonatology (Harvard Medical School, Beth Israel Deaconess Medical Center, Brigham and Women's Hospital, Children's Hospital Boston) (Cloherty & Stark, 1998). The medical doctors we spoke with include Dr. Christopher Almond from Children's Hospital Boston (Boston, MA); Dr. Burak Alsan from Harvard Brigham and Women's/Children's Hospital Boston (Boston, MA); Dr. Munish Gupta from Beth Israel Deaconess Medical Center (Boston, MA); Dr. Chafen Hart from the Tufts Medical Center (Boston, MA); and Dr. Katherine Metcalf from Saint Vincent Hospital (Worchester, MA). We are very grateful for their time and feedback, but they are of course not responsible for any errors in our work.

observe, this is not "measurement error" in the traditional sense and should not be problematic for our research design.

To define treatment of observations occurring exactly at the relevant cutoffs, we rely on definitions listed in the International Statistical Classification of Diseases and Related Health Problems (ICD-9) codes. According to the ICD-9 codes, very low birth weight is defined as having birth weight strictly less than 1500 grams, and analogously (with a strict inequality) for the other thresholds we examine.

C. Should these cutoffs matter?

Of course, birth weight and gestational age are not the only factors used to assess newborn health. For example, respiratory rate, color, APGAR score (an index of newborn health), head circumference, and presence of congenital anomalies could also affect physicians' initial health assessments of infants (Cloherty & Stark, 1998). This implies that we should expect our cutoffs of interest to be "fuzzy" rather than "sharp" discontinuities (Trochim, 1984): that is, we do not expect the probability of a given treatment to fall from 1 to 0 as one moves from 1499 grams to 1501 grams, but rather expect a change in the likelihood of treatment for newborns classified into a given risk category.

¿From an empirical perspective, the fact that we observe a first stage in terms of a discontinuity in treatment provision around 1500 grams suggests that hospitals or physicians do use these cutoffs to determine treatment either through hospital protocols or as rules of thumb.[10] As one example, the 1500 gram threshold is commonly cited as a point below which diagnostic ultrasounds should be used. Diagnostic ultrasounds (also known as cranial ultrasounds) are used to check for bleeding or swelling of the brain as signs of intraventricular hemorrhages (IVH) - a major concern for at-risk newborns. The neonatal care manual used by medical staff at the Longwood Medical Area (Boston, MA) notes: "We perform routine ultrasound screens in infants with birth weight <1500gm (Cloherty & Stark, 1998)." We will investigate differences in procedure use, including diagnostic ultrasounds, below.

Hospital protocols could also potentially be informed by Current Procedural Terminology (CPT) billing codes and ICD-9 diagnosis codes that are categorized by birth weight (ICD-9

[10] As we will discuss more in Section IV, we may not observe all relevant first stage inputs. That said, we do find evidence of a first stage for summary measures of treatment, as well as for some particular procedures.

codes V21.30-V21.35 denote birth weights of 0-500, 500-999, 1000-1499, 1500-1999, 2000-2500, *etc.*). Note that if prices differ across our threshold of interest, then any discontinuous jump in charges could in part be due to changes in prices rather than changes in quantities. In practice, we argue that a substantial portion of our observed jump in charges is a "quantity" effect rather than a "price" effect, for two reasons. First, we find evidence of discontinuities in quantities, such as length of stay, implying that changes in prices do not explain our entire measured discontinuities in charges. Second, the limited evidence available to us suggests that prices do not vary discontinuously across the VLBW for many of the births in our data. We unfortunately do not observe prices directly in any of our hospital discharge record data sets. Information on the payment methods used to reimburse hospitals is notoriously complex and frequently incomplete, and the most systematic information available is for hospitalizations covered by public insurers such as Medicare and Medicaid. Medicaid is relevant for our purposes, since as noted above the latter covers a large share of deliveries nationally. A recent study of Medicaid payment systems (Quinn, 2008) found that although some states rely on payment systems that explicitly incorporate birth weight categories into the reimbursement schedules, most states - including California - rely on systems which do *not* explicitly utilize birth weight.[11] Since we empirically observe a first stage in California data, this (combined with the length of stay results) suggests that a substantial portion of our observed jump in charges is a "quantity" effect rather than a "price" effect.

Another explanation for why these cutoffs may affect treatment provision is that they may be used as informal "rules of thumb" by physicians.[12] Discussions with physicians suggest that these potential discontinuities are well-known, salient cutoffs below which newborns may be at increased consideration for receiving additional treatments. Clinicians (as well as researchers) frequently collapse continuous measures of health (such as birth weight) into binary measures such as "low birth weight" versus "normal birth weight." Some have argued that such heuristics

[11] More precisely, because birth weight is thought to be the best predictor of neonatal resource use (Lichtig *et al.*, 1989), some newer DRG-based (that is, Diagnosis Related Group) systems explicitly incorporate birth weight categories into the reimbursement schedules. However in an analysis of the use of various types of DRG systems in state Medicaid programs, Quinn (2008) finds that most states rely on either a *per diem* system or the CMS-DRG system, neither of which explicitly utilize birth weight. Almost half of states use the CMS-DRG system (CO, IA, IL, KS, KY, MI, MN, MT, NC, ND, NE, NH, NJ, NM, OH, OR, PA, SC, SD, TX, UT, WI, WV) and twelve states use a per diem system (AK, AZ, CA, FL, HI, LA, MO, MS, NV, OK, TN, VT) (Quinn, 2008).

[12] Medical malpractice environments could also be one force affecting adherence to either formal rules or informal rules of thumb.

can influence medical treatment by acting as "silent adjudicators of clinical practice" (McDonald, 1996). Others have noted that rules of thumb in medicine can be "useful and even necessary shortcuts guiding search and choice under uncertainty and time constraint" (Andre *et al.*, 2002).

III. Data

A. Data description

Our empirical analysis requires data with information on birth weight and some welfare-relevant outcome, such as medical care expenditures or health outcomes. Our primary analysis uses three data sets: first, the National Center for Health Statistics (NCHS) birth cohort linked birth/infant death files; second, a longitudinal research database of linked birth record-death certificate-hospital discharge data from California; and third, hospital discharge data from several states in the Healthcare Cost and Utilization Project (HCUP) state inpatient databases.

The NCHS birth cohort linked birth/infant death files, hereafter the "nationwide data," include data for a complete census of births occurring each year in the US, for the years 1983-1991 and 1995-2002 - approximately 66 million births.[13] The data include information reported on birth certificates linked to information reported on death certificates for infants who die within one year of birth. The birth certificate data offers a rich set of covariates (for example, mother's age and education), and the death certificate data includes a cause of death code. Beginning in 1989, these data include some treatment variables - namely, indicators for use of a ventilator for less than or (separately) greater than thirty minutes after birth.

Our other two data sources offer treatment variables beyond ventilator use. The California research database is the same data set used in Almond & Doyle (2008). These data were created by the California Office of Statewide Health Planning and Development, and include all live births in California from 1991-2002 - approximately 6 million births. The data include hospital discharge records linked to birth and death certificates for infants who die within one year of birth. The hospital discharge data include diagnoses, course of treatment, length of hospital stay, and charges incurred during the hospitalization. The data are longitudinal in nature and track hospital readmissions for up to one year from birth as long as the infants enter

[13] NCHS did not produce linked birth/infant death files from 1992-1994.

a California hospital. This longitudinal aspect of the data allows us to examine charges and length of stay even if the newborn is transferred to another hospital.[14]

The HCUP state inpatient databases allow us to analyze the universe of hospital discharge abstracts in four other states that include the birth weight variable necessary for our analysis.[15] Specifically, we use HCUP data from Arizona for 2001-2006, New Jersey for 1995-2006, Maryland for 1995-2006, and New York for 1995-2000 - approximately 10.5 million births (see Appendix Table A3 for the number of births by state and year within our pilot bandwidth of 3 ounces of the VLBW cutoff).[16] The HCUP data include variables similar to those available in the California discharge data, but unlike the California data are not linked to mortality records nor to hospital records for readmissions or transfers. Although we cannot link these discharge data with mortality data directly, we can examine mortality outcomes for these newborns using a sub-sample of our nationwide data, as our nationwide data and the HCUP discharge data relate to the same births.[17] In much of our analysis, we pool the California and HCUP data to create a "five-state sample."

B. Analysis sample

Sample selection issues are minimal. In our main specifications, we pool data from all available years, although we do separately examine results across time periods. For the main results, we limit the sample to those observations with non-missing, non-imputed birth weight information.[18] Fortunately, given the demands of our empirical approach, these data provide relatively large samples: over 200,000 newborns fall within our pilot bandwidth of 3 ounces around the

[14]The treatment measures that include transfers described below include treatment at the hospital where the newborn was initially transferred.

[15]The State Inpatient Data (SID) we analyze contain the universe of inpatient discharge records from participating states. (Other HCUP databases, such as the National Inpatient Sample, are a sub-sample of the SID data.) At present, 39 states participate in the SID. Of these 39 states, 10 report the birth weight of newborns. We have obtained HCUP data for four of the ten states with birth weight (with the exception of North Carolina, we have discharge data for the top four states by number of births: New York, New Jersey, Maryland, and Arizona).

[16]In ongoing work, we intend to supplement our analysis with additional years of HCUP data from these four states.

[17]Note that our nationwide data include births that took place outside of hospitals, whereas our California and HCUP discharge data by construction only capture deliveries taking place in hospitals. In practice, 99.2% of deliveries in our national sample occurred in a hospital. In some specifications we limit our nationwide data to the sample of hospitalized births, for greater comparability.

[18]In practice this sample selection criteria excludes a very small number of our observations. For the full NCHS data, for example, dropping observations with missing or imputed birth weights drops only 0.12% of the sample. We also exclude a very small number of observations in early years of our data that lack information on the time of death.

1500 gram threshold in the nationwide data, and we have approximately 30,000 births in the same interval when we consider the hospital discharge data. We discuss bandwidth selection below.

IV. Empirical framework and estimation

A. Empirical framework

Consider the following structural equation for the effect of an input I (for example, charges for medical treatments) on an outcome variable Y (for example, one-year mortality):

$$Y = f(Z, I, e), \qquad Z \in [Z^* - h, Z^* + h] \tag{1}$$

where Z is the running variable (birth weight), e is unobserved heterogeneity, and h is a bandwidth around Z^*. As can be seen by comparing Figure 2 and Figure 3, higher mortality is empirically associated with higher charges, reflecting differences in underlying health. We aim to overcome the confounding influence of underlying health with our regression discontinuity design, which we introduce in an instrumental variables framework. The instrument Z^* is an indicator that divides observations according to a threshold in the running variable. In our context, we define the instrumental variable Z^* using the VLBW threshold at 1500 grams as follows:

$$Z^* = \begin{cases} 1 & \text{if VLBW (birth weight} < 1500 \text{ grams)} \\ 0 & \text{if not VLBW (birth weight} \geq 1500 \text{ grams)} \end{cases}$$

Our first stage equation can then be written as:

$$I = g(Z^*, Z, v), \qquad Z \in [Z^* - h, Z^* + h] \tag{2}$$

where v is unobserved heterogeneity. In order for Z^* to be a valid instrument, the two usual instrumental variables conditions must hold. First, there must exist a first stage relationship

between Z^* and I; note that this relationship will be conditional on our running variable Z. Second, the exclusion restriction requires that the only mechanism through which the instrument Z^* affects our outcome variable (here, mortality), conditional on Z falling within the bandwidth, is through its effect on the input I. If we observed and were able to measure all relevant inputs in I, then we could argue for the validity of this exclusion restriction. However, for any given I that we observe, it is likely that there exists some additional health-related input that we do not observe.[19] It is unclear how important such unobserved inputs are in practice, but to the extent they are important, a first stage variable I such as medical expenditures would violate the exclusion restriction.

On one hand, our reduced form estimate of the direct impact of our instrumental variable Z^* on mortality is itself interesting and policy relevant, as this estimate includes the effects of all relevant inputs. On the other hand, our instrumental variable estimate is also of substantive interest, and thus having a sense of the potential magnitude of the effects of unobserved inputs is useful. To the extent that medical inputs are much more important relative to parental inputs in the very short run after birth (say, within twenty-four hours of birth), we can test for impacts on short run mortality measures and be somewhat assured that other unobserved parental inputs are not likely to affect these estimates. As we will discuss in Section V, we do indeed find effects on short run mortality measures.

We examine a number of different first stage variables, which we discuss at length below. Broadly, we examine first stage variables in two categories: "summary treatment measures" (such as charges and length of stay) that capture many aspects of hospital treatment, and "mechanism variables" (such as ventilation) that could be the margins through which discontinuities in summary treatments occur.

In Section V, we report first stage and reduced form estimates separately. In Section IX, we combine these estimates into two-sample estimates in which the numerator is the reduced form estimate and the denominator is the first stage estimate.[20]

[19] As an example, consider the case in which the length of time parents hold their newborns has a direct effect on mortality, that this variable is unobserved in the hospital claims data we study, and that this variable varies discontinuously across our cutoff (which could be the case if the 1500 gram classification is salient to parents).

[20] Without covariates, the two-sample estimate is equivalent to the Wald and two stage least squares estimates, given our binary instrumental variable. Even though the first stage and reduced form estimates come from different data sources, we can standardize the samples and covariates to produce the same estimates that we would attain from a single data source.

B. Estimation

To estimate the size of the discontinuity in outcomes and treatment, we follow standard methods for regression discontinuity analysis (as in, for example, Imbens & Lemieux (2008)). First, we estimate and report a local-linear regression. This estimate incorporates information from a bandwidth of 3 ounces (85 grams) above and below the threshold. We describe the selection of this bandwidth in the next subsection. We use a triangle kernel so that the weight on each observation decays with the distance from the threshold, and we report asymptotic standard errors (Cheng *et al.*, 1997; Porter, 2003).[21]

In addition, within the bandwidth h, we estimate the following model for infant i weighing g grams in year t:

$$Y_i = \alpha_0 + \alpha_1 VLBW_i + \alpha_2 VLBW_i * (g_i - 1500) + \alpha_3 (1 - VLBW_i) * (g_i - 1500) + \alpha_t + \alpha_s + X_i'\delta + \epsilon_i \quad (3)$$

where Y is an outcome such as one-year mortality, and $VLBW$ is an indicator that the newborn was classified as very low birth weight (that is, strictly less than 1500 grams). We include separate gram trend terms above and below the cutoff, parameterized so that a test of whether the trend is the same above and below the threshold is simply a matter of testing whether $\alpha_2 = \alpha_3$. In some specifications, we include indicators for each year of birth t, indicators for each state of birth s, and newborn characteristics, X_i'. The newborn characteristics that are available for all of the years in the nationwide data include an indicator that the mother was born outside the state where the infant was born, as well as indicators for mother's age, education, father's age, the newborn's sex, gestational age, race, and plurality. We estimate this model by OLS with heteroskedastic-robust standard errors. Probit results for our binary dependent variables give very similar results, as described below.

C. Bandwidth selection

Our pilot bandwidth includes newborns with birth weights within 3 ounces (85 grams) of 1500 grams, or from 1415 grams to 1585 grams. We chose this bandwidth by a cross-validation procedure where the relationships between the main outcomes of interest and birth weight were

[21] We are grateful to Doug Miller for providing code from Ludwig & Miller (2007).

estimated with local linear regressions and compared to a 4-th order polynomial model. These models were estimated separately above and below the 1500 gram threshold. The bandwidth that minimized the sum of squared errors between these two estimates between 1200 and 1800 grams tended to be between 60 and 70 grams for the mortality outcomes. For the treatment measures, the bandwidth tended to be closer to 40 grams. Given that we are estimating the relationship at a boundary, a larger bandwidth is generally warranted. We chose to use a pilot bandwidth of 85 grams - 3 ounces[22] - for the main results. This larger bandwidth incorporates more information that can improve precision, but of course including births further from the threshold departs from the assumption that newborns are nearly identical on either side of the cutoff. That said, our local linear estimates allow the weight on observations to decay with the distance from the threshold. In addition, the results are qualitatively similar across a wide range of bandwidths (see Table 6). To give a clearer sense of how our data look graphically, our figures report means for a slightly wider bandwidth - namely, the 5 ounces above and below.

V. Results

A. Frequency of births by birth weight

Figure 1 reports a histogram of births between 1350 grams and 1650 grams in the nationwide sample, which has several notable characteristics.[23] First, there are pronounced reporting heaps at the gram equivalents of ounce intervals. Although there are also reporting heaps at "round" gram numbers (such as multiples of one hundred), these heaps are much smaller than those observed at gram equivalents of ounce intervals. Discussions with physicians suggest that birth weight is frequently measured in ounces, although typically also measured in grams as well for purposes of billing and treatment recommendations. Given the nature of the variation inherent in the reporting of our birth weight variable, our graphical results will focus on data which has been collapsed into one-ounce bins.[24]

Second, we do not observe irregular reporting heaps around our 1500 gram threshold of

[22] As discussed in the next section, our birth weight variable has pronounced reporting heaps at gram equivalents of ounce intervals. We specify the bandwidth in ounces to ensure that the sample sizes are comparable above and below the discontinuity, given these trends in reporting.

[23] See Appendix Figure A1 for a wider set of births.

[24] Specifically, we construct one-ounce bins radiating out from our threshold of interest (*e.g.* 0-28 grams from the threshold, 29-56 grams from the threshold, *etc.*).

interest, consistent with women being unable to predict birth weight in advance of birth with the accuracy necessary to move their newborn (via birth timing) from just above 1500 grams to just below 1500 grams. The lack of heaping also suggests that physicians or hospitals do not manipulate reported birth weight so that, for example, more newborns fall below the 1500g cutoff and justify higher reimbursements. In particular, the frequency of births at 1500 grams is very similar to the frequency of births at 1400 grams and at 1600 grams, and the ounce markers surrounding 1500 grams have similar frequencies to other ounce markers.

It is worth noting that more complicated manipulations could in theory be consistent with Figure 1. For example, if doctors re-label unobservably sicker newborns weighing just above 1500 grams to be labeled as being below 1500 grams (to, for example, receive additional treatments) and symmetrically "switched" the same number of other newborns weighing just below 1500 grams to be labeled as being above 1500 grams, this could be consistent with the smooth distribution in Figure 1. This seems unlikely, particularly given that we will later show that other covariates (such as gestational age) are smooth across our 1500 gram cutoff - implying that doctors would need to not only "symmetrically switch" newborns but symmetrically switch newborns who are identical on all of the covariates we observe. The assumption that such switching does not occur is an assumption we argue is plausible.[25]

More formally, McCrary (2008) suggests a direct test for possible manipulation of the running variable - in our case, birth weight. We implement his test by collapsing our nationwide data to the gram level - keeping a count of the number of newborns classified at each gram - and then regressing that count as the outcome variable in the same framework as our regression discontinuity estimates. Using this test, we find no evidence of manipulation of the running variable around the VLBW threshold.[26]

[25] Note that to the extent that hospitals or physicians may have a larger incentive to categorize relatively costly newborns as VLBW to justify greater charge amounts, such gaming would tend to lead to higher mortality rates just prior to the threshold, contrary to our main findings.

[26] Specifically, for 1500 grams we estimate a coefficient of -2,100 ($s.e.$=1500, p=0.1898). This test is useful under the assumption that the distribution of births would be smooth in the absence of the incentives created by the VLBW cutoff. An alternative test for gaming that does not rely on this assumption is to test whether newborns look similar on observable variables above and below the cutoff, which we examine in Figure 5.

B. Health outcomes

Figure 2 reports mean mortality for all infants in one-ounce bins close to the VLBW threshold. Note that the one-year mortality rate is relatively high for this at-risk population: close to 6%. The figure shows that even within our relatively small bandwidth, there is a general reduction in mortality as birth weight increases, reflecting the health benefits associated with higher birth weight. The increase in mortality observed just above 1500 grams appears to be a level shift, with the slope generally similar above and below the threshold.[27] The mean mortality rate in the ounce bin just above the threshold is 6.15%, which is 0.46 percentage points larger than mean mortality just below the threshold of 5.69%. We see a similar .48 percentage point difference for 28-day mortality - between 4.39% above the threshold and 3.91% below the threshold. This suggests that most of the observed gains in 28-day mortality persist to one year.

Table 1 reports the main results. The first reported outcome is one-year mortality, and the local-linear regression estimate is -0.0121. This implies a 22% reduction in mortality compared to a mean mortality rate of 5.53% in the 3 ounces above the threshold (the "untreated" group in this regression discontinuity design). OLS point estimates are slightly smaller, but still large: -0.0095. The estimates are similar when probit models are used as well.[28]

For our OLS models, we report heteroskedastic-robust standard errors. To address potential concerns about discreteness in our dependent variable, we perform the standard error correction suggested by Card & Lee (2008). In our application, this correction amounts to clustering at the gram-level. Doing so increases the magnitude of our standard errors, but our main results remain statistically significant at conventional levels.[29]

The gram-trend terms reflect the overall downward slope in mortality. The point estimates

[27] Note that in this graph there is also a smaller, visible "jump" in mortality around 1600 grams, an issue we address in several ways. First, if we construct graphs analogous to Figure 2 which focus on 1600 grams as a potential discontinuity, there is no visible jump at 1600 grams. Exploration of this issue reveals that the slightly different groupings which occur when one-ounce bins are radiated out from 1500 grams relative to when one-ounce bins are radiated out from 1600 grams explain this difference, implying that small-sample variation is producing this visible "jump" at 1600 grams in Figure 2. Reassuringly, the "jump" at 1500 grams is also visible in the graph which radiates one-ounce bins from 1600 grams, suggesting that small-sample variation is not driving the visible discontinuity at 1500 grams. Finally, when we estimate a discontinuity in a formal regression framework at 1600 grams we find no evidence of either a first stage or a reduced form effect at 1600 grams.

[28] A probit model predicts a difference of -0.0095 in a model with no controls other than the trend terms (evaluated at the cutoff), and -0.0065 in a model with full controls evaluated at the cutoff and the mean of the other control variables.

[29] For example, our OLS estimate of -0.0095 ($s.e.$=0.0022) for one-year mortality has a standard error of 0.0048 when we cluster at the gram level.

suggest a steeper slope after the threshold.[30] In terms of the covariates, the largest impact on our main coefficient of interest is found when we introduce year indicators, likely because medical treatments and associated survival rates have changed so much over time. The estimated change in mortality around the threshold in the specification with the year indicators decreases to -0.0076. Again, this remains a large effect, and we also consider heterogeneity across time periods below. When we include the full set of covariates, the results are largely unchanged.[31] To be conservative, in the rest of our analysis, we always report a specification that includes the full set of covariates.

The remaining outcomes in Table 1 are mortality measures at shorter time intervals. The 28-day mortality coefficient is similar in magnitude to the one-year mortality coefficient, despite a smaller mean mortality rate of .0383%. Given different mean mortality rates, the estimate implies a 23% reduction in 28-day mortality as compared to a 17% reduction in one-year mortality. As discussed above, the similarity between the one-year and 28-day mortality rates implies that any effects of being categorized as very-low birth weight are seen in the first month of life - a time when these infants are largely receiving medical care (as described more below in our length of stay results). Similarly, 7-day and 24-hour mortality rates are lower for those who are very low birth weight compared to those just above the threshold, with differences of between 16% and 19% compared to the mean mortality rate for infants above the threshold. Finally, 1-hour mortality rates (not shown) are also higher for those born just above the threshold.[32]

The following two subsections consider the extent to which newborns classified as very low birth weight receive discontinuously more medical treatments relative to newborns just above 1500 grams. While the universe of births in the natality file allows us to consider mortality effects with a large sample, these data do not include summary measures of treatment. As described above, we are able to examine summary measures of treatment in our hospital discharge data

[30] In the OLS specification estimating a treatment effect of -0.0095 in Table 1, for example, we reject the hypothesis that the two slope coefficients are the same at the 5% level; the p-value of the test is .0435.

[31] The estimated coefficients on many of these covariates as well as one of the year indicators are reported in Appendix Table A1. Characteristics that are associated with lower mortality include prenatal visits, mother's age, and African American. Higher mortality is associated with mother's first birth, males, longer gestational age (where greater age in this bandwidth suggests that the newborn is particularly small for gestational age), and singleton births (where the low birth weight in this bandwidth is not explained by the presence of multiple births). We only show a sub-set of the coefficients on these covariates in order to keep the table to one page.

[32] In a probit model with full controls, the main marginal effect of interest, evaluated at the cutoff and the mean of the other control variables, is -0.0012 (s.e.=0.0005) compared to a mean 1-hour mortality rate of 0.0055 just above the threshold.

from five states (Arizona, California, Maryland, New Jersey, and New York). These states appear to have broadly representative mortality outcomes. When we estimate our mortality results separately within each state and rank them by the estimated coefficient scaled by mean mortality just above the threshold, each of the states in our 5-state sample falls toward the middle of the distribution.[33]

Our mortality results are similar when limit our nationwide data to only these five states. On this sample of nearly 50,000 births, we estimate that mortality falls by 1.2 percentage points ($s.e.$=0.42) compared to a mean of 5.4% (as reported in Table 5).[34]

C. Differences in summary measures of treatment

Figure 3A reports mean hospital charges in one-ounce bins. The measure appears fairly flat at $94,000 for the three ounces prior to the threshold, then falls discontinuously to $85,000 after the threshold and continues on a downward trend, consistent with fewer problems among relatively heavier newborns. This flattening before the threshold is suggestive that newborns who are up to three ounces from the threshold may receive additional treatment due to the VLBW categorization.

As the large mean charges suggest, this measure is right skewed. The results are similar, however, when we estimate the relationship using median comparisons and when the charges are transformed by the natural logarithm to place less weight on large charge amounts, as described below.[35]

As noted in Section IIC, if prices differ across our threshold of interest, then any discontinuous jump in charges could in part be due to changes in prices rather than changes in quantities. One way to test whether differences in quantities of care are driving the main results is to consider a

[33]Specifically, we estimate separate models with no control variables in all available years in the nationwide data in each of the 50 states and the District of Columbia. With the most negative value of the estimated coefficient scaled by mean mortality just above the threshold ranked as 1, Arizona ranks 25 (-0.0101/.0618), California ranks 21 (-.0185/.0682), Maryland ranks 19 (-.0148/.0511), New Jersey ranks 16 (-.0153/.0423), and New York ranks 15 (-.0208/.0569). The scaled coefficients mitigate differences in the underlying mean mortality just above the threshold: in a separate ranking of states by mean mortality, New Jersey ranks the lowest of any state, but Arizona, California, New Jersey, and New York are more to the middle of the distribution, ranking 26, 30, 7, and 13 respectively.

[34]The mortality outcomes were also considered in these five states from 1991-2002 - the overlap of the years between the state data and the nationwide data. As expected, the results are imprecisely estimated with the smaller sample, and the point estimates are lower as well.

[35]Note that our sample sizes vary somewhat when looking at charges variables in levels or in logs due to observations with missing or zero charges. Graphing the mean probability that charges are missing or zero across 1500 grams does not reveal a discontinuous change across this threshold.

quantity measure that is consistently measured across hospitals: length of stay in the hospital. Figure 3B shows that average length of stay drops from just over 27.3 days immediately prior to the threshold to 25.7 days immediately after the threshold. Of course, length of stay and charges are not independent measures, as longer stays accrue higher charges both in terms of room charges and are associated with a greater number of services provided. We further investigate the differences in such service provision measures below.

As in the reduced form results for infant mortality, we compare treatment measures above and below the threshold using a local linear regression and a bandwidth of 3 ounces (85 grams). In addition, we estimate OLS models following equation 3, where the outcome, Y, now represents hospital charges or length of stay. Figure 3 shows that the linear trend controls within 3 ounces above and below the threshold appear reasonable.[36] Again, results are similar when we estimate alternative models, such as count models for length of stay.[37]

Table 2 reports the regression results. The first column reports estimates from the local linear regression and hospital charges are \$9,450 higher just before the threshold. This is relatively large compared to the mean charges of \$82,000 above the threshold. The remaining columns report the OLS results. The first column mimics the local linear regression but now places equal weight on the observations up to 3 ounces on either side of the threshold. Without controls, the estimate decreases somewhat to \$9,021; with full controls the estimated increase in charges for infants categorized as very low birth weight is largely unchanged (\$9065, s.e.=\$2,297). These estimates imply a difference of approximately 11% compared to the charges accrued by infants above the threshold.

We report the coefficients on the controls in Appendix Table A1. There are fewer controls in the 5-state sample than there are in the nationwide sample, as the discharge data do not include the birth certificate data. Results are qualitatively similar in a separate analysis of California, which allows for a wider set of controls from the linked birth certificate data. The coefficients suggest that preterm births, multiple births, and cesarean sections accrue higher charges.

In related models of length of stay, we find that newborns weighing just under 1500 grams

[36] Estimation of our first stage and reduced form results with quadratic rather than linear trends in birth weight gave similar results, with the estimates from linear trend specifications most often being slightly smaller in magnitude and hence more conservative than the estimates from quadratic trend specifications.

[37] The estimated incidence-rate ratio for the indicator of being just below the threshold is 1.076 (s.e.=0.018) in a model with year indicators and 1.068 (s.e.=0.18) in a model with full controls.

have stay lengths that are between 1.5 and 2 days longer, depending on the model.[38] This represents a difference of 6-8% compared to the mean length of stay of 25 days above the threshold.

Note that our first stage variables could be censored in two ways. First, newborns just above 1500 grams have higher mortality rates (as shown above) and thus mechanically may receive less medical treatment. This effect would make our first stage effect look less conservative, but would make the cost per year of life saved estimates we derive below more conservative (since larger first stage effects will make our marginal return estimates look less cost effective). Further, it appears that the length of stay differences are too large to be explained entirely by such censoring.[39]

A second potential source of censoring in our 5-state sample is the possibility that newborns just below the cutoff were transferred to another hospital. In the HCUP data we do not observe charges across hospital transfers, implying our first stage measures could be censored. In the California data, however, we can observe hospital transfers and our results are similar when such transfers are included in the treatment measures in California - as shown in Table 4.[40] Further, those born just below the threshold are slightly less likely to experience a transfer to another hospital (Table 4). This difference in transfers would tend to understate the treatment of newborns just below the threshold and bias the results away from finding an increase in care for these infants.

To ensure that our results are not driven by a small number of newborns with large charge amounts, we investigated median differences in charges as well as the natural logarithm of charges, which places less weight on large amounts and greater weight on smaller ones. Appendix Figure A2 reports these measures separately for the five-state sample and California - a state where longitudinal data allows calculations of treatment measures across hospitals if the newborn were transferred. The relationship is similar to the mean comparisons when median charges are

[38] Note that we define our length of stay variable such that the smallest value is 1 - a value of 2 indicates that the stay continued beyond the first day, and so forth. This definition allows us to include observations in our log length of stay variable that are less than one full day.

[39] If the length of stay difference of 2 days was driven entirely by the 1 percentage point reduction in mortality, then the uncensored length of stay for the newborns in question would have to be 2/0.01 or 200 days long. While not impossible, only 11 newborns within our bandwidth have stays greater than 200 days.

[40] When the dependent variable is length of stay including transfers, the main coefficient of interest is 1.36 (s.e.=0.561) in a model with year indicators and 1.24 (s.e.=0.553) in a model with full controls, compared to a mean of 27.0 for those born just above the threshold.

considered: the ounce just below the threshold has median charges of $68,000 and the ounce just after the threshold has median charges of $59,000, implying a jump of $9,000, or 15% of median charges above the threshold. Further, Table 4 shows results from a median regression with full controls, and median charges are found to be $9,415 lower just after the threshold.

Figure A2B reports the means for log charges, which are relatively noisy around 10.6 log points from 5 ounces above the threshold, with an increase just prior to the threshold. The upward slope is largely driven by newborns with few charges.[41] Further, when data from hospitals where newborns were transferred are included using the longitudinal data from California (Appendix Figure A2D), log charges are relatively flat at 11.17 in the one-ounce bins just before the threshold and drop to 11.00 after the threshold. Appendix Figures A2E and A2F report similar estimates to the main results when median log charges are compared in both the 5-state sample and in the California sample (differences of 14 and 12 log points, respectively).

Table 4 reports results when the dependent variable is log charges. When we impose a linear trend in grams, charges are found to be 28 log points higher just before the threshold. If we do not extrapolate using a linear trend, as the upward trend prior to the threshold within our bandwidth likely reflects noise rather than a true increase, then a difference between 10.77 log points just prior to the threshold and a return to 10.6 just after the threshold would suggest an alternative estimate of 17 log points. Further, when we use the longitudinal data from California to calculate total charges for newborn treatment including charges associated with hospital transfers, and the upward trend prior to the threshold is not present, the estimate is 22 log points (Table 4). Meanwhile, when log length of stay is the dependent variable, newborns weighing just below the threshold are found to stay 14 log points fewer days in care, and 11 log points when length of stay including transfers is compared in the California data.

An alternative measure to hospital charges is a measure of hospital costs.[42] This measure

[41] The upward slope disappears when the sample is restricted to newborns with greater the $3,000 in charges. A plot of an indicator that the newborn accrued charges of less than $3,000 against birth weight revealed a fairly noisy series. An OLS model with the standard controls suggests that such low charge amounts are less likely for newborns just below the threshold, with an estimated coefficient on birth weight less than 1500g of -0.042 (s.e. = 0.0065), compared to a mean above the threshold of 0.093.

[42] The Centers for Medicare and Medicaid Services (CMS) report cost-to-charge ratios for each hospital in each year beginning in 1996 and the data are available through 2005. Our charge data are available from 1991-2006 (see Appendix Table A3). To include the information from all of these years, the 2000 cost-to-charge ratios were used to deflate charges in all states but New York where the first year of data is 2001 and the 2001 cost-to-charge ratio is used. Further, we followed a CMS suggestion to replace the hospital's cost-to-charge ratio with the state median if the cost-to-charge ratio is beyond the 5th or 95th percentile of the state's distribution. Results were similar, though noisier, when the sample was restricted to 1996-2005 and each hospital-year cost-to-charge ratio

is known to introduce noise into the results, as the hospital-level cost to charge ratios are a rough measure of the difference between costs and charges and the ratios are volatile over time within hospitals. That said, the estimated cost is closer to the marginal cost of interest. While these costs may not represent social costs for such care - the nurses, physicians, and capital expenditures may not be affected by the births of a small number of very low birth weight infants - they represent our best summary measurement of the difference in treatment that VLBW classification affords. For this reason, we use costs rather than charges in our interpretation of the social costs of medical treatment for these newborns, as in Section IX below.

The figures for hospital costs in levels and in logs are similar to the main results, with a trend that is somewhat more steep after the threshold when levels are considered. As reported in Table 4, the cost measure is $4,200 higher just before the threshold in a model without controls, and $3,800 higher in a model with controls, compared to an average cost of nearly $40,000 just above the threshold, or approximately 10%. A median regression (not shown) suggests a difference of $4,800 compared to a median level of $30,100 within 3 ounces above the threshold: a 16% difference.

In summary, we find differences in summary treatment measures of approximately 10-15% with some variation in the estimate depending on the treatment measure. In terms of charges, the difference across the discontinuity is approximately $9,500. When we deflate charges by a cost-to-charge ratio, this difference is closer to $4,000.

D. Mechanisms: Differences in types of care

The discharge data include procedure codes that can be used to investigate the types of care that differ for infants on either side of the very low birth weight threshold. We explore the data for such differences, with a special focus on common perinatal procedures. Like the mortality differences, however, the smaller 5-state sample tends to make such differences difficult to find. Table 3 and Figure 4 present differences for measures of common procedures.

One of the most common procedures is some form of ventilation.[43] Figure 4A does not offer

was employed.

[43] We observe several measures of assisted ventilation, including continuous positive airway pressure (CPAP) ventilation, a procedure which can be thought of as less serious than a traditional ventilator; intubation, which involves an endotrachial breathing tube being inserted through a newborn's mouth into her lungs; and several

compelling evidence of a meaningful difference in ventilation by birth weight, however. Further, the nationwide data include ventilation measures, so sample size considerations do not play a role, but in those data we still find little evidence of an increase in ventilation among these newborns.

Another common measure of resource utilization that *a priori* seems to be a likely source of the differences we observe in our summary treatment measures is admission to a neonatal intensive care unit (NICU). Since care provided in such units is costly, it seems plausible that the threshold could be used to gain entry into such a unit. We find little difference on this margin in our data, however. First, we examine the California data, which includes a variable on whether or not the infant spent at least 24 hours in a NICU or died in the NICU in less than 24 hours. We include newborns born in hospitals that did not have a NICU for comparability to our main results, which also include such newborns. Table 3 suggests a modest increase in NICU use (approximately 3 percentage points as compared to a mean just above the threshold of 44 percentage points), but Figure 4B shows little evidence of a discontinuous change. Second, we examine the Maryland HCUP data, which records the number of days in the NICU, but again we find little evidence of a difference at the threshold.[44] Our results are consistent with a study of NICU referrals, in which very low birth weight was not listed among the common reasons for triage to a NICU.[45]

Perhaps the most compelling differences we find are for two relatively common procedures: diagnostic ultrasound of the infant and operations on the heart. As noted above, diagnostic ultrasounds are used to check for bleeding or swelling of the brain and some physician manuals cite 1500 grams as a threshold below which diagnostic ultrasounds are suggested. Figure 4C suggests a jump in ultrasounds of roughly 2 percentage points compared to a mean of approximately 25%. Table 3 suggests a similar estimate of between 2 and 3 percentage points.

other measures of ventilator use. Within these subcategories, we found little support for any discontinuous change. Some oxygen may be provided before birth weight is measured, although to the best of our knowledge we are not able to separate this from ventilation provided after birth weight is measured in our data.

[44]The New Jersey HCUP data include a field for NICU charges, but this variable proves unreliable: the fraction of newborns with non-missing NICU charges for this at-risk population is only 2%. Recent nationwide birth certificate data include an indicator for NICU admission for a handful of states. We do not see a visible discontinuity in these data, albeit potentially due to the small sample of births in the years for which we observe this variable.

[45]The most common was birth complications, followed by transitional respiratory distress, hyperbilirubinemia, prematurity and postmaturity, congenital anomalies, and "small for gestational age," or SGA (Zupancic & Richardson, 1998). Interestingly, Zupancic & Richardson (1998) concluded that "little of triage care has a strong base of evidence, potentially leaving more discretion – and thus variability – in diagnosis and management."

The pattern of the "operations on the heart" indicator shows an upward pre-trend in the procedures prior to the threshold and what appears to be a discontinuous drop after the threshold. Table 3 suggests that the jump is between 1.5 and 2.4 percentage points, or roughly 8% higher than the mean rate for those born above the threshold in this sample. If we do not extrapolate the pre-trend, Figure 4D shows a cardiac surgery rate of 30.0% prior to the threshold and 26.7% after the threshold, a difference of 3.3 percentage points.

In summary, we examine several possible treatment mechanisms at the discontinuity. The strongest evidence that we find is for operations on the heart and diagnostic ultrasounds, for which we estimate an approximate 10% increase in usage just prior to the VLBW threshold. We show that these differences are statistically significant when the usual heteroskedastic-robust standard errors are used for comparability to the main results, but these standard errors could be corrected to account for the search across procedures. When we apply a Bonferroni correction for the different types of procedures, our estimates remain statistically significant.[46] We find little evidence of differences in NICU usage or other common procedures such as ventilation, however, including an examination of ventilation in the nationwide data with much larger sample sizes.

VI. Robustness & specification checks

In this section, we test for evidence of differences in covariates across our VLBW threshold (sub-section A), discuss potential sample selection issues (sub-section B), discuss the sensitivity of our results to alternative bandwidths (sub-section C), and examine our mortality results by cause of death (sub-section D).

[46] To consider differences in potentially costly care, we searched for differences in procedures used to define NICU quality levels in California (Phibbs *et al.*, 2007). Excluding NICU measures, which we examined separately, these included two summary measures of cardiac care, seven different measures of ventilation procedures, and diagnostic ultrasound. The p-values on diagnostic ultrasound and cardiac care using the reported standard errors are 0.002 and 0.018, respectively. Using the conservative Bonferroni correction by multiplying these p-values by 10 implies that the estimates are significant at the 2% and 18% levels; multiplying these p-values by three for the three categories of exploration would imply that the estimates are significant at the 1% and 4% levels. A separate exploration considered an additional 5 categories of procedures that were among the top-25 most common primary and secondary procedures in our data: injection of medicines, excision of tissue, repair of hernia, and two additional diagnostic procedures. Again, we found little visual evidence of a change in these procedures at the 1500 gram cutoff.

A. Testing for evidence of differences in covariates across 1500 grams

As discussed above, it is thought that birth weight cannot be predicted in advance of birth with the accuracy needed to change (via birth timing) the classification of a newborn from being just above 1500 grams to being just below 1500 grams. As such, we expect that the newborns will be similar above and below the threshold in both observable and unobservable characteristics. Moreover, as discussed in Section VA, most forms of strategic recategorization of newborns based on birth weight around 1500 grams should be detectable in our histograms of birth frequencies by gram of birth weight. That said, it is still of interest to directly compare births on either side of our threshold based on observable characteristics.

Table A2 compares means of observable characteristics above and below the threshold, controlling for linear trends in grams from the threshold, as in the main analysis. A summary measure is the predicted mortality rate from a probit model of mortality on all of the controls (specifically, the newborn characteristics X_i' described above together with year indicators). Most of the comparisons show similar levels across the threshold, with few that appear to be meaningfully different. Given the large sample size, however, some of the differences are statistically significant.

To further consider these differences, Figure 5 compares covariates of interest in the 5 ounces around the VLBW threshold.[47] Here, the comparisons appear even more stable across the threshold. In particular, gestational age, which is particularly related to birth weight, is generally smooth through the threshold. Prenatal visit rates appear different in the linear-trend specifications in Table A2, but Figure 5G reveals little difference. While Table A2 suggests a statistically significant difference in predicted mortality, Figure 5J, which is in the same scale as actual mortality Figure 2, suggests little difference across the threshold. It appears that newborns are nearly identical based on observable variables regardless of whether they weighed in at a level just below or just above the VLBW threshold.

B. Sample selection

One possible source of sample selection is the possibility that very sick infants are discontinuously reported more frequently as fetal deaths across our cutoffs of interest. Although we would expect

[47] The list was selected for ease of presentation and includes the major covariates of interest. Similar results were found for additional covariates as well.

such sample selection to be apparent in the births histogram we examined above, we can also test for such sample selection directly using data on fetal death reports from the National Center for Health Statistics (NCHS) perinatal mortality data for 1995 to 2002. There are two types of fetal death classifications in these data: greater than or equal to 20 gestational weeks, and less than 20 gestational weeks. In practice, the frequency of fetal deaths for less than 20 gestational weeks is negligible for all birth weights above 1000 grams, and thus near our discontinuities only the greater than or equal to 20 gestational week fetal deaths are relevant. As in the births data, there are reporting spikes for fetal deaths at ounce-equivalent gram intervals, but graphical analyses of these data does not suggest any spikes in fetal death reporting that are discontinuous across our cutoffs of interest.

C. Bandwidth sensitivity

The local-linear regression results are qualitatively similar for a wide range of bandwidths (see Table 6). The magnitude of the mortality estimates decreases with the bandwidth, suggesting that our relatively large bandwidth is conservative. When the bandwidth includes only one ounce on either side of the threshold (h = 30 grams), the difference in 1-year mortality is -2.7 percentage points; when h = 150 grams, the estimate decreases to -0.8 percentage points, which is similar to our main results.

In terms of the treatment measures in the 5-state sample, the hospital charges are estimated to decrease by approximately $8,000 when 1 ounce is used as the bandwidth and $6500 when 150 grams is used, compared to our estimate of just over $9,000. When hospital costs or length of stay are examined, our estimates are similar regardless of bandwidth: *e.g.* $4,300 for hospital costs and 2.4 days for a one-ounce bandwidth; $2600 and 1.14 days when h = 150. Meanwhile, when log charges are considered (not shown), the estimates decrease as the bandwidth increases.

D. Causes of death

For several reasons, it is of interest to ask which causes of death appear to account for our observed mortality effect. First, if our mortality effect appeared to be driven by so-called "external" causes of death (such as accidents), this would be concerning since it would be difficult to link deaths from those causes to differences in medical inputs. Second, from a policy

perspective, it is of substantive interest to know which causes of death appear to be driving our results.

On the first point, a graphical examination of trends in "external" causes of death as a share of all deaths across our discontinuities does not suggest any spikes in external deaths that are discontinuous across our cutoffs of interest. Formalizing this in our regression framework, we find no statistically significant change in external deaths across our cutoffs of interest (see Appendix Table A4).

On the second point, we examine cause of death for newborns around 1500 grams as follows. We group causes of death into broad, mutually exclusive categories that are consistent over time based on the 9th and 10th revisions of the ICD cause of death classes.[48] Graphically, a visually discernable jump is primarily noticeable for deaths due to perinatal conditions (such as jaundice and respiratory distress syndrome) and, to a lesser extent, for deaths due to nervous system and sense organ disorders. These results support the notion that differences in care received in the hospital are likely driving the main results, as these conditions are known to be relatively common neonatal infections. Formalized regression estimates (see Appendix Table A4) confirm an economically significant and modestly statistically significant jump in perinatal conditions, as well as for nervous system and sense organ disorders.[49]

We also examine a few individual causes of death - namely, respiratory distress syndrome (RDS), sudden infant death syndrome (SIDS), and jaundice. We find no visible discontinuities for RDS or SIDS (nor any statistically significant regression estimates). For jaundice, the graphical results are noisy but consistent with an increase in deaths due to jaundice above 1500 grams, a result confirmed in regression estimates which find a modestly statistically significant jump in jaundice deaths when moving from just below to just above 1500 grams.[50]

[48] In particular, the 9th-revision 61 infant cause of death and 10th-revision 130 infant cause of death recode groups.were used to create the following categories: (1) infectious and parasitic diseases (such as meningococcal diseases); (2) neoplasms (that is, cancers); (3) endocrine, nutritional, metabolic, immunity, and blood disorders (such as cystic fibrosis); (4) nervous system and sense organ disorders (such as meningitis); (5) respiratory system disorders (such as pneumonia); (6) digestive system disorders (such as gastritis); (7) congenital anomalies (such as congenital malformations of the heart); (8) perinatal conditions (such as jaundice); (9) symptoms, signs, and ill-defined conditions (such as SIDS, or sudden infant death syndrome); and (10) other causes of death (such as external deaths).

[49] The nervous system and sense organ disorder category includes meningitis, which is a common form of morbidity and mortality among newborns - particularly in the first month of life (Levene *et al.*, 2008).

[50] Neonatal jaundice is a common problem among newborns (Levene *et al.*, 2008), and it should be detected during the initial hospital stay as opposed to after discharge for newborns in our bandwidth. A reference in the medical literature gives the following background information on jaundice: "Jaundice is observed during the first week of life in approximately 60% of term infants and 80% of preterm infants...Jaundice, consisting of indirect or

VII. Variation in treatment effects

Several potential sources of heterogeneity in our estimated treatment effects are of interest. In this section, we discuss over time (sub-section A), across hospitals (sub-section B), and across sub-groups of newborns (sub-section C).

A. Time period & technology changes

It is possible that that there is heterogeneity in outcomes and treatment across time - due, for example, to technological changes in medical treatments for at-risk newborns over the time period we examine. For example, one major technological innovation occurring during our sample is the development of artificial surfactant. Respiratory problems in newborns are frequently caused by insufficient pulmonary surfactant, a "wetting" protein on the surface of the alveoli (Beers, 2003). Surfactant helps to keep the alveoli open, and a lack of surfactant is often referred to as respiratory distress syndrome (RDS). The use of artificial surfactant began in the early 1990s, but unfortunately we do not observe the use of surfactant in our primary data sets. However, even without data on surfactant use we can still examine differences in our first stage and reduced form results across different time periods to test for general evidence of a "surfactant" effect.[51]

Table 5 reports 1-year mortality results for four time periods available in our data: 1983-1987; 1988-1991; 1995-1998; and 1999-2002. In general, we find a reduction in mortality associated with VLBW status in each period, even though average mortality just above the threshold declines from 8.13% in the first period to 3.78% in the last period. The estimates suggest 18%, 12%, 2%, and 18% lower mortality rates compared to the mean rates for infants born within 3 ounces above the threshold. In general, these trends over time are not consistent with a "surfactant" story or any other clear medical technology story of which we are aware.

The five-state discharge data also allow us to consider changes in our first stage over time.

direct bilirubin, that is present at birth or appears within the first 24 hours of life requires immediate attention...." (Behrman *et al.*, 2000).

[51] The more recent birth certificate data referenced above include an indicator for the use of artificial surfactant. We do not see a visible discontinuity in this variable, again potentially due to the small sample of births. These data also include an indicator for steroid medication administered to mothers prior to birth to help their newborns' lungs produce more surfactant and mature more quickly; although ideally we would use this indicator as the basis for a placebo test (since the medication is provided prior to the measurement of birth weight), the fact that we did not observe a visible discontinuity on this variable may again be due to the small sample of births for which we observe this variable.

We again divide the sample into four periods, in part to coincide with the time periods studied in the nationwide data: 1991-1994, 1995-1998, 1999-2002, and 2003-2006. Here, hospital charges appear to be greater for those born just below the threshold in all time periods, with a smaller difference in the 2003-2006 period.[52] In general, this suggests that both our first stage and reduced form results are relatively unchanged over time.[53]

B. Hospital types

We also consider potential heterogeneity in outcomes and treatment across hospitals, which our regression discontinuity design allows us an opportunity to rigorously address. In contexts without a regression discontinuity, comparisons of births in higher quality hospitals to births in lower quality hospitals could be biased: on one hand, a positive correlation could arise if healthier mothers choose to give births at better hospitals; on the other hand, a negative correlation could arise if riskier mothers choose to give birth at better hospitals, knowing that their infant will need more care than an average newborn. However, as discussed above, because birth weight should not be predictable in advance of birth with the accuracy needed to move a birth from just above to just below our 1500 gram threshold of interest, selection should not be *differential* across our discontinuity - implying that we can calculate internally valid estimates for different types of hospitals and consider how the quality of the hospital affects the results.

One natural grouping of hospitals, given our population under study, is the level of neonatal care available in an infant's hospital of birth. For our California data, classifications of neonatal care availability by hospital by year are available during our time period due to analysis by Phibbs *et al.* (2007).[54] In the sample of newborns within our bandwidth, 10% of births occur at hospitals with no NICU, just over 12% at hospitals with a Level 0-2 NICU, and the remainder at hospitals with Level 3A-3D NICUs.[55]

[52] Part of this difference arises because the states in our sample change with time because of data availability (see Appendix Table A3) - the last period does not include California, for example.

[53] Plotting first stage estimates by year against reduced form estimates by year for our five-state sample, normalizing each coefficient by the mean outcome for newborns above 1500 grams within our bandwidth, suggests that years with larger first stage estimates are associated with larger reduced form estimates.

[54] We are grateful to Christopher Afendulis and Ciaran Phibbs for sharing this data with us. Phibbs *et al.* (2007) used the same California data we study to identify the quality level of NICUs (Levels 1 to 3D) by hospital by year, in part based on NICU quality definitions from the American Academy of Pediatrics (definitions which in turn are primarily based on whether hospitals offer specific types of procedures, such as specific types of ventilation and surgery).

[55] In particular, 1.2% at hospitals with a Level 0 NICU, 0.04% at hospitals with a Level 1 NICU, 11.3% at hospitals with a Level 2 NICU, 10.9% at hospitals with a Level 3A NICU, 28.7% at hospitals with a Level 3B

While we can examine our reduced form estimates by NICU quality level (no NICU, and levels 0, 1, 2, 3A, 3B, 3C, and 3D), it is worth noting that we expect to lack sufficient sample size within these NICU quality level sub-samples to give clean estimates of these effects for our one-year mortality outcome.[56] Regression estimates which interact our regression discontinuity variable as well as our linear birth weight trends with indicators for the NICU quality level available in a newborn's hospital of birth generally do not give statistically significant estimates for our one-year mortality outcome, with the exception of Level 0/1/2 NICU hospitals - for which we estimate a negative, statistically significant coefficient.[57] Qualitatively, the mortality coefficients are negative for non-NICU hospitals as well as Level 0/1/2 and Level 3D NICU hospitals (although the magnitude of the coefficient for Level 3D NICU hospitals is orders of magnitude smaller than the other two coefficients), whereas the mortality coefficients for other hospitals are positive (again, not statistically significant).

Using charges as a first stage outcome in the same regression framework, we estimate economically and statistically significant positive coefficients for non-NICU hospitals as well as Level 0/1/2 and Level 3B hospitals; coefficients for the other hospitals do not produce statistically significant coefficients.[58]

Clearly we can only give a very cautious interpretation of these trends given that many of our estimates are not statistically significant at conventional levels. That said, Figure 6 plots one descriptive analysis - namely, plotting first stage estimates by hospital against reduced form estimates by hospital, normalizing each coefficient by the mean outcome for newborns above 1500 grams within our bandwidth for that type of hospital. Hospitals with larger first stage estimates have larger reduced form estimates, which provides further evidence that treatment differences are driving the outcome differences. In addition, this analysis provides suggestive evidence that

NICU, 21.6% at hospitals with a Level 3C NICU, and 16.2% at hospitals with a Level 3D NICU. Because of the low number of births we observe in Level 0 or Level 1 NICUs, we create a combined category for births in Level 0, 1, an 2 NICU hospitals.

[56] In the California data, we estimate a treatment coefficient of -0.0027 (s.e.=0.0068, p=0.6909). When all years in the nationwide data are included for California, however, we do find a statistically significant effect: coefficient of -0.014 (s.e.=0.0066).

[57] Specifically, we estimate the following: for non-NICU hospitals, -0.0304 (s.e.=0.0223, p=0.1735); for Level 0/1/2 NICU hospitals, -0.0436 (s.e.=0.0186, p=0.0193); for Level 3A NICU hospitals, 0.0032 (s.e.=0.0209, p=0.8773); for Level 3B NICU hospitals, 0.0141 (s.e.=0.0116, p=0.2244); for Level 3C NICU hospitals, 0.0076 (s.e.=0.0149, p=0.6099); and for Level 3D hospitals, -0.0017 (s.e.=0.0177, p=0.9224).

[58] Specifically, we estimate the following: for non-NICU hospitals, 15000 (s.e.=7800, p=0.0534); for Level 0/1/2 NICU hospitals, 19000 (s.e.=5300, p=0.0003); for Level 3A NICU hospitals, 10000 (s.e.=9600, p=0.2967); for Level 3B NICU hospitals, 15000 (s.e.=7100, p=0.0400); for Level 3C NICU hospitals, 453 (s.e.=11000, p=0.9678); and for Level 3D hospitals, -12000 (s.e.=9900, p=0.2306).

the non-NICU and Level 0/1/2 NICU hospitals are the hospitals where our estimated effects are largest.[59]

C. Sub-group analyses

The final set of results compares the mortality results across different values of our main covariates. We concentrate on the mortality outcomes because we can take advantage of the large sample sizes in the nationwide data to split the sample into subgroups. The estimates provide suggestive evidence on the sources of our main results. In particular, we find statistically significant differences for less educated mothers; newborns with missing father's information (a proxy for single parenthood in our data, which otherwise lacks a stable marital status indicator); single births (where low birth weight may point to greater developmental problems); and male patients (who are known to be more vulnerable). The first stage estimates by subgroup exhibit similar differences, with a larger first stage for male newborns and singleton births.

VIII. Alternative thresholds

In this section we discuss alternative birth weight thresholds (sub-section A) as well as gestational age-based thresholds (sub-section B).

A. Alternative birth weight thresholds

A main limitation to our analysis is that the returns are estimated at a particular point in the birth weight distribution. To the extent that we find large returns to treatment for newborns just below the 1500 gram threshold, the evidence suggests that the threshold should be moved to a higher birth weight where the cost of saving a statistical life is likely closer to the estimates of the value of a statistical life.

We can also examine other points in the birth weight distribution where differences in treatment may be expected. The presence of discontinuities at other thresholds need not invalidate our main findings at 1500 grams. Rather, other discontinuities could provide an opportunity

[59] Another way to consider treatment intensity in the nationwide data is to compare states that have higher end-of-life spending levels according to the Dartmouth Atlas of Healthcare: a resource that considers Medicare spending. When the 1996 state rankings are used (the earliest year available, although the rankings are remarkably stable over the years 1996-2005), the mortality effects are found in the bottom two and top two quintiles, suggesting that the results are fairly robust across different types of hospital systems that vary by spending levels.

to trace out marginal returns for wider portions of the overall birth weight distribution. At points in the distribution where we do not anticipate treatment differences, however, economically and statistically significant jumps of magnitudes similar to our VLBW treatment effects could suggest that the discontinuity we observe at 1500 grams may be due to natural variation in treatment and mortality in our data.

Discussions with physicians and readings of the medical literature suggest that other cutoffs may be relevant. The ICD-9 codes, for example, list separate diagnosis codes (V21.30-V21.35) that depend on birth weight in grams: 0-500, 500-999, 1000-1499, 1500-1999, 2000-2500, *etc.* Cloherty & Stark (1998) report separate recommendations for the care of "extremely low-birth weight infants (ELBW)" with birth weight lower than 1000 grams. In addition, there may be other salient thresholds in gestational age that physicians use to determine care.

To investigate other potential thresholds, we estimate differences in mortality and hospital charges for each 100 gram interval between 1000 and 3000 grams. We use local linear regression estimates because they are less sensitive to observations far from the thresholds, and our pilot bandwidth of 3 ounces for comparability.

In terms of the mortality differences, the largest difference in mortality compared to the mean at the cutoff is found at 1500 grams (23%), other than one found at 1800 grams (27%).[60] A 5% reduction in mortality (relative to the mean) is found at 1000 grams and a 16% reduction in mortality if found at 2500 grams. That said, the differences at 1000 and 2500 grams appear to be driven by the inappropriate use of linear trends before and after these thresholds, as graphs do not reveal convincing discontinuities in mortality at these, or other, cutoffs.

When we considered hospital charges, again 1500 grams stands out with a relatively large discontinuity, especially compared to discontinuities at birth weights between 1100 and 2500g. A 12% increase in charges (relative to the mean) is found for newborns classified as extremely low birth weight (1000 grams), with similarly large differences for 800 and 900 gram thresholds. These differences at and below 1000 grams are not robust to alternative specifications such as the transformation of charges by the natural logarithm, however. One explanation for the lack of stability in the estimates at these alternative birth weight thresholds is that there are fewer

[60] 1800 grams is a commonly cited threshold for changes in feeding practices (Cloherty & Stark, 1998). However, we cannot observe changes in feeding practices in our data, and, as discussed in the next paragraph, we do not observe a correspondingly large discontinuity at 1800 grams in our hospital charges measure.

newborns to study and the spending levels are particularly susceptible to outliers given the large charge amounts. In summary, we find striking discontinuities in treatment and mortality at the VLBW threshold, but less convincing differences at other points of the distribution.

B. Gestational age and SGA analyses

As motivated by the discussion in Section IIB, we examine heterogeneity in outcomes and treatment by gestational age across the 37-week threshold. In graphical analyses using the nationwide sample, measures of average mortality by gestational week appear smooth across the 37-week threshold. Similarly, in graphical analyses using the California data, which report gestation in days, measures of average mortality, charges, and length of stay by gestational day appear smooth across this threshold. Corresponding regression results yield statistically significant coefficients of the expected sign, but we do not emphasize them here given the lack of a visibly discernable discontinuity in the graphical analysis.[61]

Even though we do not observe meaningful discontinuities in outcomes or treatment at 37 gestational weeks, there is still reason to investigate the *interaction* between birth weight and gestational age through the "small for gestational age" (SGA) classification: newborns below the 10th percentile of birth weight for a given gestational age. From conversations with physicians, we have reason to believe that doctors use SGA charts such as that established by Fenton (2003).[62] On this chart, 2500 grams is almost exactly the 10th percentile of birth weight for a gestational age of 37 weeks. If physicians treat based on SGA cutoffs, we expect discontinuities in outcomes and treatment at 2500 grams to be most pronounced exactly at 37 weeks and less pronounced at other values of gestational weeks, although we are agnostic about the pattern of decline. In regression results (not shown) in which we fully interact the low birth weight indicator and linear trends with dummies for each value of gestational week, we indeed find that the effect of the low birth weight designation is largest in magnitude at 37 weeks, and it declines as gestational weeks move away from 37 in both directions. Several of the interacted low birth

[61] Specifically, the coefficient on an indicator variable for "below 37 gestational weeks" is -0.00070 (robust s.e.=0.0001277) in a specification that includes linear trends, run on an estimation sample of 21,562,532 observations within a 3 week bandwidth around 37 weeks. Mean mortality above the threshold is 0.0032. To address the concern that discontinuities could be obscured in cases where gestational age can be manipulated, we also estimate a specification which includes only vaginal births that are not induced or stimulated and find similar results.

[62] The SGA chart by Fenton (2003) updates the previous work of Babson & Benda (1976), which was available during our sample period.

weight indicators are statistically different from zero. Overall, this evidence is consistent with treatment based on SGA around 2500 grams.

We also examine a potential SGA treatment threshold around the 1500 gram discontinuity. In the Fenton (2003) chart, 1500 grams is considered SGA for newborns with between 32 and 33 gestational weeks. In regression results (not shown), we see that discontinuities in mortality around 1500 grams are most pronounced at 29 weeks and decrease on either side of 29 weeks, which is not clearly consistent with treatment based on the Fenton (2003) definition of SGA around 1500 grams.

IX. Estimating returns to medical spending

In this section, for comparability to the existing literature we present a Cutler & Meara (2000)-style time series estimate of the returns to large changes in spending over time for newborns in our bandwidth (sub-section A). We then combine our first stage and reduced form estimates to derive two-sample estimates of the marginal returns to medical spending for newborns near 1500 grams (sub-section B).

A. Comparison to time-series estimates of returns to medical spending

As one benchmark, we can compare our marginal return estimate to the type of return estimate calculated by Cutler & Meara (2000). The spirit of the Cutler-Meara calculation is to assume that within-birth weight changes in survival over time are primarily due to improvements in medical technologies, and to thus value medical improvements by looking at changes over time in within-birth weight expenditures and health outcomes. For comparability to our marginal returns calculation, we undertake this calculation in our California data as a "long difference" in costs (in 2006 dollars) and one-year mortality from 1991 to 2002. Within our bandwidth, we estimate a $30,000 increase in costs and a 0.0295 decline in one-year mortality over this period, which implies a return under the Cutler-Meara assumptions of $1 million dollars. By this metric, as we will see below, our marginal return estimates appear to be similar or slightly more cost-effective than time-series returns to large changes in spending for newborns in our bandwidth.

B. Two-sample estimates of marginal returns to medical spending

As discussed in Section IV, we can combine our results to produce two-sample estimates of the effect of treatments on health outcomes around the VLBW threshold. To do so, we need to invoke the exclusion restriction that the VLBW designation can only affect mortality through treatments captured by our treatment measure. As discussed above, this assumption is more plausible for our summary treatment measures (charges, costs, and length of stay) than it is for our mechanism variables. We focus on costs to summarize treatment in terms of dollars.

Because we examine health outcomes and summary treatments in different data sources, we require additional assumptions to combine our estimates. To be as conservative as possible in these assumptions, we can restrict our combined estimate to California, the only population for which we observe one-year mortality and costs in the same data set. (As discussed above, we do not observe one-year mortality in the other states in our 5-state sample.) In the California data, in a specification without covariates, we estimate that costs increase by \$2900 (se = 1770) as birth weight approaches the VLBW threshold from above, but we do not estimate a statistically significant change in mortality across the threshold. Thus, we do not have enough statistical power to estimate the effect of higher treatment levels on mortality using our regression discontinuity design in this sample.

For another conservative combination of estimates with more power, we can combine mortality and cost estimates based only on the states in the 5-state sample. Because we only observe one-year mortality for these states in the national data, we obtain the one-year mortality estimate on the national data, restricted so that it contains only those newborns in the 5-state sample in available years. We standardize covariates across the two samples, so that if we had the exact same newborns in the two samples, our two-sample estimate would be identical to a one-sample estimate on the complete data.[63] Coefficients are shown in the last column of Table 5, where \$4,550 in additional costs are associated with a 0.74 percentage point reduction in mortality.

If we are willing to assume that costs differences in the 5-state sample in the available years

[63]Specifically, we restrict the national data to the 5 states in the years 1991 and 1995-2002. Also, for comparability with the 5-state sample, we restrict the national sample to contain only in-hospital births. Because we do not have individual-level identifiers, we cannot restrict the national sample to contain the exact same newborns as the 5-state sample, but the agreement is very good. The restricted national sample contains 23,698 infants, and the 5-state sample contains 21,479.

(1991-2006) are broadly representative of what we would observe in the full national sample in available years (1983-2002), we can compare our main results: a difference of $3,812 in costs and a one-year mortality reduction of 0.73 percentage points as birth weight approaches the VLBW threshold from above.

Equivalently, we can compute a measure of dollars per newborn life saved. In such a calculation, the numerator is our hospital costs estimate: $3,812 for each VLBW newborn in the full 5-state sample. The denominator is our mortality estimate: a 0.73 percentage point reduction in mortality among VLBW newborns in the full sample. These estimates imply that the cost per newborn life saved is $522,191 ($3,812/.0073). In the 5-state sample over the years that overlap with the nationwide data, we attain a slightly larger estimate of costs per newborn life saved of $615,270 ($4,553/.0074). Following Inoue & Solon (forthcoming), we calculate an asymptotic $95% confidence interval on this estimate of approximately $30,000 to $1.20 million. Note that this confidence interval for the estimate from the restricted sample is conservative relative to the analogous confidence interval for the more precise estimate we obtain from the full samples.[64]

We can compare these estimates of the cost per newborn life saved to a variety of potential benchmarks. If we take the very conservative view that the mortality effects that we estimate do not persist beyond one year, it is relevant to compare our costs estimates to estimates of the value of one year of life from the literature, which are generally around $100,000 (see Cutler (2004)). Based on this comparison, the interventions that we observe do not appear to be cost-effective. However, it is generally agreed that one-year mortality effects likely persist well beyond one year. In this case, we can compare our cost estimates to estimates of the value of an entire life. If our one-year mortality effects persist but life span and quality of life are reduced, we can compare our estimates to quality-adjusted value of newborn life calibrations, such as those from Cutler & Meara (2000). Specifically, in their analysis, Cutler & Meara (2000) calibrate the value of life in 1990 for newborns born between 1,000 and 2,499 grams (based on life expectancy and expected quality of life, including disabilities associated with being low birth weight) to be approximately $2,700,000, implying our estimated expenditures are cost effective, even at the upper bound of our more conservative 95% confidence interval. If we take the even less conservative view that our one-year mortality effects persist and that the newborns who are saved do not experience

[64] Using the full samples with common covariates, we obtain an estimate of $537,640 with an approximate 95% confidence interval of $30,000 to $1.05 million.

decreases in life span or quality of life, the relevant benchmark is approximately $3 to $7 million dollars (Cutler, 2004).[65] Comparison with this benchmark suggests that the treatments that we observe are very cost-effective.

X. Conclusion

In the universe of all births in the US over 20 years, we estimate that newborns weighing just below 1500 grams have substantially lower mortality rates than newborns that weigh just over 1500 grams, despite a general decline in health associated with lower birth weight. Specifically, one-year mortality falls by approximately one percentage point as birth weight crosses 1500 grams from above, which is large relative to mean one-year mortality of 5.5% just above 1500 grams. Robustness tests suggest some variation around this point estimate, but a reduction in mortality of close to 0.7 percentage points for newborns just below the threshold is generally found.

It appears that infants categorized as "very low birth weight" have a lower mortality rate because they receive additional treatment. Using all births from five states that report treatment measures and birth weight - states that have a similar mortality discontinuity to the nationwide sample - we find that treatment differences are on the order of $9,500 in hospital charges, or $4,000 when these charges are converted into costs. While these costs may not represent social costs for such care - the nurses, physicians, and capital expenditures may not be affected by the births of a small number of very low birth weight infants - they represent our best summary measurement of the difference in treatment that the VLBW classification affords. Taken together, our estimates suggest that the cost of saving a statistical life for newborns near 1500 grams is approximately $550,000 with an upper bound of approximately $1.2 million in 2006 dollars, suggesting that greater levels of spending for at-risk infants near 1500 grams would be expected to yield benefits that outweigh their costs.

[65] An alternative interpretation of our results would assume that the VLBW threshold was currently set "optimally" from a social perspective, which would imply that society places a relatively low value on lives of infants near this threshold. However, an innovation in this paper is the consideration of the universe of newborns in the US over a 20 year period, which allows us to precisely detect the mortality difference. It is possible that such effects are simply unknown to physicians or other institutions that determine treatment thresholds.

References

Almond, Douglas, & Doyle, Joseph. 2008. After midnight: A regression discontinuity design in length of postpartum hospital stays. *National Bureau of Economic Research (NBER) working paper 13877*.

Andre, Malin, Borgquist, Lars, Foldevi, Mats, & Molstad, Sigvard. 2002. Asking for 'rules of thumb': a way to discover tacit knowledge in medical practice. *Family Medicine*, **19**(6), 617–622.

Anspach, Renee. 1993. *Deciding Who Lives: Fateful Choices in the Intensive-Care Nursury*. University of California Press.

Babson, S. Gorham, & Benda, Gerda. 1976. Growth graphs for the clinical assessment of infants of varying gestational age. *Journal of Pediatrics*, **89**(5), 814–820.

Baicker, Katherine, & Chandra, Amitabh. 2004. Medicare spending, the physician workforce, and the quality of care received by Medicare beneficiaries. *Health Affairs*, **W4**, 184–197.

Beers, Mark. 2003. *The Merck Manual of Medical Information*. Second edn. Whitehouse, Station, New Jersey: Merck Research Laboratories.

Behrman, Richard, Kliegman, Robert, & Jenson, Hal. 2000. *Nelson Textbook of Pediatrics*. 16th edn. Philadelphia: W.B Saunders Company.

Card, David, & Lee, David. 2008. Regression discontinuity inference with specification error. *Journal of Econometrics*, **142**(2), 655–674.

Cheng, Ming-Yen, Fan, Jianqing, & Marron, J.S. 1997. On automatic boundary corrections. *Annals of Statistics*, **25**(4), 1691–1708.

Cloherty, John, & Stark, Ann. 1998. *Manual of Neonatal Care: Joint Program in Neonatology (Harvard Medical School, Beth Israel Deaconess Medical Center, Brigham and Women's Hospital, Children's Hospital Boston)*. 4th edn. Lippincott-Raven.

Cutler, David. 2004. *Your Money or Your Life: Strong Medicine for America's Health Care System*. Oxford University Press.

Cutler, David, & McClellan, Mark. 2001. Is technological change in medicine worth it? *Health Affairs*, **20**(5), 11–29.

Cutler, David, & Meara, Ellen. 2000. The technology of birth: Is it worth it? *Frontiers in Health Policy Research*, **3**(3).

Cutler, David, McClellan, Mark, Newhouse, Joseph, & Remler, Dahlia. 1998. Are medical prices declining? Evidence for heart attack treatments. *Quarterly Journal of Economics*, **113**(4), 991–1024.

Cutler, David, Rosen, Allison, & Vijan, Sandeep. 2006. The value of medical spending in the United States, 1960-2000. *New England Journal of Medicine*, **355**(9), 920–927.

Enthoven, Alain. 1980. *Health Plan: The Only Practical Solution to the Soaring Cost of Medical Care*. Addison Wesley.

Fenton, Tanis. 2003. A new growth chart for preterm babies: Babson and Benda's chart updated with recent data and a new format. *BMC Pediatrics*, **3**(1), 13.

Fisher, Elliott, Wennberg, John, Stukel, Therese, & Sharp, Sandra. 1994. Hospital readmission rates for cohorts of Medicare beneficiaries in Boston and New Haven. *New England Journal of Medicine*, **331**(15), 989–995.

Fuchs, Victor. 2004. More variation in use of care, more flat-of-the-curve medicine. *Health Affairs*, **23**(6), 104–107.

Goodman, David, Fisher, Elliott, Little, George, Stukel, Therese, hua Chang, Chiang, & Schoendorf, Kenneth. 2002. The relation between the availability of neonatal intensive care and neonatal mortality. *The New England Journal of Medicine*, **346**(20), 1538–1544.

Grumbach, Kevin. 2002. Specialists, technology, and newborns – Too much of a good thing? *The New England Journal of Medicine*, **346**(20), 1574–1575.

Horbar, Jeffrey, Badger, Gary, Carpenter, Joseph, Fanaroff, Avroy, Kilpatrick, Sarah, LaCorte, Meena, Phibbs, Roderic, & Soll, Roger. 2002. Trends in mortality and morbidity for very low birth weight infants, 1991-1999. *Pediatrics*, **110**(1), 143–151.

Imbens, Guido, & Lemieux, Thomas. 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, **142**(2), 615–635.

Inoue, Atsushi, & Solon, Gary. forthcoming. Two-sample instrumental variables estimators. *Review of Economics and Statistics*.

Kessler, Daniel, & McClellan, Mark. 1996. Do doctors practice defensive medicine? *Quarterly Journal of Economics*, **111**(2), 353–390.

Levene, Malcolm, Tudehope, David, & Sinha, Sunil. 2008. *Essential Neonatal Medicine*. Blackwell Publishing.

Lichtig, Leo, Knauf, Robert, Bartoletti, Albert, Wozniak, Lynn-Marie, Gregg, Robert, Muldoon, John, & Ellis, William. 1989. Revising diagnosis-related groups for neonates. *Pediatrics*, **84**(1), 49–61.

Luce, Brian, Mauskopf, Josephine, Sloan, Frank, Ostermann, Jan, & Paramore, L. Clark. 2006. The return on investment in health care: From 1980 to 2000. *Value in Health*, **9**(3), 146–156.

Ludwig, Jens, & Miller, Douglas L. 2007. Does head start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*, **122**(1), 159–208.

McClellan, Mark. 1997. The marginal cost-effectiveness of medical technology: A panel instrumental-variables approach. *Journal of Econometrics*, **77**(1), 39–64.

McCrary, Justin. 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, **142**(2), 698–714.

McDonald, Clement. 1996. Medical heuristics: The silent adjudicators of clinical practice. *Annals of Internal Medicine*, **124**(1), 56–62.

Murphy, Kevin, & Topel, Robert. 2003. *Measuring the Gains from Medical Research*. The University of Chicago Press. Kevin M. Murphy and Robert H. Topel (editors). Chap. The economic value of medical research, pages 41–73.

Nordhaus, William. 2002. The health of nations: The contribution of improved health to living standards. *National Bureau of Economic Research (NBER) working paper 8818*.

O'Connor, Gerald, Quinton, Hebe, Traven, Neal, Ramunno, Lawrence, Dodds, Andrew, Marciniak, Thomas, & Wennberg, John. 1999. Geographic variation in the treatment of acute myocardial infarction: The Cooperative Cardiovascular Project. *Journal of the American Medical Association*, **281**(7), 627–633.

Paneth, Nigel. 1995. The problem of low birth weight. *The Future of Children*, **5**(1), 19–34.

Phibbs, Ciaran, Baker, Laurence, Caughey, Aaron, Danielson, Beate, Schmitt, Susan, & Phibbs, Roderic. 2007. Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. *New England Journal of Medicine*, **356**(21), 2165–2175.

Pilote, Louise, Califf, Robert, Sapp, Shelly, Miller, Dave, Mark, Daniel, Weaver, Douglas, Gore, Joel, Armstrong, Paul, Ohman, Magnus, & Topol, Eric. 1995. Regional variation across the United States in the management of acute myocardial infarction. *New England Journal of Medicine*, **333**(9), 565–572.

Porter, Jack. 2003. Estimation in the regression discontinuity model. *Working Paper*.

Pressman, Eva, Bienstock, Jessica, Blakemore, Karin, Martin, Shari, & Callan, Nancy. 2000. Prediction of birth weight by ultrasound in the third trimester. *Obstetrics & Gynecology*, **95**(4), 502–506.

Quinn, Kevin. 2008. New directions in Medicaid payment for hospital care. *Health Affairs*, **27**(1), 269–280.

Russell, Rebecca, Green, Nancy, Steiner, Claudia, Meikle, Susan, Howse, Jennifer, Poschman, Karalee, Dias, Todd, Potetz, Lisa, Davidoff, Michael, Damus, Karla, & Petrini, Joann. 2007. Cost of hospitalization for preterm and low birth weight infants in the United States. *Pediatrics*, **120**(1), e1–e9.

Stukel, Therese, Lucas, Lee, & Wennberg, David. 2005. Long-term outcomes of regional variations in the intensity of invasive versus medical management of Medicare patients with acute myocardial infarction. *New England Journal of Medicine*, **293**(11), 1329–1337.

Trochim, William. 1984. *Research Design for Program Evaluation: The Regression-Discontinuity Design*. Sage Publications.

Tu, Jack, Pashos, Chris, Naylor, David, Chen, Erluo, Normand, Sharon-Lise, Newhouse, Joseph, & McNeil, Barbara. 1997. Use of cardiac procedures and outcomes in elderly patients with myocardial infarction in the United States and Canada. *New England Journal of Medicine*, **336**(21), 1500–1505.

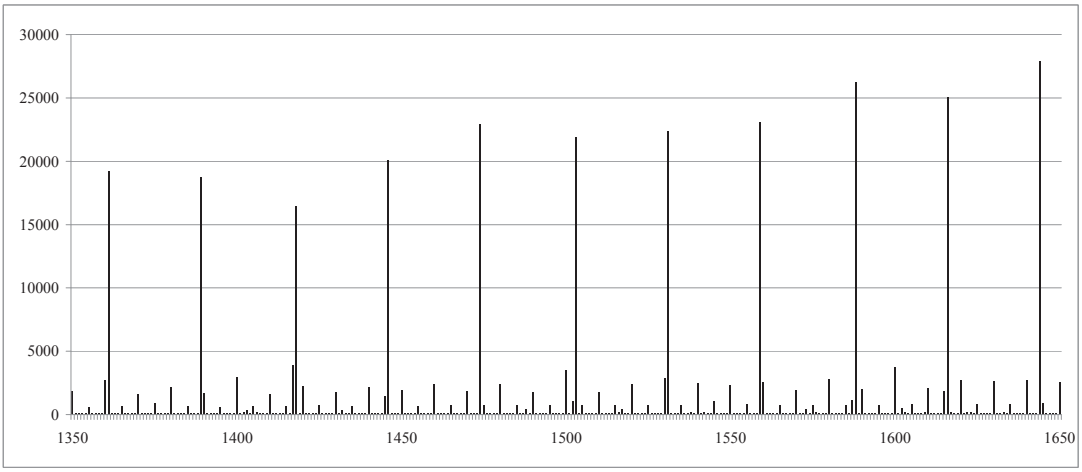
United States Congress, Office of Technology Assessment. 1981. *The Implications of Cost-Effectiveness Analysis of Medical Technology, Background Paper 2: Case studies of medical technologies; Case study 10: The costs and effectiveness of neonatal intensive care*. Author.

United States Institute of Medicine. 1985. *Preventing Low Birthweight*. National Academies Press.

Williams, Ronald, & Chen, Peter. 1982. Identifying the source of the recent decline in perinatal mortality rates in California. *New England Journal of Medicine*, **306**(4), 207–214.

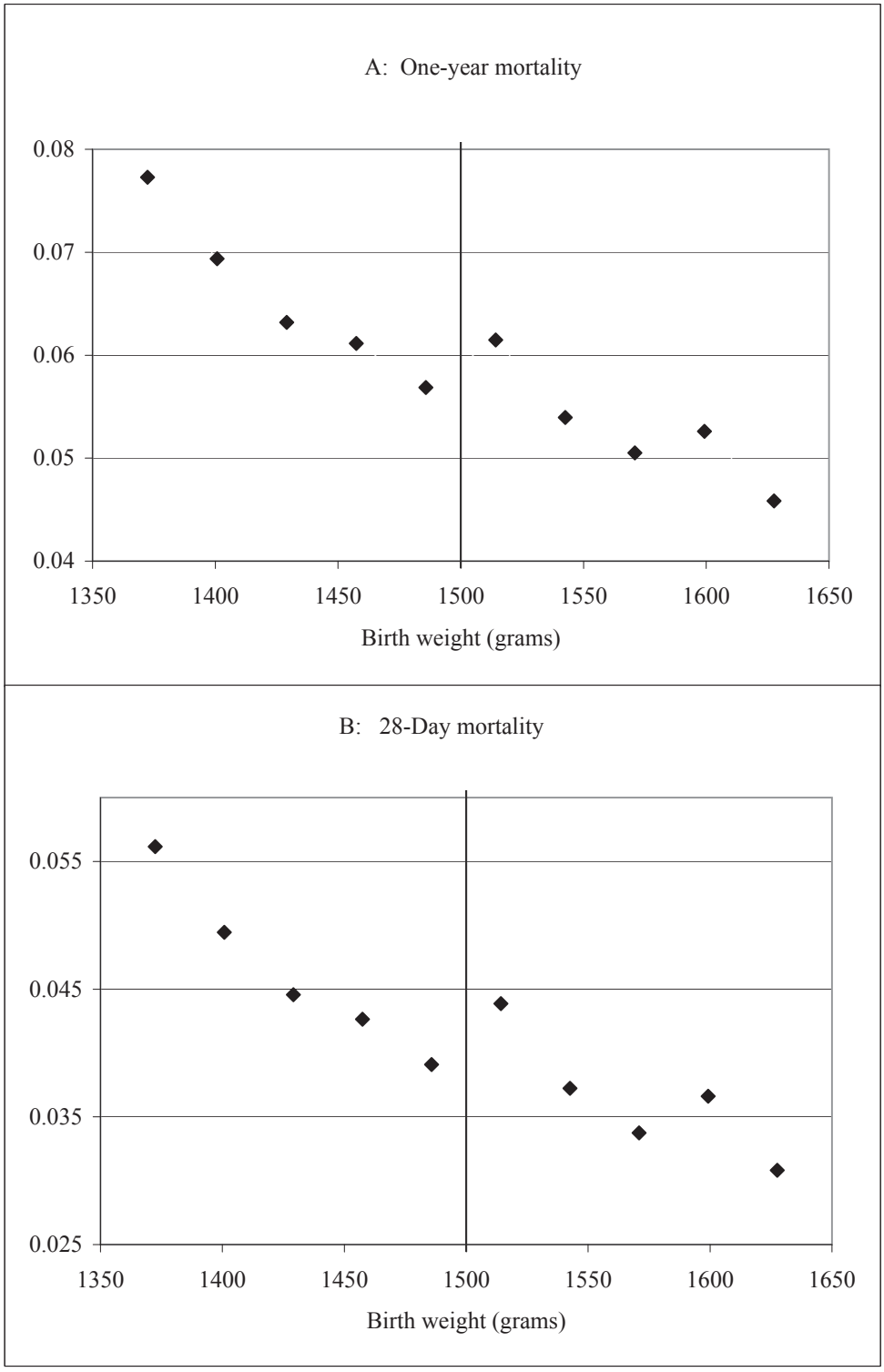
Zupancic, John, & Richardson, Douglas. 1998. Characterization of the triage process in neonatal intensive care. *Pediatrics*, **102**(6), 1432–1436.

Figure 1: Frequency of births by gram: Population of US births between 1350-1650 grams



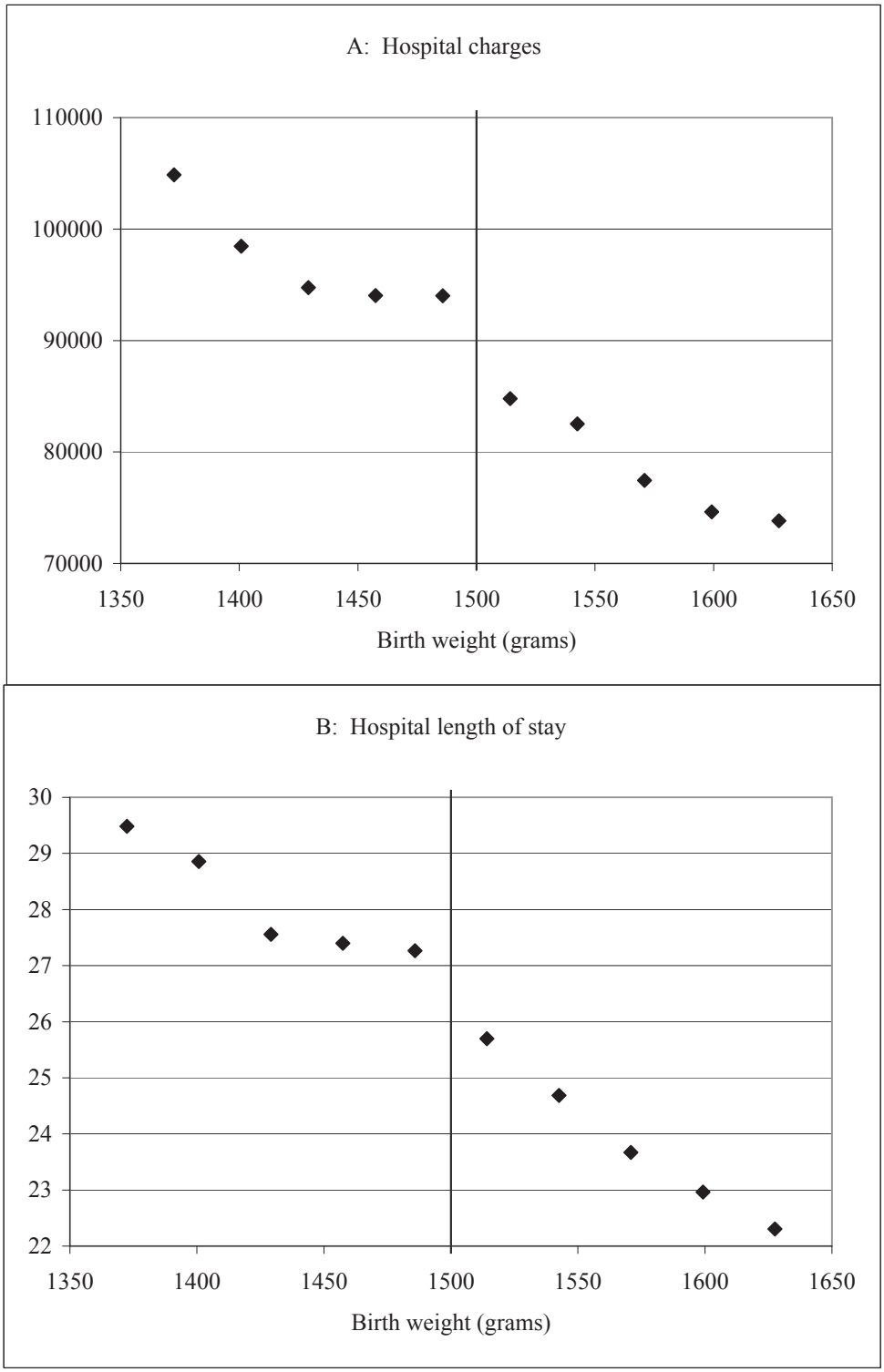
Notes: NCHS birth cohort linked birth/infant death files, 1983-1991 and 1995-2003, as described in the text.

Figure 2: One-year and 28-day mortality around 1500 grams



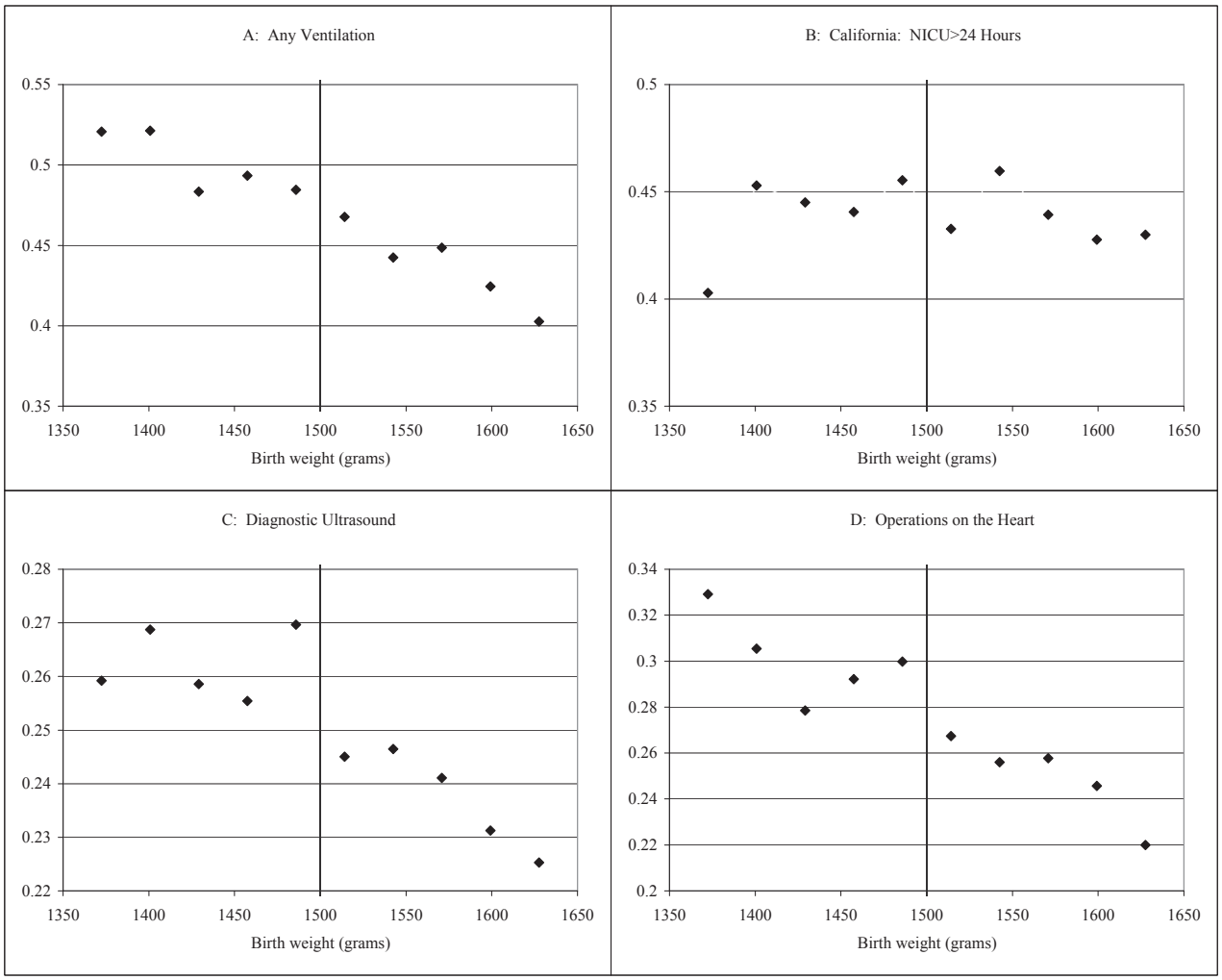
Notes: NCHS birth cohort linked birth/infant death files, 1983-1991 and 1995-2003, as described in the text. Points represent means in one-ounce bins radiating from 1500 grams.

Figure 3: Summary treatment measures around 1500 grams



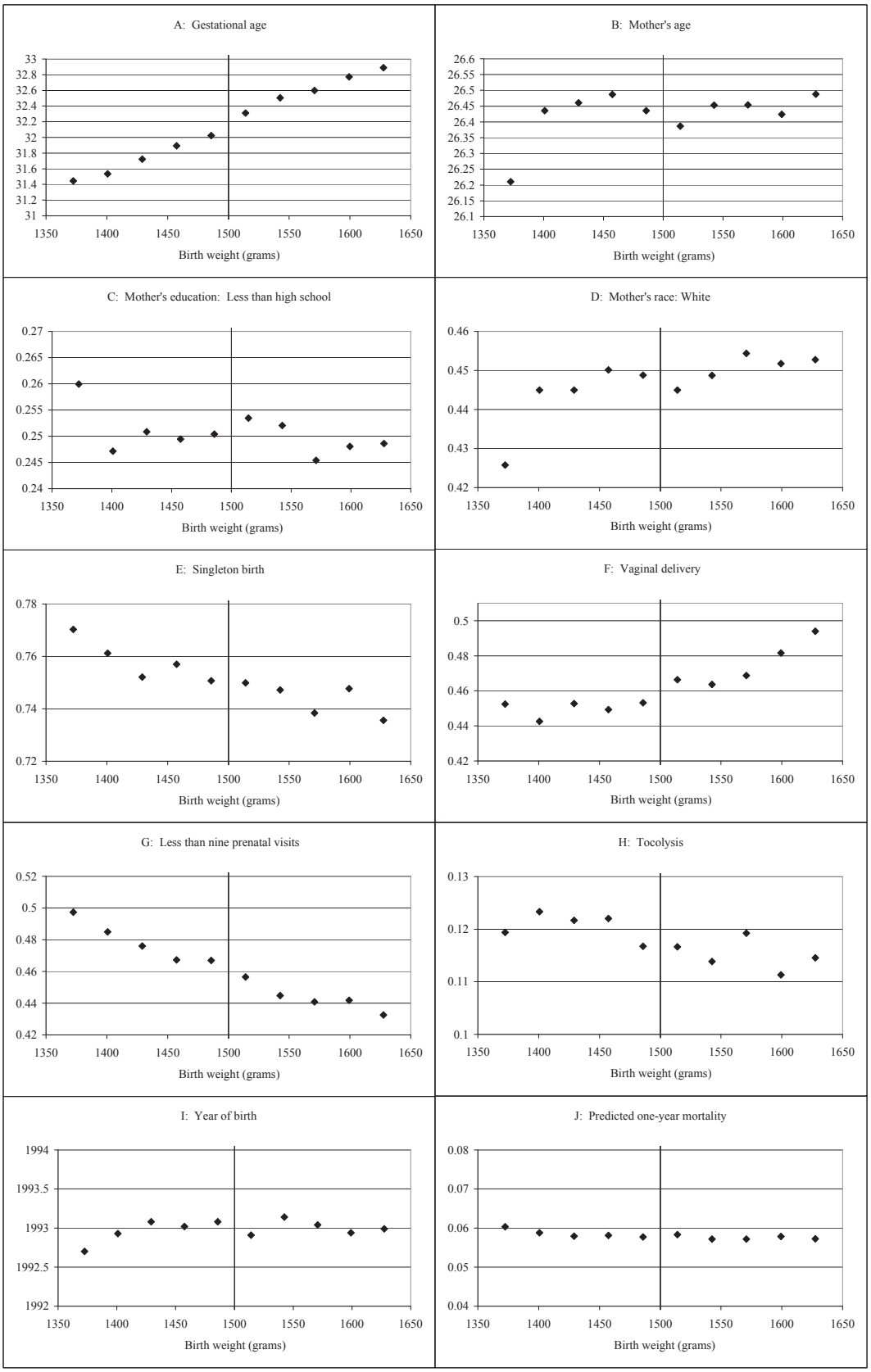
Notes: Data are all births in the 5-state sample (AZ, CA, MD, NY, and NJ), as described in the text. Charges are in 2006 dollars. Points represent means in one-ounce bins radiating from 1500 grams.

Figure 4: Specific treatment measures around 1500 grams



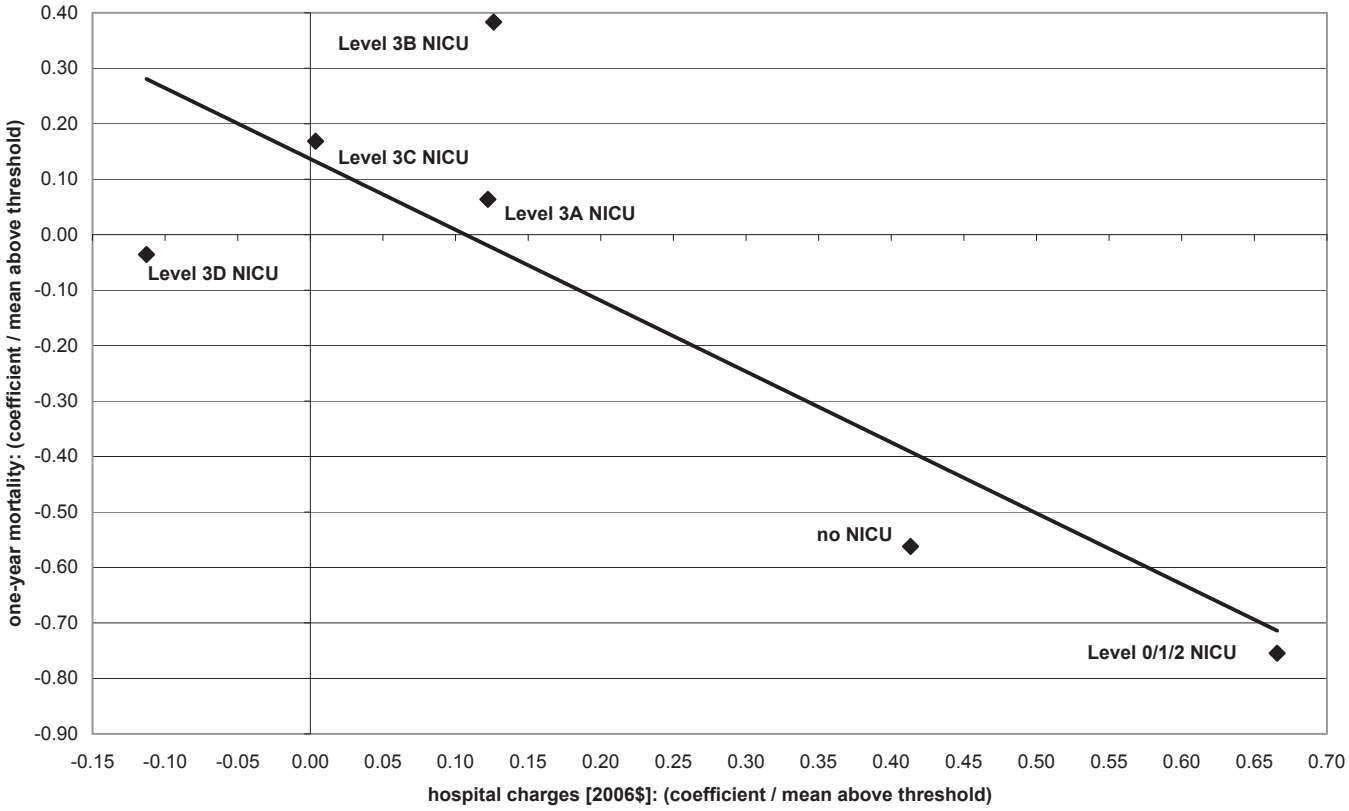
Notes: Data are all births in the 5-state sample (AZ, CA, MD, NY, and NJ), as described in the text. Points represent means in one-ounce bins radiating from 1500 grams.

Figure 5: Covariates around 1500 grams



Notes: NCHS birth cohort linked birth/infant death files, 1983-1991 and 1995-2003. Points represent means in one-ounce bins radiating from 1500 grams.

Figure 6: First stage versus reduced form, by NICU quality level



Notes: Plot of first stage coefficients (for 2006 charges, in levels) and reduced form coefficients (for one-year mortality) by NICU level in our California data. See text for details on the NICU classifications.

Table 1: Infant mortality by very low birth weight status, National data, 1983-2002 (available years)

Dependent variable:	one-year mortality				28-day mortality			
Model:	local linear	OLS	OLS	OLS	local linear	OLS	OLS	OLS
Birth weight < 1500g	-0.0121	-0.0095	-0.0076	-0.0073	-0.0107	-0.0088	-0.0074	-0.0073
	(0.0023)**	(0.0022)**	(0.0022)**	(0.0022)**	(0.0019)**	(0.0018)**	(0.0018)**	(0.0018)**
Birth weight < 1500g * Grams from cutoff		-0.0001	-0.0001	-0.0001		-0.0001	-0.0001	-0.0001
		(0.0000)**	(0.0000)**	(0.00004)**		(0.0000)**	(0.0000)**	(0.00003)**
Birth weight >= 1500g * Grams from cutoff		-0.0002	-0.0002	-0.0002		-0.0002	-0.0002	-0.0002
		(0.0000)**	(0.0000)**	(0.00003)**		(0.0000)**	(0.0000)**	(0.00002)**
Year controls		No	Yes	Yes		No	Yes	Yes
Main controls		No	No	Yes		No	No	Yes
Mean of dependent variable above cutoff:	0.0553				0.0383			

Dependent variable:	7-day mortality				24-hour mortality			
Model:	local linear	OLS	OLS	OLS	local linear	OLS	OLS	OLS
Birth weight < 1500g	-0.0068	-0.0060	-0.0049	-0.0048	-0.0068	-0.0043	-0.0036	-0.0035
	(0.0017)**	(0.0016)**	(0.0016)**	(0.0016)**	(0.0017)**	(0.0013)**	(0.0013)**	(0.0013)**
Birth weight < 1500g * Grams from cutoff		-0.0001	-0.0001	-0.0001		-0.0001	-0.0001	-0.00004
		(0.0000)**	(0.0000)**	(0.00002)**		(0.0000)*	(0.0000)	(0.00002)
Birth weight >= 1500g * Grams from cutoff		-0.0001	-0.0001	-0.0001		-0.0001	-0.0001	-0.00009
		(0.0000)**	(0.0000)**	(0.00002)**		(0.0000)**	(0.0000)**	(0.00002)**
Year controls		No	Yes	Yes		No	Yes	Yes
Main controls		No	No	Yes		No	No	Yes
Mean of dependent variable above cutoff:	0.0301				0.0191			
Observations	202071							

Notes: Local linear regressions use a bandwidth of 3 ounces (85 grams). OLS models estimated on a sample within 3 ounces above and below the VLBW threshold. "Main controls" are listed in Table A1, as well as indicators for 5-year intervals of mother's age, 5-year intervals of father's age, gestational week, state of residence, year, as well as missing-information indicators for the prenatal, birth order, gestational age and mother's race categories. * significant at 5%; ** significant at 1%. Robust standard errors in parentheses.

Table 2: Summary treatment measures by very low birth weight status, 5-state sample, 1991-2006

Dependent variable:	hospital charges				length of stay			
Model:	local linear	OLS	OLS	OLS	local linear	OLS	OLS	OLS
Birth weight < 1500g	9450	9021.92	8205.34	9065.17	1.97	1.78	1.76	1.46
	(2710)**	(2,448)**	(2,416)**	(2,297)**	(0.451)**	(0.417)**	(0.417)**	(0.411)**
Birth weight < 1500g * Grams from cutoff		-17.30	-31.76	6.17		-0.0010	-0.0014	-0.0058
		(37.00)	(36.47)	(34.63)		(0.0065)	(0.0065)	(0.0064)
Birth weight >= 1500g * Grams from cutoff		-73.30	-86.84	-79.51		-0.0231	-0.0238	-0.0260
		(30.18)*	(29.78)**	(28.23)**		(0.0052)**	(0.0053)**	(0.0052)**
Year controls		No	Yes	Yes		No	Yes	Yes
Main controls		No	No	Yes		No	No	Yes
Mean of dependent variable above cutoff:	81566				24.68			
Observations	28928				30935			

Notes: Local linear regressions use a bandwidth of 3 ounces (85 grams). OLS models estimated on a sample within 3 ounces above and below the VLBW threshold. Five states include AZ, CA, MD, NY, and NJ (various years). "Main controls" are listed in Table A1, as well as indicators for each year. Some observations have missing charges, as described in the text. * significant at 5%; ** significant at 1%. Robust standard errors in parentheses.

Table 3: Specific treatment measures by very low birth weight status: Five-state sample, 1991-2006

Dependent variable:	Ventilation (various methods)			California: >24 hours in NICU		
Model:	local linear	OLS	OLS	local linear	OLS	OLS
Birth weight < 1500g	0.0357	0.0380	0.0274	0.0372	0.0282	0.0265
	(0.0125)**	(0.0115)**	(0.0112)*	(0.0170)*	(0.0157)	(0.0156)
Birth weight < 1500g * Grams from cutoff		0.0001	0.00001		0.0003	0.0003
		(0.0002)	(0.0002)		(0.0002)	(0.0002)
Birth weight >= 1500g * Grams from cutoff		-0.0001	-0.0001		0.0003	0.0003
		(0.0002)	(0.0001)		(0.0002)	(0.0002)
Year controls		No	Yes		No	Yes
Main controls		No	Yes		No	Yes
Mean of dependent variable above cutoff:	0.511			0.444		
Observations	30935			16528		

Dependent variable:	Diagnostic ultrasound of infant			Operations on the heart		
Model:	local linear	OLS	OLS	local linear	OLS	OLS
Birth weight < 1500g	0.0196	0.0166	0.0297	0.0147	0.0155	0.0236
	(0.0109)	(0.0101)	(0.0095)**	(0.0112)	(0.0104)	(0.0100)*
Birth weight < 1500g * Grams from cutoff		0.00004	0.00012		-0.0000	0.0001
		(0.00015)	(0.00014)		(0.0002)	(0.0001)
Birth weight >= 1500g * Grams from cutoff		-0.00006	0.00018		-0.0003	-0.0002
		(0.00013)	(0.00012)		(0.0001)*	(0.0001)
Year controls		No	Yes		No	Yes
Main controls		No	Yes		No	Yes
Mean of dependent variable above cutoff:	0.244			0.260		
Observations	30935			30935		

Notes: Local linear regressions use a bandwidth of 3 ounces (85 grams). OLS models estimated on a sample within 3 ounces above and below the VLBW threshold. Five states include AZ, CA, MD, NY, and NJ (various years). "Main controls" are listed in Table A1, as well as indicators for each year. The dependent variable in the NICU models is only available in our California data, and equals one if the infant spent more than 24 hours in a NICU or died in the NICU at less than 24 hours. * significant at 5%; ** significant at 1%. Robust standard errors in parentheses.

Table 4: Specification and robustness checks

Dependent variable:	hospital costs	hospital costs	log(hospital costs)	log (hospital charges)	median regression hospital charges	log(length of stay)
Birth weight < 1500g	4206	3812	0.263	0.282	9415	0.140
	(1068)**	(1033)**	(0.106)**	(0.037)**	(1593)**	(0.0272)**
Sample	5-State	5-State	5-State	5-State	5-State	5-State
Controls	No	Yes	Yes	Yes	Yes	Yes
Mean of dependent variable above cutoff:	39628	39628	9.91	10.58	81566	2.78
Observations	28769	28769	28769	28769	28928	30935

Dependent variable:	hospital transfer	hospital charges including transfers	hospital costs including transfers	log(charges) including transfers	log(length of stay) including transfers
Birth weight < 1500g	-0.011	7297	2872	0.223	0.1088
	(0.0067)	(4313)	(1776)	(0.045)**	(0.0319)**
Sample	5-State	California	California	California	California
Controls	Yes	Yes	Yes	Yes	Yes
Mean of dependent variable above cutoff:	0.100	109421	45141	11.0	2.99
Observations	30935	14560	14560	14560	16528

Notes: All models are OLS, estimated on a sample within 3 ounces above and below VLBW threshold. All models include the gram-trend variables and our "main controls," which are listed in Table A1, as well as indicators for each year. Charges are in $2006. Some observations have missing or zero charges, as described in the text. Five states include AZ, CA, MD, NY, and NJ (various years). * significant at 5%; ** significant at 1%. Robust standard errors in parentheses.

Table 5: Results by year and for overlap of NCHS/five-state data

A. NCHS Nationwide Data

Dependent variable: 1-year mortality	Years:				In-hospital births only	
					Five states, all NCHS years	Five states, years in NCHS and multi-state data
	1983-1987	1988-1991	1995-1998	1999-2002	1983-2002 (available years)	1991; 1995-2002
Birth weight < 1500g	-0.0144	-0.0077	-0.00081	-0.0069	-0.0122	-0.0074
	(0.0051)**	(0.0048)	(0.0038)	(0.0035)*	(0.0042)**	(0.0051)
Mean of dependent variable above cutoff:	0.0813	0.0622	0.0410	0.0378	0.054	0.039
Observations	50947	47545	49989	53590	49839	23698

B. Multi-State Sample

Dependent variable: hospital charges	hospital charges	hospital charges	hospital charges	In-hospital births only	
				hospital charges	hospital costs
	Years:			Years in NCHS and multi-state data	
	1991-1994	1995-1998	1999-2002	2003-2006	1991; 1995-2002	1991; 1995-2002
Birth weight < 1500g	12055	3515	16985	582	10108	4553
	(4,538)**	(3,167)	(4,930)**	(6,151)	(2738)**	(1242)**
Mean of dependent variable above cutoff:	69566	71392	93717	96124	80721	39946
Observations	5018	10711	9504	3695	21479	21479

Notes: All models are OLS, estimated on a sample within 3 ounces above and below VLBW threshold. All models include the gram-trend variables. The first four columns include our "main controls," which vary by the sample used and are described in the notes in the previous tables. The last two columns include common covariates across samples: indicators for whether the baby is male, preterm, black,"other" race, a twin, or a non-twin multiple birth, as well as state indicators and year indicators. Although in theory the births included in the NCHS birth records in the state-years available in our multi-state sample should be the same as the births included in the multi-state sample, in practice the samples are slightly different (as evidenced by the difference in sample size), largely due to 300-400 fewer births in the discharge data in each year from 2000-2002. Some observations have missing charges, as described in the text. * significant at 5%; ** significant at 1%. Robust standard errors in parentheses.

Table 6: Bandwidth sensitivity

A. NCHS Nationwide Data

Dependent variable:	**1-year Mortality**				
Bandwidth	30	60	90	120	150
Birth weight < 1500g	-0.0267	-0.0162	-0.0114	-0.00911	-0.00865
	(0.00382)**	(0.00269)**	(0.0022)**	(0.00190)**	(0.00170)**
Mean of dependent variable above cutoff:	0.0607	0.0562	0.0545	0.0532	0.0515
Observations	72937	163415	233880	304630	376400

Dependent variable:	**28-Day Mortality**				
Bandwidth	30	60	90	120	150
Birth weight < 1500g	-0.0228	-0.0146	-0.0101	-0.00828	-0.00773
	(0.00322)**	(0.00227)**	(0.00185)**	(0.00160)**	(0.00143)**
Mean of dependent variable above cutoff:	0.0431	0.0390	0.0377	0.0367	0.0352
Observations	72937	163415	233880	304630	376400

B. 5-State Sample

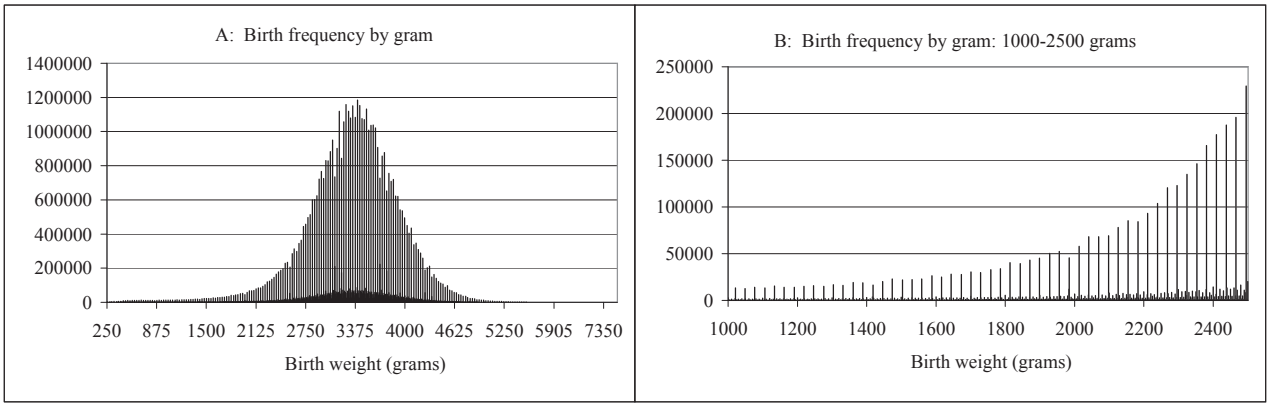
Dependent variable:	**Hospital Charges**				
Bandwidth	30	60	90	120	150
Birth weight < 1500g	7670	8380	9290	8070	6490
	(4300)*	(3210)**	(2630)**	(2270)**	(2030)**
Mean of dependent variable above cutoff:	83890	81527	80235	79092	77158
Observations	10533	21404	31990	42012	52471

Dependent variable:	**Hospital Costs**				
Bandwidth	30	60	90	120	150
Birth weight < 1500g	4460	3620	3970	3410	2580
	(1880)*	(1390)**	(1140)**	(985)**	(881)**
Mean of dependent variable above cutoff:	41063	39321	38572	38028	37094
Observations	10533	21404	31990	42012	52471

Dependent variable:	**Length of Stay**				
Bandwidth	30	60	90	120	150
Birth weight < 1500g	2.38	1.84	1.91	1.53	1.14
	(0.743)**	(0.536)**	(0.439)**	(0.379)**	(0.340)**
Mean of dependent variable above cutoff:	25.7	24.8	24.3	24.0	23.5
Observations	11254	22877	34183	44868	56067

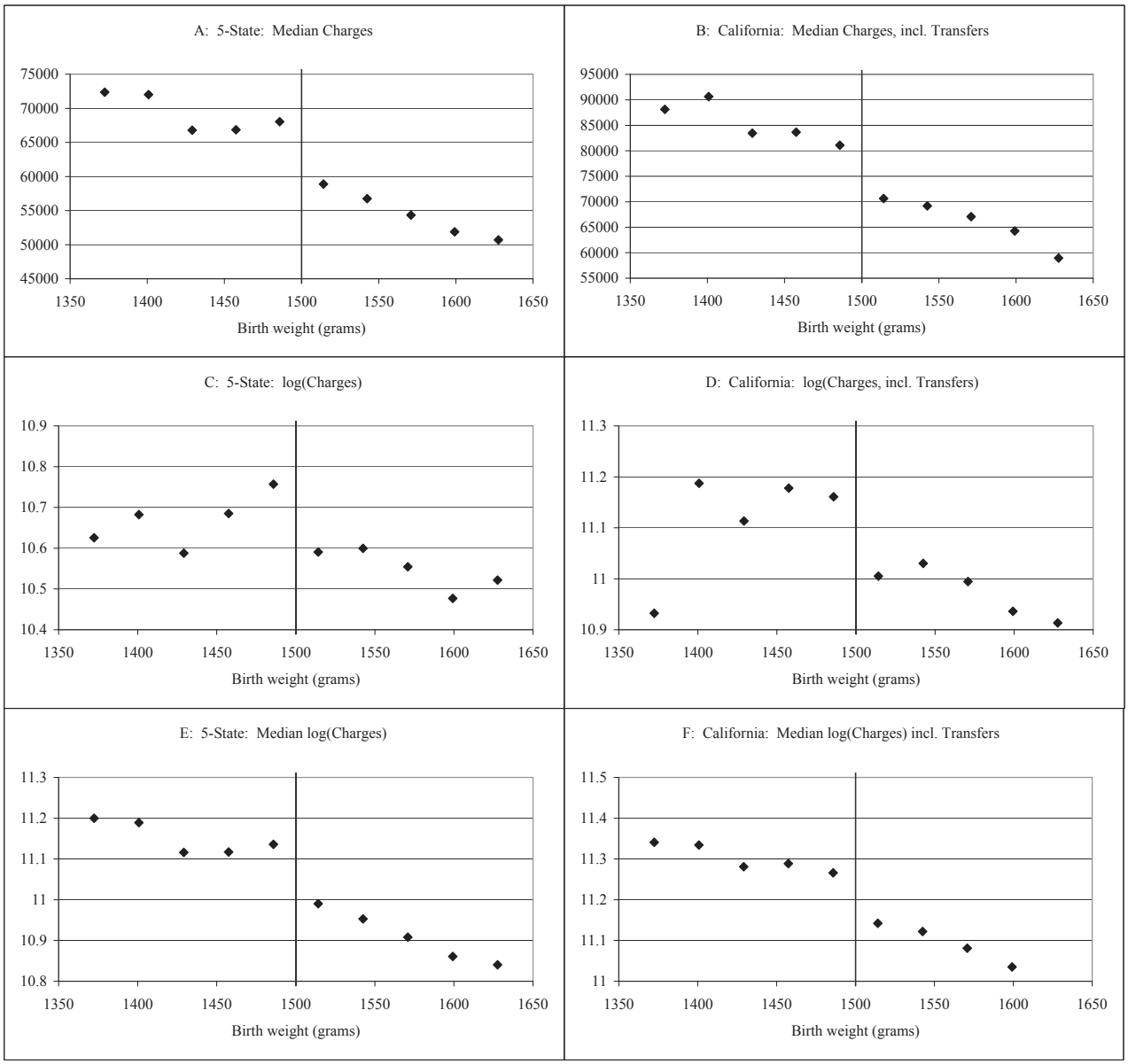
Notes: All models are OLS, estimated on a sample within 3 ounces above and below VLBW threshold. All models include the gram-trend variables and our "main controls," which vary by the sample used and are described in the notes in the previous tables. Some observations have missing charges, as described in the text. * significant at 5%; ** significant at 1%. Robust standard errors in parentheses.

Figure A1: Birth frequencies for wider bandwidths



Notes: NCHS birth cohort linked birth/infant death files, 1983-1991 and 1995-2003, as described in the text.

Figure A2: Alternative summary treatment measures around 1500 grams



Notes: Data are all births in the 5-state sample (AZ, CA, MD, NY, and NJ), as described in the text. Some observations have missing or zero charges, as described in the text. Charges are in 2006 dollars. Points represent means in one-ounce bins radiating from 1500 grams.

Table A1: Coefficients on selected covariates

	Sample: NCHS Nationwide Data Dependent variable: 1-year mortality		Multi-State Sample hospital charges
Birth weight < 1500g	-0.0073 (0.0022)**	Birth weight < 1500g	9065 (2,297)**
Birth weight < 1500g * Grams from cutoff	-0.00011 (0.0000)**	Birth weight < 1500g * Grams from cutoff	6.175 -34.631
Birth weight >= 1500g * Grams from cutoff	-0.00018 (0.0000)**	Birth weight >= 1500g * Grams from cutoff	-79.513 (28.23)**
Prenatal Visits: 9-14	-0.0041 (0.0012)**	=1 if newborn is male and not missing	11611 (1,145)**
Prenatal Visits: >=15	-0.0028 (0.0017)	Pre-term birth	22958 (1,688)**
Mother born outside state	-0.0011 (0.0011)	Mother's Race/Ethnicity: African American	1827 -1481
First birth	0.0193 (0.0012)**	Mother's Race/Ethnicity: Other	4533 (1,600)**
Mother's Age: 31-35 (compared to <16)	-0.0132 (0.0049)**	Twin birth	3405 (1,346)*
Mother's Age: 36-40	-0.0122 (0.0051)*	Multiple (non-twin) birth	11835 (2,354)**
Mother's Age: 41+	-0.0011 (0.0065)	Cesarean Section	2770 (1,199)*
Mother's Education: High School	0.0001 (0.0015)	Arizona (compared to NJ)	-1653 -2484
Mother's Education: Some College	-0.0029 (0.0017)	California	101580 (1,805)**
Mother's Education: College+	0.0032 (0.0019)	New Jersey	87235 (1,608)**
Moher's Education: missing	0.0176 (0.0028)**	New York	60591 (1,500)**
Father's Age: 31-35 (compared to <16)	-0.0013 (0.0187)	Year = 1991 (compared to 2003)	-92968 (4,690)**
Father's Age: 36-40	-0.0016 (0.0188)	Year = 2006	3937 -4694
Father's Age: 41+	-0.0044 (0.0188)	Constant	31557 (4,237)**
Father's Age: missing	0.0021 (0.0187)		
Male	0.0144 (0.0010)**		
Gestational Age: 37 weeks (compared to <31)	0.0252 (0.0038)**		
Gestational Age: 40 weeks	0.0121 (0.0053)*		
Gestational Age: 41 weeks	0.0118 (0.0069)		
Mother's Race/Ethnicity: African American	-0.0189 (0.0014)**		
Mother's Race/Ethnicity: Hispanic	-0.0034 (0.0019)		
Singleton birth	0.0446 (0.0018)**		
Twin birth	0.0114 (0.0018)**		
Year = 2002 (compared to 1984)	-0.0356 (0.0035)**		
Constant	0.0470 (0.0243)		
Mean of dependent variable above cutoff:	0.0553		81566.3098
Observations	202071		28928

Notes: All models are OLS, estimated on a sample within 3 ounces above and below VLBW threshold. Charges are in $2006. Some observations have missing charges, as described in the text. Five states include AZ, CA, MD, NY, and NJ (various years) * significant at 5%; ** significant at 1%. Robust standard errors in parentheses.

Table A2: Selected covariate comparison (controlling for trends in birth weight)

	adjusted mean below threshold	raw mean above threshold	p-value
Fewer than 9 Prenatal visits	0.478	0.443	(0.000)**
First birth	0.4145	0.4145	(0.860)
Mother's Age	26.41	26.44	(0.122)
Mother's Education: <High School	0.2515	0.2495	(0.158)
Mother's Education: High School	0.3385	0.3365	(0.195)
Mother's Education: Some College	0.1738	0.1728	(0.574)
Mother's Education: College+	0.1466	0.1506	(0.001)**
Mother's Education: Missing	0.0896	0.0906	(0.465)
Mother born outside state	0.4116	0.4076	(0.217)
Mother's Race: White	0.4426	0.4506	(0.000)**
Mother's Race: African American	0.2504	0.2524	(0.524)
Mother's Race: Hispanic	0.1315	0.1275	(0.058)
Father's Age	29.85	29.86	(0.646)
Missing Father's Age	0.245	0.242	(0.076)
Father's Education: <High School	0.1276	0.1266	(0.459)
Father's Education: High School	0.2636	0.2656	(0.371)
Father's Education: Some College	0.1019	0.0999	(0.287)
Father's Education: College+	0.1066	0.1086	(0.243)
Father's Education: Missing	0.4003	0.3993	(0.698)
Male	0.4983	0.5003	(0.574)
Gestational Age	31.73	32.62	(0.000)**
Singleton Birth	0.7357	0.7437	(0.006)**
Twin Birth	0.2074	0.2214	(0.000)**
Multiple (non-twin) Birth	0.0349	0.0349	(0.969)
Vaginal Birth	0.4504	0.4754	(0.000)**
Obstetric Procedures: Amnioscentesis	0.0480	0.0510	(0.004)**
Obstetric Procedures: Induction	0.0874	0.1004	(0.000)**
Obstetric Procedures: Stimulation	0.0604	0.0664	(0.000)**
Obstetric Procedures: Tocolysis	0.1210	0.1150	(0.000)**
Obstetric Procedures: Ultrasound	0.6454	0.6484	(0.091)
Obstetric Procedures: Other	0.0624	0.0634	(0.151)
Year of Birth	1992.97	1993.00	(0.110)
Predicted 1-year Mortality	0.0586	0.0576	(0.000)**

Notes: Sample is NCHS national data. For most covariates, the number of observations is 341,140. Delivery method is available for 229,843 births; obstretric procedures are available for 229,175 births. * significant at 5%; ** significant at 1%.

Table A3: Five-state sample: Data summary

Year	AZ	CA	MD	NJ	NY	Total
1991	0	1,430	0	0	0	1,430
1992	0	1,428	0	0	0	1,428
1993	0	1,346	0	0	0	1,346
1994	0	1,410	0	0	0	1,410
1995	0	1,365	251	433	921	2,970
1996	0	1,400	232	372	797	2,801
1997	0	1,317	212	408	838	2,775
1998	0	1,380	211	412	772	2,775
1999	0	1,333	259	649	882	3,123
2000	0	1,387	237	395	842	2,861
2001	138	1,380	245	383	0	2,146
2002	176	1,352	249	393	0	2,170
2003	184	0	271	404	0	859
2004	262	0	250	409	0	921
2005	271	0	249	385	0	905
2006	325	0	293	397	0	1,015
Total	1,356	16,528	2,959	5,040	5,052	30,935

Notes: Table displays years for which each of our state data sets are available, and the relevant sample sizes for births within 3 ounces of 1500 grams.

Table A4: One-year mortality results by cause of death

	one-year mortality, by cause										
Dependent variable:	all cause	infectious and parasitic diseases	neoplasms	endocrine, nutritional, metabolic, immunity, blood disorders	nervous system, sense organ disorders	respiratory system disorders	digestive system disorders	congenital anomalies	perinatal conditions	symptoms, signs, ill-defined conditions	other
Model:	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS
Birth weight < 1500g	-0.0095	-0.0031	0.0009	0.0029	-0.0131	0.0017	-0.0006	0.0091	-0.0294	0.0173	0.0144
	(0.0022)**	(0.0055)	(0.0024)	(0.0031)	(0.0050)**	(0.0070)	(0.0050)	0.0191	(0.0188)	(0.0108)	(0.0093)
Mean of dependent variable above cutoff:	0.0553	0.0225	0.0038	0.0055	0.0177	0.0339	0.0161	0.4024	0.3567	0.0831	0.0582
Observations	202071	11090	11090	11090	11090	11090	11090	11090	11090	11090	11090

	one-year mortality, by cause				
Dependent variable:	"external" cause	respiratory distress syndrome (RDS)	sudden infant death syndrome (SIDS)	jaundice	meningitis
Model:	OLS	OLS	OLS	OLS	OLS
Birth weight < 1500g	0.0045	-0.0054	0.0148	-0.0034	-0.0016
	(0.0048)	(0.0109)	(0.0095)	(0.0018)	(0.0030)
Mean of dependent variable above cutoff:	.0141	.0878	.0608	.0022	0.0052
Observations	11090	11090	11090	11090	11090

Notes: The ten cause of death classifications (other than all cause mortality) in the first row were constructed to be categories which could be defined consistently over time, across a change in cause of death coding which occurs partway through our sample; these broad categories partition non-missing causes of death. The second row extracts some individual causes of death from these broad categories. We exclude observations with missing information on the timing or cause of death. OLS models estimated on a sample within 3 ounces above and below the VLBW threshold. All models include the gram-trend variables. * significant at 5%; ** significant at 1%. Robust standard errors in parentheses.