ESTIMATING MARGINAL RETURNS TO MEDICAL CARE: EVIDENCE FROM AT-RISK NEWBORNS*

DOUGLAS ALMOND
JOSEPH J. DOYLE, JR.
AMANDA E. KOWALSKI
HEIDI WILLIAMS

A key policy question is whether the benefits of additional medical expenditures exceed their costs. We propose a new approach for estimating marginal returns to medical spending based on variation in medical inputs generated by diagnostic thresholds. Specifically, we combine regression discontinuity estimates that compare health outcomes and medical treatment provision for newborns on either side of the very low birth weight threshold at 1,500 grams. First, using data on the census of U.S. births in available years from 1983 to 2002, we find that newborns with birth weights just below 1,500 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 1,500 grams from above, which is large relative to mean infant mortality of 5.5% just above 1,500 grams. Second, using hospital discharge records for births in five states in available years from 1991 to 2006, we find that newborns with birth weights just below 1,500 grams have discontinuously higher charges and frequencies of specific medical inputs. Hospital costs increase by approximately $4,000 as birth weight crosses 1,500 grams from above, relative to mean hospital costs of $40,000 just above 1,500 grams. Under an assumption that observed medical spending fully captures the impact of the "very low birth weight" designation on mortality, our estimates suggest that the cost of saving a statistical life of a newborn with birth weight near 1,500 grams is on the order of $550,000 in 2006 dollars.

*We thank Christine Pal and Jean Roth for assistance with the data, Christopher Afendulis and Ciaran Phibbs for data on California neonatal intensive care units, and doctors Christopher Almond, Burak Alsan, Munish Gupta, Chafen Hart, and Katherine Metcalf for helpful discussions regarding neonatology. David Autor, Amitabh Chandra, Janet Currie, David Cutler, Dan Fetter, Amy Finkelstein, Edward Glaeser, Michael Greenstone, Jonathan Gruber, Jerry Hausman, Guido Imbens, Lawrence Katz, Michael Kremer, David Lee, Ellen Meara, Derek Neal, Joseph Newhouse, James Poterba, Jesse Rothstein, Gary Solon, Tavneet Suri, the editor and four referees, and participants in seminars at Harvard, the Harvard School of Public Health, MIT, Princeton, 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, the Maryland Health Services Cost Review Commission, the New Jersey Department of Health and Senior Services, and the New York State Department of Health. Funding from the National Institute on Aging, Grant T32-AG000186 to the National Bureau of Economic Research, is gratefully acknowledged (Doyle, Kowalski, Williams).

The Quarterly Journal of Economics, May 2010

I. Introduction

Medical expenditures in the United States are high and increasing. Do the benefits of additional medical expenditures exceed their costs? The tendency for patients in worse health to receive more medical inputs complicates empirical estimation of the returns to medical expenditures. Observational studies have used cross-sectional, time-series, and panel data techniques to attempt to identify patients who are 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 may not be surprising, as these studies have measured returns on many different margins of care. The return to a dollar of medical spending likely differs across medical technologies and across patient populations, and in any given context the return to the first dollar of medical spending likely differs from the return to the last dollar of 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. Although 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 of this paper is a novel research design that, under explicit assumptions, permits direct estimation of the marginal returns to medical care. Implementation of our research design requires a setting with an observable, continuous

1. See, for example, McClellan (1997); Cutler et al. (1998); Cutler and McClellan (2001); Nordhaus (2002); Murphy and Topel (2003); Cutler, Rosen, and Vijan (2006); and Luce et al. (2006).

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

measure of health risk and a diagnostic threshold (based on this risk variable) that generates a discontinuous probability of receiving additional treatment.[3] In such 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 care 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. Under the assumption that observed medical spending fully captures the impact of a "higher risk" designation on mortality, combining these cost and benefit estimates 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, the welfare implications of 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. 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 and 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 childbirth is one of the most common reasons for hospital admission in the United States. This patient population also provides samples large enough to detect effects of additional treatment on infant mortality.

3. Such criteria are common in clinical medicine. For example, diabetes diagnoses are frequently made based on a threshold fasting glucose level, hypertension diagnoses based on a threshold systolic blood pressure level, hypercholesterolemia diagnoses based on a threshold cholesterol level, and overweight diagnoses based on a threshold body mass index. Nevertheless, there is "little evidence" that the regression discontinuity framework has been used to evaluate triage criteria in clinical medicine (Linden, Adams, and Roberts 2006). Similarly, Zuckerman et al. (2005, p. 561) note that "program evaluation in health services research has lacked a formal application" of the regression discontinuity approach.

We focus on the "very low birth weight" (VLBW) classification at 1,500 g (just under 3 pounds, 5 ounces)—a designation frequently referenced 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 (p. 23), for example, notes that "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 1,500 g to being just below 1,500 g. Thus, although we empirically investigate our assumption that the position of a newborn just above 1,500 g relative to just below 1,500 g is "as good as random," the medical literature also suggests that this assumption is reasonable.

To preview our main results, using data on the census of U.S. births in available years from 1983 to 2002, we find that one-year mortality decreases 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 1,500 g. This sharply contrasts with the overall increase in mortality as birth weight falls, and to the extent that lighter newborns are less healthy in unobservable ways, the mortality change we observe is all the more striking. Second, using hospital discharge records for births in five states in available years from 1991 to 2006, we estimate a $4,000 increase in hospital costs for infants just below the 1,500-g threshold, relative to mean hospital costs of $40,000 just above 1,500 g. As we discuss in Section VIII, this estimated cost difference may not capture all of the relevant mortality-reducing inputs, but it is our best available summary measure of health inputs. Under the assumption that hospital costs fully capture the impact of the VLBW designation on mortality, our estimates suggest that the cost of saving a statistical life for newborns near 1,500 g is on the order of $550,000—well below most value-of-life estimates for this group of newborns.

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 and gives a brief

background on the at-risk newborn classifications we study. Section III describes our data and analysis sample, and Section IV outlines our empirical framework and bandwidth selection. Section V presents our main results, and Section VI discusses several robustness and specification checks. Section VII examines variation in our estimated treatment effects across hospital types. In Section VIII we combine our main estimates to calculate two-sample estimates of marginal returns, and Section IX concludes.

II. Background

II.A. Costs and Benefits of Medical Care for At-Risk Newborns

Medical treatments for at-risk newborns have been expanding tremendously in recent years, at high cost. For example, in 2005 the U.S. 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 United States in 2001, preterm and low–birth weight diagnoses accounted for 8% of newborn admissions, but 47% of costs for all infant hospitalizations (at $15,100 on the average). Despite their high and highly concentrated costs, use of new neonatal technologies has continued to expand.[5]

These high costs 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 that 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; Goodman et al. 2002; Grumbach 2002). On the other hand, Cutler and 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.[6]

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

5. 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).

6. 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

II.B. "At-Risk" Newborn Classifications

Birth weight and gestational age are the two most common metrics of newborn health, and continuous measures of these variables are routinely collapsed into binary classifications. We focus on the VLBW classification at 1,500 g (just under 3 pounds, 5 ounces). We also examine other birth weight classifications—including the "extremely low birth weight" (ELBW) classification at 1,000 g (just over 2 pounds, 3 ounces) and the "low birth weight" (LBW) classification at 2,500 g (just over 5 pounds, 8 ounces)—as well as gestational age–based measures such as the "prematurity" classification at 37 weeks, where gestation length is usually based on the number of weeks since the mother's last menstrual period. Below, we briefly describe the evolution of these classifications.[7]

Physicians had begun to recognize and assess the relationships among inadequate growth (LBW), shortened gestation (prematurity), and mortality by the early 1900s. The 2,500-g LBW classification, for example, has existed since at least 1930, when a Finnish pediatrician advocated 2,500 g as the birth weight below which infants were at high risk of adverse neonatal outcomes. Over time, interest increased in the fate of the smallest infants, and "very low birth weight" infants were conventionally defined as those born weighing less than 1,500 g (United States Institute of Medicine 1985).[8]

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 (p. 22):

> 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

medical care in the immediate postnatal period (Williams and Chen 1982; Paneth 1995). 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). For comparison to the results obtained with our methodology, we present results based on the Cutler and Meara methodology in our data in Section VIII.A.

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

8. In our empirical work, 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, VLBW is defined as having birth weight strictly less than 1,500 g, and analogously (with a strict inequality) for the other thresholds we examine.

> 2,501 grams on the basis of birth weight alone ... As with the 2,500 grams limit, designation of very low birth weight infants as those weighing 1,500 grams or less reflected convention rather than biologic criteria.

Gestational age classifications, such as the "prematurity" classification at 37 gestational weeks, have also been emphasized. Although gestational age is a natural consideration when determining treatment for newborns with low birth weights, applying our research design to gestational age introduces some additional complications. Gestational age is known to women in advance of giving birth, and women can choose to time their births (for example, through an induced vaginal birth or through a C-section) based on gestational age. Thus, 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 1,500 g to being just below 1,500 g; this assertion has been confirmed from conversations with physicians,[9] as well from studies such as Pressman et al. (2000).

Of course, birth weight and gestational age are not the only factors used to assess newborn health.[10] 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 1,499 to 1,501 g, but rather expect a change in the likelihood of treatment for newborns classified into a given risk category.

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 and 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.

10. 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 and Stark 1998).

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. From an empirical perspective, the fact that we will observe a discontinuity in treatment provision around 1,500 g suggests that hospitals or physicians do use these cutoffs to determine treatment either through hospital protocols or as rules of thumb. As an example of a relevant hospital protocol, the 1,500 g threshold is commonly cited as a point below which diagnostic ultrasounds should be used.[11] Such classifications could also affect treatment provision through use as more informal "rules of thumb" by physicians.[12] As we discuss below, it is likely that VLBW infants receive a bundle of mortality-reducing health inputs, not all of which we can observe.[13] Moreover, because several procedures are given simultaneously, our research design does not allow us to measure marginal returns to specific procedures. This motivates our focus on summary measures—such as charges and length of stay—that are our best available measures of differences in health inputs.

Differential reimbursement by birth weight is another potential source of observed discontinuities in summary spending measures. For example, some Current Procedural Terminology (CPT) billing codes and ICD-9 diagnosis codes are categorized by birth weight (ICD-9 codes V21.30–V21.35 denote birth weights of 0–500, 500–999, 1,000–1,499, 1,500–1,999, 2,000–2,500, etc.). If prices differ across our threshold of interest, then any discontinuous jump in charges could in part be due to mechanical changes in the "prices" of services rather than to changes in the "quantities" of the services performed. In practice, we argue that a

11. 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 and Stark 1998, p. 508). We investigate differences in the use of diagnostic ultrasounds and other procedures below.

12. For a recent contribution on this point in the economics literature, see Frank and Zeckhauser (2007). In the medical literature, see McDonald (1996) and Andre et al. (2002). Medical malpractice environments could also be one force affecting adherence to either formal rules or informal rules of thumb.

13. A recent review article (Angert and Adam 2009) on care for VLBW infants offered several examples of health inputs that we would likely not be able to detect in our hospital claims data. For example, the authors note (p. 32): "To decrease the risk for intraventricular hemorrhage and brain injury during resuscitation, the baby should be handled gently and not placed in a head down or Trendelenburg position."

substantial portion of our observed jump in charges is a quantity effect rather than a price effect, for three reasons. First, the limited qualitative evidence available to us suggests that prices do not vary discontinuously across the VLBW threshold for many of the births in our data.[14] Second, we empirically observe a discontinuity in charges within California, a state where the Medicaid reimbursement scheme does not explicitly utilize birth weight during the time period of our study. Third, we find evidence of discontinuities in a summary quantity measure—length of stay—as well as quantities of specific procedures. These three reasons suggest that a substantial portion of our observed jump in charges is a "quantity" effect rather than a "price" effect. Furthermore, if the pricing effect were purely mechanical, we should not observe the empirical discontinuity in mortality.

III. Data

III.A. Data Description

Our empirical analysis requires data 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 United States for the years 1983–1991 and 1995–2002—approximately 66 million births.[15] The data include information reported on birth certificates linked to information reported on death certificates for infants who die

14. We unfortunately do not observe prices directly in any of our data sets. 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 either a *per diem* system or the CMS-DRG system, neither of which explicitly utilizes birth weight. 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.

15. NCHS did not produce linked birth/infant death files from 1992 to 1994.

within one year of birth. The birth certificate data offer a rich set of covariates (for example, mother's age and education), and the death certificate data include 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) more 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 and Doyle (2008). These data were collected by the California Office of Statewide Health Planning and Development and include all live births in California from 1991 to 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 diagnosis, 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 are admitted to 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.[16]

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.[17] 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 Table A1 in the Online Appendix for the number of births by state and year within our pilot bandwidth of three ounces of the VLBW cutoff). 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 the HCUP data with mortality records directly, we can examine mortality outcomes

16. The treatment measures that include transfers described below include treatment at the birth hospital and the hospital where the newborn was initially transferred.

17. 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 subsample of the SID data.) At present, 39 states participate in the SID. Of these 39 states, 10 report the birth weights of newborns. We have obtained HCUP data for 4 of the 10 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.

for these newborns using a subsample of our nationwide data, as our nationwide data and the HCUP discharge data relate to the same births.[18] In much of our analysis, we pool the California and HCUP data to create a "five-state sample."

Both the California and the HCUP data report hospital charges. These charges are used in negotiations for reimbursement and are typically inflated well over costs. We consider these charges our best available summary of the difference in treatment that the VLBW classification affords. When calculating the returns to medical spending, we adjust hospital charges by a cost-to-charge ratio.[19] The main text focuses on charges rather than costs because charges are available for all years of data, whereas cost-to-charge ratios are available for only a subset of years and are known to introduce noise into the results.

III.B. Analysis Sample

Sample selection issues are minimal. In our main specifications, we pool data from all available years, although in the Online Appendix we separately examine results across time periods. For the main results, we limit the sample to those observations with nonmissing, nonimputed birth weight information.[20] Fortunately, given the demands of our empirical approach, these data provide relatively large samples: over 200,000 newborns fall within our pilot bandwidth of three ounces around the 1,500-g threshold in the nationwide data, and we have approximately 30,000 births in the same interval when we consider the five-state sample. We discuss bandwidth selection below.

18. 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 hospitals. In some robustness checks we limit our nationwide data to the sample of hospitalized births, for greater comparability.

19. The Centers for Medicare and Medicaid Services (CMS) report cost-to-charge ratios for each hospital in each year beginning in 1996 and continuing through 2005. When we use the cost-to-charge ratios, so that we can include information from all years, we use the 2000 cost-to-charge ratios 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 follow 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 was employed.

20. 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.

IV. Empirical Framework and Estimation

IV.A. Empirical Framework

To estimate the size of the discontinuity in outcomes and treatment, we follow standard methods for regression discontinuity analysis (as in, for example, Imbens and Lemieux [2008] and Lee and Lemieux [forthcoming]).

First, we restrict the data to a small window around our threshold (85 g) and estimate a local-linear regression. We describe the selection of this bandwidth in the next section. 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, Fan, and Marron 1997; Porter 2003).[21]

Second, within our bandwidth, we estimate the following model for infant i weighing g grams in year t:

$$(1)\quad Y_i = \alpha_0 + \alpha_1 \text{VLBW}_i + \alpha_2 \text{VLBW}_i \times (g_i - 1500) + \alpha_3(1 - \text{VLBW}_i) \times (g_i - 1500) + \alpha_t + \alpha_s + \delta X_i' + \epsilon_i,$$

where Y is an outcome or treatment measure such as one-year mortality or costs, and VLBW is an indicator that the newborn was classified as VLBW (that is, strictly less than 1,500 g). We include separate gram trend terms above and below the cutoff, parameterized so that $\alpha_2 = \alpha_3$ if the trend is the same above and below the cutoff. 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, and we report two sets of standard errors.[22] First, we report heteroscedastic-robust standard errors. Second, to address potential concerns about discreteness in birth weight, we perform the standard error correction suggested by Card and Lee (2008). In our application, this

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

22. Probit results for our binary dependent variables give very similar results, as described below.

correction amounts to clustering at the gram level. Estimation of our outcome and treatment results with quadratic (or higher-order) rather than linear trends in birth weight gives similar results (see Online Appendix Table A4).

In Section V, we report outcome and treatment estimates separately. Our reduced-form estimate of the direct impact of our VLBW indicator on mortality is itself interesting and policy relevant, as this estimate includes the effects of all relevant inputs.

Under an additional assumption, we can combine our outcome and treatment estimates into two-sample estimates of the return to an increment of additional spending in terms of health benefits. In the language of instrumental variables, the discontinuity in mortality is the reduced-form estimate and the discontinuity in health inputs is the first-stage estimate.[23] In this framework, the instrument is the VLBW indicator. For our VLBW indicator to be a valid instrument, the two usual instrumental variables conditions must hold. First, there must exist a first-stage relationship between our VLBW indicator and our measure of health inputs; note that this relationship will be conditional on our running variable (birth weight). Second, the exclusion restriction requires that the only mechanism through which the instrument VLBW affects the mortality outcome, conditional on birth weight falling within the bandwidth, is through its effect on our measure of health inputs. If our summary measures allow us to observe and capture all relevant health inputs, then we can argue for the validity of this exclusion restriction. However, for any given measure of health inputs that we observe, it is likely that there exists some additional health-related input that we do not observe (see Section II.B). It is unclear how important such unobserved inputs are in practice, but to the extent that they are important, a violation of the exclusion restriction would occur.

We present two-sample estimates in Section VIII, using our most policy-relevant available summary measure of treatment (hospital costs) as our first-stage variable, but we stress that the interpretation of these estimates relies on an assumption that

23. 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.

hospital charges capture all relevant medical inputs. We also attempt to gauge the magnitude of unobserved inputs by testing for effects on short-run mortality. To the extent that medical inputs are much more important than parental or other unobserved inputs in the very short run after birth (say, within 24 hours of birth), we can test for impacts on short-run mortality measures and be somewhat assured that unobserved parental or other 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.

IV.B. Bandwidth Selection

Our pilot bandwidth includes newborns with birth weights within 3 ounces (85 g) of 1,500 g, or from 1,415 to 1,585 g. We chose this bandwidth by a cross-validation procedure where the relationships between the main outcomes of interest and birth weight were estimated with local linear regressions and compared to a fourth-order polynomial model. These models were estimated separately above and below the 1,500-g threshold. The bandwidth that minimized the sum of squared errors between these two estimates between 1,200 and 1,800 g tended to be between 60 and 70 g for the mortality outcomes. For the treatment measures, the bandwidth tended to be closer to 40 g. 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 g—three ounces[24]—for the main results. This larger bandwidth incorporates more information, which 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 Online Appendix Table A3). To give a clearer sense of how our data look graphically, our figures report means for a slightly wider bandwidth—namely, the five ounces above and below the threshold.

24. 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.

V. Results

V.A. Frequency of Births by Birth Weight

Figure I is a histogram of births between 1,350 and 1,650 g in the nationwide sample, which has several notable characteristics.[25] 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 100), 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 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 that have been collapsed into one-ounce bins.[26]

Second, we do not observe irregular reporting heaps around our 1,500-g threshold of 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 1,500 g to just below 1,500 g. 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 1,500-g cutoff and justify higher reimbursements. In particular, the frequency of births at 1,500 g is very similar to the frequency of births at 1,400 g and at 1,600 g, and the ounce markers surrounding 1,500 g have frequencies similar to those of other ounce markers.

More formally, McCrary (2008) suggests a direct test for possible manipulation of the running variable. We implement his test by collapsing our nationwide data to the gram level—keeping 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.[27]

Fetal deaths are not included in the birth records data, and hence one possible source of sample selection is the possibility that

25. See Online Appendix Figure A1 for a wider set of births.

26. Specifically, we construct one-ounce bins radiating out from our threshold of interest (e.g., 0–28 g from the threshold, 29–56 g from the threshold).

27. For 1,500 g we estimate a coefficient of −2,100 (s.e. = 1,500).

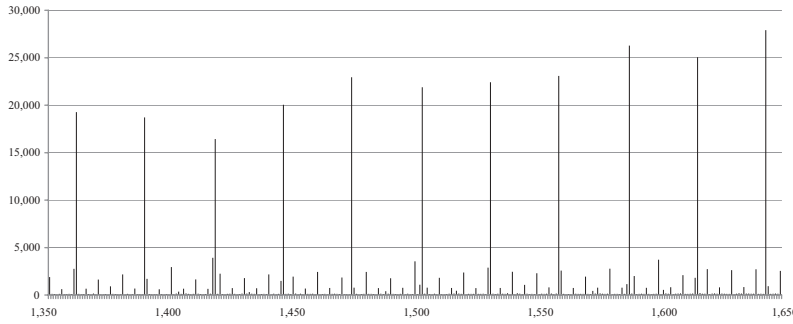

FIGURE I
Frequency of Births by Gram: Population of U.S. Births between 1,350 and 1,650 g
NCHS birth cohort linked birth/infant death files, 1983–1991 and 1995–2003, as described in the text.

very sick infants are discontinuously reported more frequently as fetal deaths across our cutoff of interest (and are thus excluded from our sample). We test for this type of sample selection directly using a McCrary test with data on fetal death reports from the National Center for Health Statistics (NCHS) perinatal mortality data for 1985 to 2002. We would be concerned if we observed a positive jump in fetal deaths for VLBW infants, but in fact the estimated coefficient is negative and not statistically significant.[28] Graphical analysis of the data is consistent with this formal test.

More complicated manipulations of birth weight could in theory be consistent with Figure I. For example, if doctors relabeled unobservably sicker newborns weighing just above 1,500 g as being below 1,500 g (to receive additional treatments, for example) and symmetrically "switched" the same number of other newborns weighing just below 1,500 g to be labeled as being above 1,500 g, this could be consistent with the smooth distribution in Figure I. This seems unlikely, particularly given that we will later show that other covariates (such as gestational age) are smooth across our 1,500-g cutoff—implying that doctors would need to not only "symmetrically switch" newborns but symmetrically switch

28. As above, we implement this test by collapsing the NCHS perinatal data to the gram level—keeping count of the number of fetal deaths classified at each gram—and then regressing that count as the outcome variable in the same framework as our regression discontinuity estimates. We estimate a coefficient of −106.59 (s.e. = 78.32).

newborns who are identical on all of the covariates we observe. We hold that the assumption that such switching does not occur is plausible.[29]

V.B. Health Outcomes

Figure II 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 1,500 g appears to be a level shift, with the slope slightly less steep below the threshold.[30] The mean mortality rate in the ounce bin just above the threshold is 6.15%, which is 0.46 percentage points higher than mean mortality just below the threshold of 5.69%. We see a similar 0.48–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 I reports the main results that account for trends and other covariates. 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 three ounces above the threshold (the "untreated" group in this regression discontinuity design). The OLS estimate in the second column mimics the local linear regression but now places equal weight on the observations up to three ounces on

29. Note also that to the extent that hospitals or physicians may have an 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.

30. Note that in this graph there is also a smaller, visible "jump" in mortality around 1,600 g, an issue we address in several ways. First, if we construct graphs analogous to Figure II that focus on 1,600 g as a potential discontinuity, there is no visible jump at 1,600 g. Exploration of this issue reveals that the slightly different groupings that occur when one-ounce bins are radiated out from 1,500 g relative to when one-ounce bins are radiated out from 1,600 g explain this difference, implying that small-sample variation is producing this visible "jump" at 1,600 g in Figure II. Reassuringly, the "jump" at 1,500 g is also visible in the graph which radiates one-ounce bins from 1,600 g, suggesting that small-sample variation is not driving the visible discontinuity at 1,500 g. Finally, when we estimate a discontinuity in a formal regression framework at 1,600 g, we find no evidence of either a first-stage or a reduced-form effect at 1,600 g.

TABLE I
INFANT MORTALITY BY VERY-LOW-BIRTH-WEIGHT STATUS, NATIONAL DATA, 1983–2002 (AVAILABLE YEARS)

	One-year mortality				28-day mortality			
	Local linear model	OLS	OLS	OLS	Local linear model	OLS	OLS	OLS
Birth weight < 1,500 g	−0.0121	−0.0095	−0.0067	−0.0072	−0.0107	−0.0088	−0.0074	−0.0073
	(0.0023)**	(0.0022)**	(0.0022)**	(0.0022)**	(0.0019)**	(0.0018)**	(0.0018)**	(0.0018)**
		[0.0048]*	[0.0040]	[0.0040]		[0.0038]*	[0.0031]*	[0.0031]*
Birth weight < 1,500 g × grams from cutoff (100s)		−0.0136	−0.0119	−0.0111		−0.0114	−0.0102	−0.0097
		(0.0032)**	(0.0032)**	(0.0032)**		(0.0027)**	(0.0027)**	(0.0027)**
		[0.0062]*	[0.0024]**	[0.0018]**		[0.0055]*	[0.0027]**	[0.0022]**
Birth weight ≥ 1,500 g × grams from cutoff (100s)		−0.0224	−0.0196	−0.0184		−0.0199	−0.0179	0.0172
		(0.0029)**	(0.0029)**	(0.0029)**		(0.0024)**	(0.0024)**	(0.0024)**
		[0.0081]**	[0.0074]**	[0.0074]*		[0.0060]**	[0.0056]**	[0.0055]**
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			

TABLE I
(CONTINUED)

	7-day mortality				24-hour mortality			
	Local linear model	OLS	OLS	OLS	Local linear model	OLS	OLS	OLS
Birth weight < 1,500 g	−0.0068	−0.0060	−0.0049	−0.0047	−0.0068	−0.0043	0.0036	−0.0035
	(0.0017)**	(0.0016)**	(0.0016)**	(0.0016)**	(0.0017)**	(0.0013)**	(0.0013)**	(0.0013)**
		[0.0032]	[0.0027]	[0.0027]		[0.0023]	[0.0020]	[0.0020]
Birth weight < 1,500 g × grams from cutoff (100s)		−0.0078	−0.0068	−0.0066		−0.0042	−0.0036	−0.0036
		(0.0024)**	(0.0024)**	(0.0024)**		(0.0019)*	(0.0019)	(0.0019)
		[0.0047]	[0.0026]**	[0.0023]**		[0.0031]	[0.0018]*	[0.0015]*
Birth weight ≥ 1,500 g × grams from cutoff (100s)		−0.0137	−0.0120	−0.0116		−0.0098	−0.0088	−0.0086
		(0.0022)**	(0.0022)**	(0.0022)**		(0.0017)**	(0.0017)**	(0.0017)**
		[0.0049]**	[0.0046]*	[0.0046]*		[0.0036]**	[0.0034]*	[0.0034]*
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	202,071							

Notes. Local linear regressions use a bandwidth of 3 ounces (85 g). OLS models are estimated on a sample within 3 ounces above and below the VLBW threshold. "Main controls" are listed in Online Appendix Table A5, in addition to indicators for five-year intervals of mother's age, five-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. Local linear models report asymptotic standard errors. OLS models report heteroscedastic-robust standard errors in parentheses and standard errors clustered at the gram level in brackets.

* Significant at 5%; ** significant at 1%.

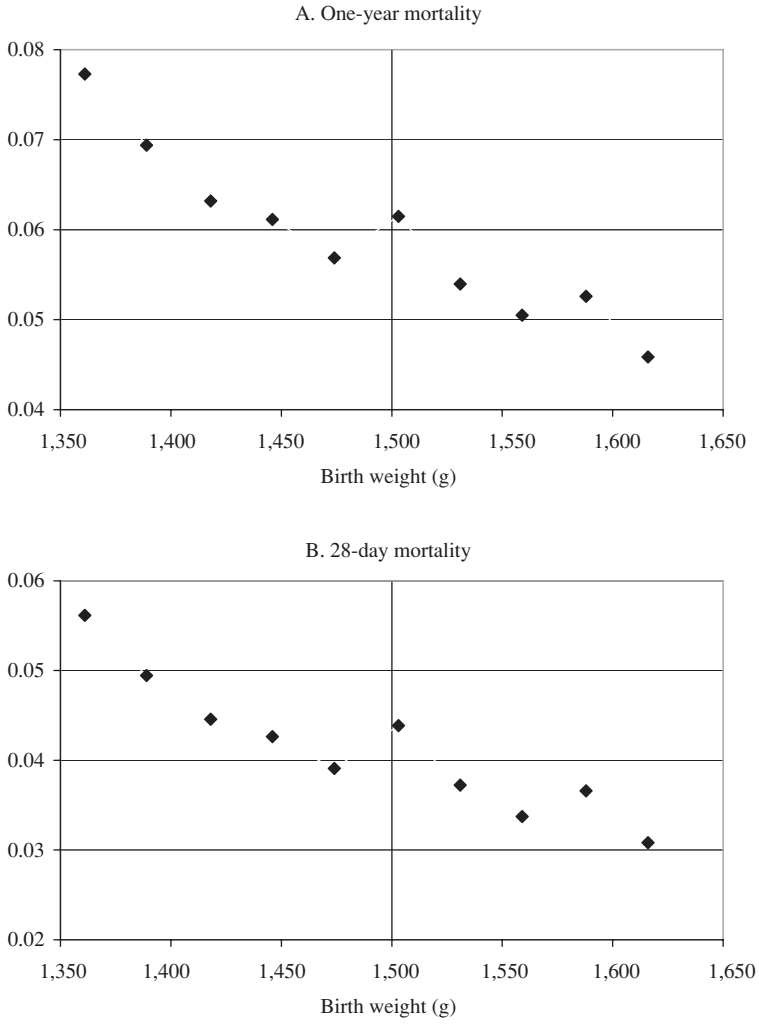

FIGURE II
One-Year and 28-Day Mortality around 1,500 g

NCHS birth cohort–linked birth/infant death files, 1983–1991 and 1995–2003, as described in the text. Points represent gram-equivalents of ounce intervals, with births grouped into one-ounce bins radiating from 1,500 g; the estimates are plotted at the median birth weight in each bin.

either side of the threshold. The point estimate is slightly smaller, but still large: −0.0095. The probit model estimate is similar.[31]

31. A probit model with no controls other than the trend terms predicts a difference of −0.0095 evaluated at the cutoff. A probit model with full controls

The trend terms reflect the overall downward slope in mortality. The point estimates suggest a steeper slope after the threshold. This trend difference could result from greater treatment levels that extend below the cutoff at a diminishing rate. Our estimate of the discontinuity in models that account for trends will not take treatment of inframarginal VLBW infants into account.

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, levels of associated survival rates, and trends in 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. When we include the full set of covariates, the results are largely unchanged.[32] 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 I 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 3.83%. 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 coefficients implies that any effects of being categorized as VLBW 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). Within the first month of life, timing of the mortality gains varies, but the percentage reduction in mortality for VLBW infants relative to the rate above the threshold is consistent with that at 28 days. The 7-day and 24-hour mortality rates are 16% and 19% compared to the mean mortality rate for infants above the threshold. Finally, 1-hour mortality rates (not shown) are also lower for those born just below the threshold.[33]

predicts an average difference across the actual values of the covariates of −0.0069 evaluated at the cutoff.

32. The estimated coefficients on many of these covariates are reported in Online Appendix Table A5.

33. In a probit model with no controls other than the trend terms, the main marginal effect of interest, evaluated at the cutoff, is −0.0018 (s.e. = 0.0007) compared to a mean 1-hour mortality rate of 0.0055 just above the threshold. In a model with full controls, the average marginal effect evaluated at the cutoff is −0.0016.

The following two sections consider the extent to which newborns classified as VLBW receive discontinuously more medical treatments than newborns just above 1,500 g. Although 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 from five states (Arizona, California, Maryland, New Jersey, and New York), which appear to have broadly representative mortality outcomes.[34] When we replicate the results in Table I limiting our nationwide data to these five states (a sample of nearly 50,000 births), we estimate that mortality falls by 1.1 percentage points (s.e. = 0.42) compared to a mean of 5.4% (as reported in Online Appendix Table A7).

V.C. Differences in Summary Measures of Treatment

Figure IIIA 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.[35]

Table II reports the regression results.[36] The first column reports estimates from the local linear regression, which suggests that hospital charges are $9,450 higher just before the threshold—relatively large compared to the mean charges of $82,000 above the threshold. The remaining columns report the OLS results. Without controls, the estimate decreases somewhat to $9,022; with full controls the estimated increase in charges for infants categorized as VLBW is largely unchanged ($9,065, s.e. = $2,297).

34. 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 five-state sample falls toward the middle of the distribution. Further, Online Appendix Table A7 also considers mortality outcomes in these five states in the (smaller) overlap of the years between the HCUP data and the nationwide data. As expected, the results are more imprecisely estimated with the smaller sample, and the point estimates are lower as well.

35. 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.

36. Results are similar when we estimate alternative models, such as count models for length of stay. Note that there are fewer controls in the five-state sample than there are in the nationwide sample, as the discharge data do not include the birth certificate data. Results (not shown) are qualitatively similar in a separate analysis of California, which allows for a wider set of controls from the linked birth certificate data.

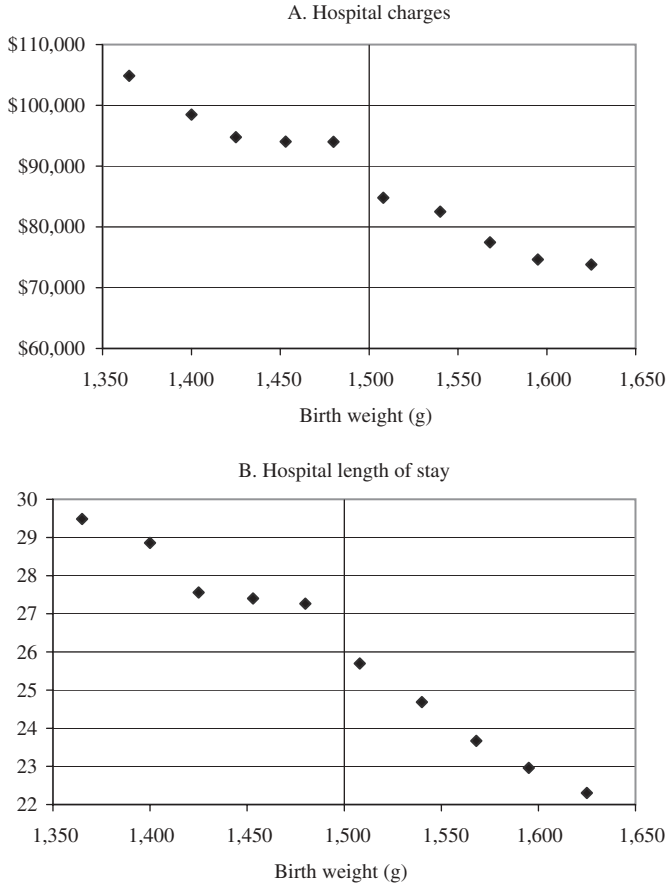

FIGURE III
Summary Treatment Measures around 1,500 g

Data are all births in the five-state sample (AZ, CA, MD, NY, and NJ), as described in the text. Charges are in 2006 dollars. Points represent gram-equivalents of ounce intervals, with births grouped into one-ounce bins radiating from 1,500 g; the estimates are plotted at the median birth weight in each bin.

These estimates imply a difference of approximately 11% compared to the charges accrued by infants above the threshold.

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

TABLE II
SUMMARY TREATMENT MEASURES BY VERY-LOW-BIRTH-WEIGHT STATUS, FIVE-STATE SAMPLE, 1991–2006

	Hospital charges				Length of stay			
	Local linear model	OLS	OLS	OLS	Local linear model	OLS	OLS	OLS
Birth weight < 1,500 g	9,450	9,022	8,205	9,065	1.97	1.7768	1.7600	1.4635
	(2,710)**	(2,448)**	(2,416)**	(2,297)**	(0.451)**	(0.4165)**	(0.4166)**	(0.4107)**
		[3,538]*	[3,174]*	[5,094]		[1.0024]	[0.9775]	[0.7928]
Birth weight < 1,500 g × grams from cutoff (100s)		−1,728	−3,176	617.4876		−0.1012	−0.1356	−0.5766
		(3,700)	(3,647)	(3,463)		(0.6482)	(0.6467)	(0.6366)
		[8,930]	[7,937]	[8,447]		[1.9397]	[1.8419]	[1.4858]
Birth weight ≥ 1,500 g × grams from cutoff (100s)		−7,331	−8,684	−7,951		−2.3130	−2.3779	−2.5993
		(3,018)*	(2,978)**	(2,823)**		(0.5245)**	(0.5250)**	(0.5174)**
		[5,022]	[4,337]*	[7,562]		[1.4366]	[1.4117]	[1.1464]*
Year controls		No	Yes	Yes		No	Yes	Yes
Main controls		No	No	Yes		No	No	Yes
Mean of dependent variable above cutoff	81,566				24.68			
Observations	28,928				30,935			

Notes. Local linear regressions use a bandwidth of 3 ounces (85 g). OLS models are estimated on a sample within three ounces above and below the VLBW threshold. Five states include AZ, CA, MD, NY, and NJ (various years). Charges are in 2006 dollars. "Main controls" are listed in Online Appendix Table A5, as well as indicators for each year. Some observations have missing charges, as described in the text. Local linear models report asymptotic standard errors. OLS models report heteroscedastic-robust standard errors in parentheses and standard errors clustered at the gram level in brackets.

* Significant at 5%; ** significant at 1%.

large charge amounts, as shown in Online Appendix Figure A2 and Online Appendix Table A6.[37]

As noted in Section II.B, 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 quantity measure that is consistently measured across hospitals: length of stay in the hospital.[38] Figure IIIB 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. Corresponding regression results shown in Table II show that newborns weighing just under 1,500 g have stay lengths that are between 1.5 and 2 days longer, depending on the model, representing a difference of 6%–8% compared to the mean length of stay of 25 days above 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 as associated with a greater number of services provided. We further investigate the differences in such service provision measures below.

The first-stage variables in the five-state sample could be censored from above if newborns were transferred to another hospital, because we do not observe charges and procedures across hospital transfers in the HCUP data. This censoring is only problematic insofar as newborns on either side of the cutoff are more likely to be transferred to another hospital. In the discharge data, we do observe hospital transfers, and we do not find a statistically significant difference in transfers across the threshold. The first-stage results are also similar when we use the longitudinal data available from California to consider treatment provided at both the hospital of birth and any care provided in a subsequent hospital following a transfer (Online Appendix Table A6).

It can also be argued that if treatment is effective at reducing mortality, newborns just below 1,500 g will receive more medical treatment than newborns just above 1,500 g because they are

37. 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 1,500 g does not reveal a discontinuous change across this threshold.

38. We define our length of stay variable so 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.

more likely to be alive. Such treatment is unlikely to drive the first-stage results, however, as it is provided to only 1% of newborns below the cutoff, who appear to have longer lives due to their VLBW status (as in Figure II). Nevertheless, any additional care provided to these newborns is part of the total cost of treatment. Our two-sample instrumental variable estimate of the returns to care discussed in Section VIII.B takes into account these additional costs. To the extent that some of this additional care does not contribute to an improvement in mortality, then our estimate will attribute the reduction in mortality to both care that is effective and care that is ineffective. This will lead to estimated returns that are smaller than they otherwise would be if the ineffective care were excluded.

Taken together, these results show 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 $9,000.

V.D. Mechanisms: Differences in Types of Care

The hospital discharge data include procedure codes that can be used to investigate the types of care that differ for infants on either side of the VLBW threshold. We explore the data for such differences, with a special focus on common perinatal procedures.[39] As in the mortality analysis in the smaller five-state sample, however, such differences are difficult to find. Often, for the same procedures, statistically significant regression results do not correspond to convincing graphical results, and convincing graphical results do not correspond to convincing regression results. Table III and Figure IV present regression and graphical results for four common types of treatment.

One of the most common procedures is some form of ventilation.[40] Although Table III provides some evidence of a

39. Specifically, we searched for differences in procedures used to define NICU quality levels in California (Phibbs et al. 2007), as well as five 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.

40. We observe several measures of assisted ventilation, but found little support for any discontinuous change in any of the measures. 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. As noted above, the nationwide data include some

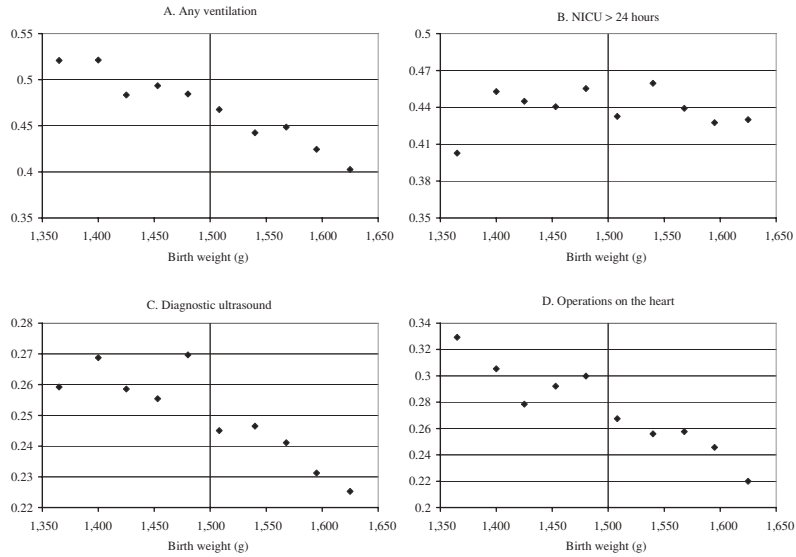

FIGURE IV

Specific Treatment Measures around 1,500 g

Data are all births in the five-state sample (AZ, CA, MD, NY, and NJ), as described in the text. Points represent gram-equivalents of ounce intervals, with births grouped into one-ounce bins radiating from 1,500 g; the estimates are plotted at the median birth weight in each bin.

discontinuous increase in ventilation for VLBW infants, Figure IVA does not offer compelling evidence of a meaningful difference.

Another common measure of resource utilization that we observe in our summary treatment measures is admission to a neonatal intensive care unit (NICU). Because care provided in such units is costly, it seems plausible that the threshold could be used to gain entry into such a unit. However, our data reveal little difference on this margin. 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. Although Table III suggests a modest increase in this NICU use measure (approximately 3 percentage points as compared to a mean just above the threshold of

ventilation measures, but we also find little evidence of an increase in ventilation among VLBW newborns in those data.

TABLE III
SPECIFIC TREATMENT MEASURES BY VERY-LOW-BIRTH-WEIGHT STATUS: FIVE-STATE SAMPLE, 1991–2006

	Ventilation (various methods)			California: >24 hours in NICU		
	Local linear model	OLS	OLS	Local linear model	OLS	OLS
Birth weight < 1,500 g	0.0357	0.0380	0.0274	0.0372	0.0282	0.0265
	(0.0125)**	(0.0115)**	(0.0112)*	(0.0170)*	(0.0157)	(0.0156)
		[0.0263]	[0.0191]		[0.0214]	[0.0204]
Controls		No	Yes		No	Yes
Mean of dependent variable above cutoff	0.511			0.444		
Observations	30,935			16,528		
	Diagnostic ultrasound of infant			Operations on the heart		
	Local linear model	OLS	OLS	Local linear model	OLS	OLS
Birth weight < 1,500 g	0.0196	0.0166	0.0297	0.0147	0.0155	0.0236
	(0.0109)	(0.0101)	(0.0095)**	(0.0112)	(0.0104)	(0.0100)*
		[0.0128]	[0.0125]*		[0.0338]	[0.0146]
Controls		No	Yes		No	Yes
Mean of dependent variable above cutoff	0.244			0.260		
Observations	30,935			30,935		

Notes. Local linear regressions use a bandwidth of three ounces (85 g). OLS models are estimated on a sample within three ounces above and below the VLBW threshold, and include linear trends in grams (coefficients not reported). Five states include AZ, CA, MD, NY, and NJ (various years). "Main controls" are listed in Online Appendix Table A5, 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. Local linear models report asymptotic standard errors. OLS models report heteroscedastic-robust standard errors in parentheses and standard errors clustered at the gram level in brackets.

* Significant at 5%; ** significant at 1%.

44 percentage points), Figure IVB shows little evidence of a discontinuous change. Second, we examine the Maryland HCUP data, which record the number of days in a NICU, but again we find little evidence of a difference at the threshold.[41] Our results are consistent with a study of NICU referrals (Zupancic and Richardson 1998), in which VLBW was not listed among the common reasons for triage to a NICU.

We find some weak evidence of differences 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 1,500 g as a threshold below which diagnostic ultrasounds are suggested. Figure IVC suggests a jump in ultrasounds of roughly two percentage points compared to a mean of approximately 25%. Table III suggests estimates of similar size, although only the OLS estimates with controls are statistically significant at conventional levels.

The pattern of the "operations on the heart" indicator in Figure IVD shows an upward pretrend in the procedures prior to the threshold and what appears to be a discontinuous drop after the threshold.[42] Table III 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, although again only the OLS estimates with controls are statistically significant at conventional levels.

In summary, we examine several possible treatment mechanisms at the discontinuity. We find some weak evidence of differences for operations on the heart and diagnostic ultrasounds, for which we estimate an approximate 10% increase in usage just prior to the VLBW threshold.[43] These differences are often not

41. The New Jersey HCUP data include a field for NICU charges, but this variable proves unreliable: the fraction of newborns with nonmissing 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 relatively small sample of births in the years for which we observe this variable.

42. LBW is associated with failure of the ductus arteriosis to close, in which case surgery may be necessary. Investigating the surgical code for this particular procedure on its own as used in Phibbs et al. (2007) suggested a low mean (4.4 of 1,000 births) and no visible jump.

43. Although these differences are at best suggestive, it is worth noting that our best estimate based on limited pricing data is that these differences would not account for the majority of the measured difference in total charges. On the basis of 2007 Medi-Cal rates, we estimate that the charge for a diagnostic ultrasound is relatively inexpensive (approximately $450) and various heart operations range

statistically significant, and would be even less so if the standard errors were corrected with a Bonferroni correction to account for search across procedures. We find little evidence of differences in NICU usage or other common procedures such as ventilation. In the end, differences in our summary measures are consistent with medical care driving the mortality results, but we likely lack the statistical power to detect differences in particular procedures in our five-state sample (as evidenced by relatively noisy procedure rates across birth weight bins).

VI. Robustness and Specification Checks

In this section, we test for evidence of differences in covariates across our VLBW threshold (Section VI.A), discuss the sensitivity of our results to alternative bandwidths (Section VI.B), examine our mortality results by cause of death (Section VI.C), and discuss evidence of discontinuities at alternative birth weight and gestational age thresholds (Sections VI.D and VI.E).

VI.A. Testing for Evidence of Differences in Covariates across 1,500 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 1,500 g to being just below 1,500 g. Moreover, as discussed in Section V.A, most forms of strategic recategorization of newborns based on birth weight around 1,500 g should be detectable in our histograms of birth frequencies by gram birth weight. As such, we expect that the newborns will be similar above and below the threshold in both observable and unobservable characteristics. That said, it is still of interest to directly compare births on either side of our threshold based on observable characteristics.

Online Appendix Table A2 compares means of observable characteristics above and below the threshold, controlling for

from \$200 to \$2,200. Without more systematic data on prices, it is difficult to pin down an accurate estimate of what share of charges these two procedures could account for, but they do not appear to be able to explain most of our measured difference in charges. Another approach controlled for common procedures in our charges regression. With their inclusion in the model, the estimated difference in hospital charges falls from our main estimate of \$9,000 to \$5,100. That is, the procedures appear to explain some, but not all of the effect. Length of stay is our other summary measure of treatment, although we find that charges are higher for VLBW newborns even when controlling for length of stay (by \$2,184 (s.e. = 1,587)).

linear trends in grams from the threshold as in the main analysis. The table also includes a summary measure—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 V compares covariates of interest in the 5 ounces around the VLBW threshold.[44] Here, the comparisons appear even more stable across the threshold. In particular, gestational age—which is particularly related to birth weight and shows a statistically significant difference in Online Appendix Table A2—is generally smooth through the threshold. Similarly, Figure VJ, which is on the same scale as actual mortality in Figure II, suggests little difference in predicted mortality across the threshold. It thus appears that newborns are nearly identical based on observable variables regardless of whether they weigh just below or just above the VLBW threshold.[45]

VI.B. Bandwidth Sensitivity

The local-linear regression results are qualitatively similar for a wide range of bandwidths (see Online Appendix Table A3). 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 ($h = 30$ g), the difference in one-year mortality is -2.7 percentage points; when $h = 150$ g, the estimate decreases to -0.8 percentage points, which is similar to our main estimate at a bandwidth of 85 g. In fact, we find qualitatively similar results for bandwidths as large as 700 g. In terms of the treatment measures in the five-state sample, the discontinuity in hospital charges is largest in magnitude for our

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

45. We also investigate the possibility that newborns in our data reported as exactly 3 pounds 5 ounces (1,503 g) were treated as VLBW newborns and only appear above the threshold in our data due to rounding when the birth weight was reported to Vital Statistics. Although we prefer not to exclude the information here (the two-sample IV estimates should correct for the misclassification), when we exclude newborns at 1,503 g, we find a larger discontinuity in one-year mortality (-0.011, s.e. $= 0.0025$) and continue to find a meaningful discontinuity in charges ($5,600, s.e. $= 2,400$).

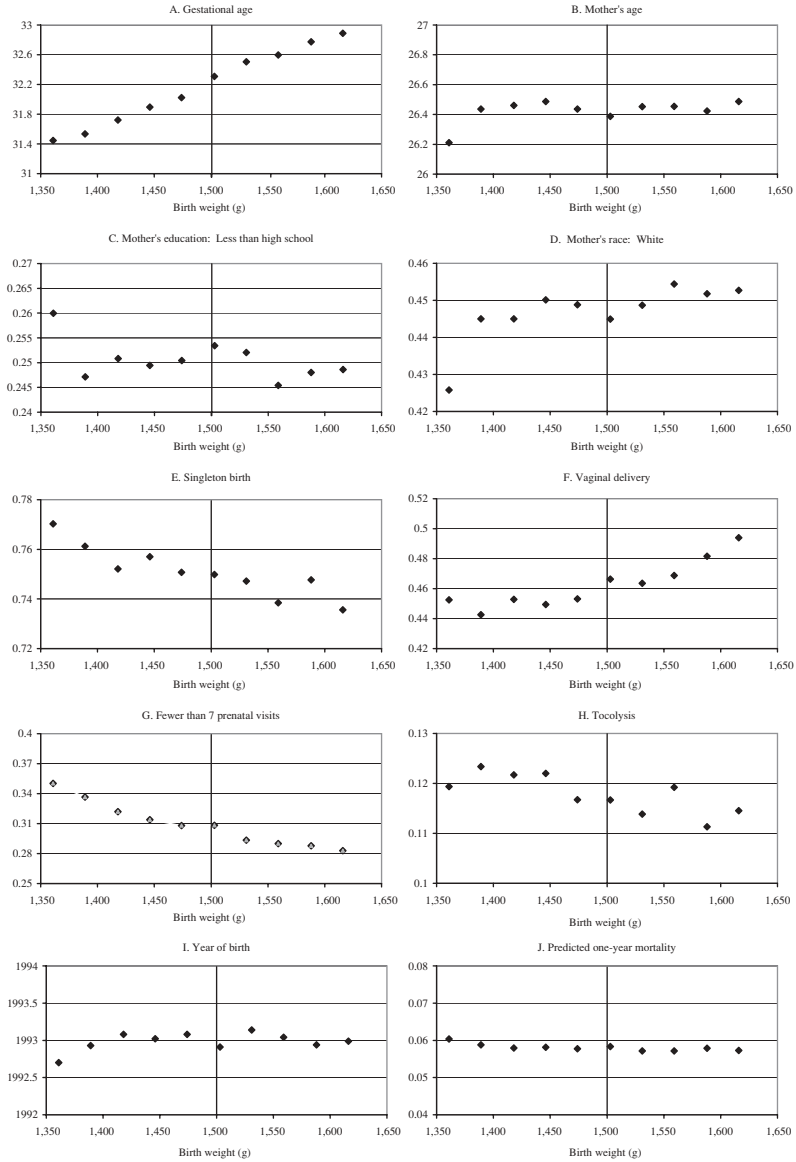

FIGURE V
Covariates around 1,500 g

NCHS birth cohort–linked birth/infant death files, 1983–1991 and 1995–2003, as described in the text. Points represent gram-equivalents of ounce intervals, with births grouped into one-ounce bins radiating from 1,500 g; the estimates are plotted at the median birth weight in each bin.

benchmark bandwidth, although qualitatively similar across the range from $h = 30$ g to $h = 150$ g.

VI.C. Causes of Death

If our mortality effect were driven by so-called "external" causes of death (such as accidents), this would be of concern, because it would be difficult to link deaths from those causes to differences in medical inputs. Reassuringly, we find no statistically significant change in external deaths across our cutoffs of interest (see Online Appendix Table A8).

Examination of our mortality results by cause of death may also be of interest from a policy perspective. When we group causes of death into broad, mutually exclusive categories, we find (see Online Appendix Table A8) effects of the largest magnitude for perinatal conditions (such as jaundice and respiratory distress syndrome), as well as for nervous system and sense organ disorders—the latter of which is a statistically significant effect at conventional levels. We also examine a few individual causes of death, and find a modestly statistically significant reduction in deaths due to jaundice for VLBW infants.[46] These results support the notion that differences in care received in the hospital are likely driving our mortality results.

VI.D. Alternative Birth Weight Thresholds

A main limitation of our analysis is that the returns are estimated at a particular point in the birth weight distribution. For two reasons, we also examine other points in the birth weight distribution. First, other discontinuities could provide an opportunity to trace out marginal returns for wider portions of the overall birth weight distribution. Second, at points in the distribution where we do not anticipate treatment differences, economically and statistically significant jumps of magnitudes similar to our VLBW treatment effects could suggest that the discontinuity we observe at 1,500 g may be due to natural variation in treatment and mortality in our data.

As noted in Section II.B, discussions with physicians and readings of the medical literature suggest that other cutoffs may be relevant. To investigate other potential thresholds, we

46. Jaundice is a common neonatal problem that should be detected during the initial hospital stay for newborns in our bandwidth. According to Behrman, Kliegman, and Jenson (2000, p. 513), "Jaundice is observed during the first week of life in approximately 60% of term infants and 80% of preterm infants."

estimate differences in mortality and hospital charges for each 100-g interval between 1,000 and 3,000 g. 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 1,500 g (23%), other than one found at 1,800 g (27%).[47] A 5% reduction in mortality (relative to the mean) is found at 1,000 g and a 16% reduction in mortality is found at 2,500 g, but graphs do not reveal convincing discontinuities in mortality at these or other cutoffs.

When we consider hospital charges, again 1,500 g stands out with a relatively large discontinuity, especially compared to discontinuities at birth weights between 1,100 and 2,500 g. A 12% increase in charges (relative to the mean) is found for newborns classified as ELPW (1,000 g), with similarly large differences for 800- and 900-g, thresholds. However, differences at and below 1,000 g are not robust to alternative specifications (such as the transformation of charges by the natural logarithm), possibly because there are fewer newborns to study at these lower thresholds and the spending levels are thus 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.[48] These results support the validity of our main findings, but they do not allow us to trace out marginal returns across the distribution.

VI.E. Gestational Age Thresholds

As motivated by the discussion in Section II.B, as an alternative to birth weight thresholds, we also 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

47. A weight of 1,800 g is a commonly cited threshold for changes in feeding practices (Cloherty and 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 1,800 g in our hospital charges measure.

48. We also undertook a more formal search method. Namely, searching for a break between 1,400 and 1,600 g, the largest discontinuity is found at 1,500 g, and that discontinuity also maximizes the F-statistic in a simple model with linear trends.

across the 37-week threshold.[49] 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.[50]

We also investigated the *interaction* between birth weight and gestational age through the "small for gestational age" (SGA) classification: newborns below the tenth percentile of birth weight for a given gestational age. Conversations with physicians suggest that doctors use SGA charts such as that established by Fenton (2003), updating the previous work of Babson and Benda (1976). On this chart, 2,500 g is almost exactly the tenth 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 2,500 g 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) we do find evidence consistent with treatment being based on SGA around 2,500 g. For 1,500 g, analogous results are not clearly consistent with treatment based on the Fenton (2003) definition of SGA around 1,500 g.[51]

VII. Variation in Treatment Effects across Hospital Types

Our regression discontinuity design allows us to assess potential heterogeneity in outcomes and treatment across hospitals.[52]

49. 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.

50. Specifically, the coefficient on an indicator variable for "below 37 gestational weeks" is −0.00070 (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 that includes only vaginal births that are not induced or stimulated and find similar results.

51. Specifically, if we run separate specifications for each value of gestational weeks, we estimate a coefficient of −.0025 (s.e. = .0009) in the 37-week specification, and the coefficient declines in magnitude in specifications that move away from 37 weeks in both directions (at 35 weeks: −.0002 (s.e. = .0012), at 39 weeks: −.0007 (s.e. = .0009)). These coefficients are not directly comparable to our main estimates because they allow separate trends by gestational week. In the Fenton (2003) chart, 1,500 g is considered SGA for newborns with between 32 and 33 gestational weeks, whereas we find that discontinuities in mortality around 1,500 g are most pronounced at 29 weeks and decrease on either side of 29 weeks.

52. We also examined how our estimated treatment effects vary over time and across subgroups of newborns (results not shown). The trends over time are not consistent with any clear medical technology story of which we are aware (see Online Appendix Table A7), such as a "surfactant effect." The more recent birth

In contexts without a regression discontinuity, an estimated relationship between hospital quality and newborn health 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 infants 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 1,500-g 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).[53] 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.[54]

Although we can examine our reduced-form estimates by NICU quality level, it is worth noting that we expect to lack sufficient sample size within these NICU quality-level subsamples to give precise estimates of these effects for our one-year mortality outcome. Perhaps unsurprisingly, regression estimates that interact with our regression discontinuity variable as well as our

certificate data referenced above include an indicator for the use of artificial surfactant which we can use to test directly for this type of effect, and we do not see a visible discontinuity in this variable—again potentially due to the small sample of births. In examining our mortality effects by subgroups, 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 LBW 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.

53. We are grateful to Christopher Afendulis and Ciaran Phibbs for sharing these 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 that in turn are primarily based on whether hospitals offer specific types of procedures, such as specific types of ventilation and surgery).

54. Because of the small number of births observed in Level 0 or Level 1 NICUs, we create a combined category for births in Level 0, 1, and 2 NICU hospitals.

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. 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.

We can only offer a cautious interpretation of these results, given that many of our estimates are not statistically significant at conventional levels. That said, Figure VI provides one descriptive analysis—plotting first-stage estimates by hospital against reduced-form estimates by hospital, normalizing each coefficient by the mean outcome for newborns above 1,500 g within our bandwidth for that type of hospital. Hospitals with larger first-stage estimates have larger reduced-form estimates, providing further evidence that treatment differences are driving the outcome differences. In addition, this analysis provides suggestive evidence that the non-NICU and Level 0/1/2 NICU hospitals are the hospitals where our estimated effects are largest.[55]

VIII. Estimating Returns to Medical Spending

In this section, for comparability to the existing literature, we present a time series estimate of the returns to large changes in spending over time for newborns in our bandwidth (Section VIII.A). We then combine our first-stage and reduced-form estimates to derive two-sample estimates of the marginal returns to

55. Similarly, we find larger first-stage results when we consider hospitals in the five-state sample that had no NICU compared to hospitals that have a NICU—a wider set of states that does not allow an investigation by NICU quality level. We also considered hospital size (calculated using the number of births in a hospital-year in our sample). Larger hospitals had higher levels of charges, so we compared log charges across quartiles in our hospital size variable. The bulk of the data are in the larger hospitals, and we find treatment differences in the second, third, and top quartiles (the bottom quartile contained fewer births ($n = 2{,}110$) and was less precisely estimated). 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.

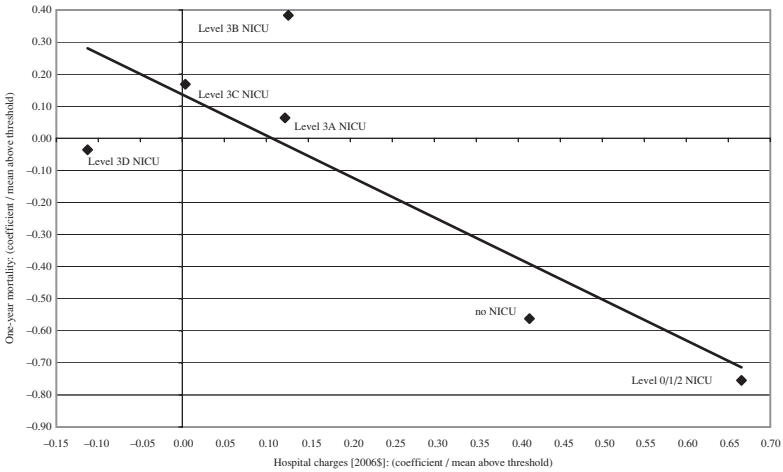

FIGURE VI
First Stage versus Reduced Form, by NICU Quality Level

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.

medical spending for newborns near 1,500 g (Section VIII.B). As noted in Section III.A, all of our spending figures in this section are hospital costs (that is, hospital charges deflated by a cost-to-charge ratio) because costs most closely approximate the true social costs of resource use.

VIII.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 and 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 in the immediate postnatal period (Williams and Chen 1982; Paneth 1995), and thus to value medical improvements by looking at changes over time in within–birth weight expenditures and health outcomes. 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 cost per newborn life 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.

VIII.B. Two-Sample Estimates of Marginal Returns to Medical Spending

As discussed in Section IV.A, 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 only affects mortality through treatments captured by our treatment measure—an assumption that is most plausible for costs, our best available summary treatment measure.

Because we examine health outcomes and summary treatments in different data sources, additional assumptions are required to combine our estimates. To be conservative, we can focus on mortality and cost estimates based solely on states in the five-state sample. We obtain the one-year mortality estimate on the nationwide data, restricted to newborns in the five-state sample in available years and standardize covariates across the two samples.[56] 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.[57] Coefficients are shown in the last column of Online Appendix Table A7, where $4,553 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 five-state sample in the available years (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 $3,795 in costs and a one-year mortality reduction of 0.72 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: $3,795 for each VLBW newborn in the full

56. Specifically, we restrict the national data to the five states in the years 1991 and 1995–2002. Also, for comparability with the five-state sample, we restrict the national sample to contain only in-hospital births.

57. Because we do not have individual-level identifiers, we cannot restrict the national sample to contain the exact same newborns as the five-state sample, but the agreement is very good. The restricted national sample contains 23,698 infants, and the five-state sample contains 21,479.

five-state sample. The denominator is our mortality estimate: a 0.72-percentage point reduction in mortality among VLBW newborns in the full sample. These estimates imply that the cost per newborn life saved is $527,083 ($3,795/0.0072). In the five-state sample over the years that overlap with the nationwide data, we obtain a slightly higher estimate of costs per newborn life saved of $615,270 ($4,553/0.0074). Following Inoue and Solon (2005), 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: $30,000 to $1.05 million.

We can compare these estimates of the cost per newborn life saved to a variety of potential benchmarks. Using data on disabilities and life expectancy, Cutler and Meara (2000) calculate a quality-adjusted value of a newborn life for newborns born in 1990 near 1,500 g to be approximately $2.7 million. If we take the less conservative view that newborns who are saved do not experience decreases in lifespan or quality of life, the relevant benchmark is approximately $3 million to $7 million dollars (Cutler 2004). Comparison with this benchmark suggests that the treatments that we observe are very cost-effective.

IX. Conclusion

Medical inputs can vary discontinuously across plausibly smooth measures of health risk—in our case, birth weight—inviting evaluation using a regression discontinuity design. The treatment threshold and estimated effects are relevant to a "marginally untreated" subpopulation. The relatively frequent use of clinical triage criteria (as discussed in Section I) and availability of micro-level data on health treatments and health outcomes imply that this type of regression discontinuity analysis may be fruitfully applied to a number of other contexts. This approach offers a useful complement to conventional approaches to estimating the returns to medical expenditures—which have generally focused on time-series, cross-sectional, or panel variation in medical treatments and health outcomes—yet has not been widely applied in either the economics literature or the health services literature to date (Zuckerman et al. 2005; Linden, Adams, and Roberts 2006).

In the universe of all births in the United States over twenty years, we estimate that newborns weighing just below 1,500 g have substantially lower mortality rates than newborns that weigh just over 1500 g, 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 1,500 g from above, which is large relative to mean one-year mortality of 5.5% just above 1,500 g. Robustness tests suggest some variation around this point estimate, but we generally find a reduction in mortality of close to 0.7 percentage points for newborns just below the threshold.

It appears that infants categorized as VLBW 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 mortality discontinuity similar to that for 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. Although 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 VLBW 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 1,500 g is on the order of $550,000 with an upper bound of approximately $1.2 million in 2006 dollars. Although the cost measures may not fully capture the additional care provided to VLBW newborns, the magnitude of the cost-effectiveness estimates suggests that returns to medical care are large for this group.

COLUMBIA UNIVERSITY AND NATIONAL BUREAU OF ECONOMIC RESEARCH
MIT AND NATIONAL BUREAU OF ECONOMIC RESEARCH
YALE UNIVERSITY AND NATIONAL BUREAU OF ECONOMIC RESEARCH
HARVARD UNIVERSITY

REFERENCES

Almond, Douglas, and Joseph Doyle, "After Midnight: A Regression Discontinuity Design in Length of Postpartum Hospital Stays," NBER Working Paper No. 13877, 2008.

Andre, Malin, Lars Borgquist, Mats Foldevi, and Sigvard Molstad, "Asking for 'Rules of Thumb': A Way to Discover Tacit Knowledge in Medical Practice," *Family Medicine*, 19 (2002), 617–622.

Angert, Robert, and Henry Adam, "Care of the Very Low-Birthweight Infant," *Pediatrics Review*, 30 (2009), 1–32.
Anspach, Renee, *Deciding Who Lives: Fateful Choices in the Intensive-Care Nursery* (Berkeley, CA: University of California Press, 1993).
Babson, S. Gorham, and Gerda Benda, "Growth Graphs for the Clinical Assessment of Infants of Varying Gestational Age," *Journal of Pediatrics*, 89 (1976), 814–820.
Baicker, Katherine, and Amitabh Chandra, "Medicare Spending, the Physician Workforce, and the Quality of Care Received by Medicare Beneficiaries," *Health Affairs*, W4 (2004), 184–197.
Behrman, Richard, Robert Kliegman, and Hal Jenson, *Nelson Textbook of Pediatrics*, 16th ed. (Philadelphia, PA: W.B Saunders Company, 2000).
Card, David, and David Lee, "Regression Discontinuity Inference with Specification Error," *Journal of Econometrics*, 142 (2008), 655–674.
Cheng, Ming-Yen, Jianqing Fan, and J.S. Marron, "On Automatic Boundary Corrections," *Annals of Statistics*, 25 (1997), 1691–1708.
Cloherty, John, and Ann Stark, *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 ed. (Philadelphia, PA: Lippincott-Raven, 1998).
Cutler, David, *Your Money or Your Life: Strong Medicine for America's Health Care System* (New York: Oxford University Press, 2004).
Cutler, David, and Mark McClellan, "Is Technological Change in Medicine Worth It?" *Health Affairs*, 20 (2001), 11–29.
Cutler, David, Mark McClellan, Joseph Newhouse, and Dahlia Remler, "Are Medical Prices Declining? Evidence for Heart Attack Treatments," *Quarterly Journal of Economics*, 113 (1998), 991–1024.
Cutler, David, and Ellen Meara, "The Technology of Birth: Is It Worth It?" *Frontiers in Health Policy Research*, 3 (2000), 33–68.
Cutler, David, Allison Rosen, and Sandeep Vijan, "The Value of Medical Spending in the United States, 1960–2000," *New England Journal of Medicine*, 355 (2006), 920–927.
Enthoven, Alain, *Health Plan: The Only Practical Solution to the Soaring Cost of Medical Care* (Reading, MA: Addison Wesley, 1980).
Fenton, Tanis, "A New Growth Chart for Preterm Babies: Babson and Benda's Chart Updated with Recent Data and a New Format," *BMC Pediatrics*, 3 (2003), 13.
Fisher, Elliott, John Wennberg, Therese Stukel, and Sandra Sharp, "Hospital Readmission Rates for Cohorts of Medicare Beneficiaries in Boston and New Haven," *New England Journal of Medicine*, 331 (1994), 989–995.
Frank, Richard, and Richard Zeckhauser, "Custom-Made versus Ready-to-Wear Treatments: Behavioral Propensities in Physicians' Choices," *Journal of Health Economics*, 26 (2007), 1101–1127.
Fuchs, Victor, "More Variation in Use of Care, More Flat-of-the-Curve Medicine," *Health Affairs*, 23 (2004), 104–107.
Goodman, David, Elliott Fisher, George Little, Therese Stukel, Chiang Hua Chang, and Kenneth Schoendorf, "The Relation between the Availability of Neonatal Intensive Care and Neonatal Mortality," *New England Journal of Medicine*, 346 (2002), 1538–1544.
Grumbach, Kevin, "Specialists, Technology, and Newborns—Too Much of a Good Thing?" *New England Journal of Medicine*, 346 (2002), 1574–1575.
Horbar, Jeffrey, Gary Badger, Joseph Carpenter, Avroy Fanaroff, Sarah Kilpatrick, Meena LaCorte, Roderic Phibbs, and Roger Soll, "Trends in Mortality and Morbidity for Very Low Birth Weight Infants, 1991–1999," *Pediatrics*, 110 (2002), 143–151.
Imbens, Guido, and Thomas Lemieux, "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142 (2008), 615–635.
Inoue, Atsushi, and Gary Solon, "Two-Sample Instrumental Variables Estimators," NBER Technical Working Paper No. 311, (2005).
Kessler, Daniel, and Mark McClellan, "Do Doctors Practice Defensive Medicine?" *Quarterly Journal of Economics*, 111 (1996), 353–390.

Lee, David, and Thomas Lemieux, "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, forthcoming.
Lichtig, Leo, Robert Knauf, Albert Bartoletti, Lynn-Marie Wozniak, Robert Gregg, John Muldoon, and William Ellis, "Revising Diagnosis-Related Groups for Neonates," *Pediatrics*, 84 (1989), 49–61.
Linden, Ariel, John Adams, and Nancy Roberts, "Evaluating Disease Management Programme Effectiveness: An Introduction to the Regression Discontinuity Design," *Journal of Evaluation in Clinical Practice*, 12 (2006), 124–131.
Luce, Brian, Josephine Mauskopf, Frank Sloan, Jan Ostermann, and L. Clark Paramore, "The Return on Investment in Health Care: From 1980 to 2000," *Value in Health*, 9 (2006), 146–156.
Ludwig, Jens, and Douglas L. Miller, "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics*, 122 (2007), 159–208.
McClellan, Mark, "The Marginal Cost-Effectiveness of Medical Technology: A Panel Instrumental-Variables Approach," *Journal of Econometrics*, 77 (1997), 39–64.
McCrary, Justin, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142 (2008), 698–714.
McDonald, Clement, "Medical Heuristics: The Silent Adjudicators of Clinical Practice," *Annals of Internal Medicine*, 124 (1996), 56–62.
Murphy, Kevin, and Robert Topel, "The Economic Value of Medical Research," in *Measuring the Gains from Medical Research*, Kevin M. Murphy and Robert H. Topel, eds. (Chicago: University of Chicago Press, 2003).
Nordhaus, William, "The Health of Nations: The Contribution of Improved Health to Living Standards," NBER Working Paper No. 8818, 2002.
O'Connor, Gerald, Hebe Quinton, Neal Traven, Lawrence Ramunno, Andrew Dodds, Thomas Marciniak, and John Wennberg, "Geographic Variation in the Treatment of Acute Myocardial Infarction: The Cooperative Cardiovascular Project," *Journal of the American Medical Association*, 281 (1999), 627–633.
Paneth, Nigel, "The Problem of Low Birth Weight," *Future of Children*, 5 (1995), 19–34.
Phibbs, Ciaran, Laurence Baker, Aaron Caughey, Beate Danielson, Susan Schmitt, and Roderic Phibbs, "Level and Volume of Neonatal Intensive Care and Mortality in Very-Low-Birth-Weight Infants," *New England Journal of Medicine*, 356 (2007), 2165–2175.
Pilote, Louise, Robert Califf, Shelly Sapp, Dave Miller, Daniel Mark, Douglas Weaver, Joel Gore, Paul Armstrong, Magnus Ohman, and Eric Topol, "Regional Variation across the United States in the Management of Acute Myocardial Infarction," *New England Journal of Medicine*, 333 (1995), 565–572.
Porter, Jack, "Estimation in the Regression Discontinuity Model," University of Wisconsin-Madison Working Paper, 2003.
Pressman, Eva, Jessica Bienstock, Karin Blakemore, Shari Martin, and Nancy Callan, "Prediction of Birth Weight by Ultrasound in the Third Trimester," *Obstetrics and Gynecology*, 95 (2000), 502–506.
Quinn, Kevin, "New Directions in Medicaid Payment for Hospital Care," *Health Affairs*, 27 (2008), 269–280.
Russell, Rebecca, Nancy Green, Claudia Steiner, Susan Meikle, Jennifer Howse, Karalee Poschman, Todd Dias, Lisa Potetz, Michael Davidoff, Karla Damus, and Joann Petrini, "Cost of Hospitalization for Preterm and Low Birth Weight Infants in the United States," *Pediatrics*, 120 (2007), e1–e9.
Stukel, Therese, Lee Lucas, and David Wennberg, "Long-Term Outcomes of Regional Variations in the Intensity of Invasive versus Medical Management of Medicare Patients with Acute Myocardial Infarction," *Journal of the American Medical Association*, 293 (2005), 1329–1337.
Trochim, William, *Research Design for Program Evaluation: The Regression-Discontinuity Design* (Beverly Hills, CA: Sage Publications, 1984).
Tu, Jack, Chris Pashos, David Naylor, Erluo Chen, Sharon-Lise Normand, Joseph Newhouse, and Barbara McNeil, "Use of Cardiac Procedures and Outcomes in Elderly Patients with Myocardial Infarction in the United States and Canada," *New England Journal of Medicine*, 336 (1997), 1500–1505.

United States Congress, Office of Technology Assessment, *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* (Washington, DC: United States Congress, Office of Technology Assessment, 1981).
United States Institute of Medicine, *Preventing Low Birthweight* (Washington, DC: National Academies Press, 1985).
Williams, Ronald, and Peter Chen, "Identifying the Source of the Recent Decline in Perinatal Mortality Rates in California," *New England Journal of Medicine*, 306 (1982), 207–214.
Zuckerman, Ilene H., Euni Lee, Anthony Wutoh, Zhenyi Xue, and Bruce Stuart, "Application of Regression-Discontinuity Analysis in Pharmaceutical Health Services Research," *Health Services Research*, 41 (2005), 550–563.
Zupancic, John, and Douglas Richardson, "Characterization of the Triage Process in Neonatal Intensive Care," *Pediatrics*, 102 (1998), 1432–1436.