

The Role of Hospital Heterogeneity in Measuring Marginal Returns to Medical Care: A Reply to Barreca, Guldi, Lindo, and Waddell*

Douglas Almond
Joseph J. Doyle, Jr.
Amanda E. Kowalski
Heidi Williams

Abstract:

In Almond, Doyle, Kowalski and Williams (2010), we describe how marginal returns to medical care can be estimated by comparing patients on either side of diagnostic thresholds. Our application examines at-risk newborns near the very low birth weight threshold at 1500 grams. We estimate large discontinuities in medical care and mortality at this threshold, with effects concentrated at “low-quality” hospitals. While our preferred estimates retain newborns near the threshold, when they are excluded the estimated marginal returns decline, although they remain large. In low-quality hospitals, our estimates are similar in magnitude regardless of whether these newborns are included or excluded.

In Almond, Doyle, Kowalski and Williams (2010, ADKW), we describe how diagnostic thresholds can provide plausibly exogenous variation in medical care for patients near the threshold. Regression discontinuity estimates for differences in medical care and health outcomes at the threshold can then be combined to estimate marginal returns to medical care. Our application is the very low birth weight (VLBW) threshold at 1500 grams: newborns weighing slightly less than 1500g receive more medical care and have lower mortality compared to those weighing slightly more. The empirical estimates suggest large returns to medical care for these at-risk newborns.

Figure I of ADKW is a histogram of reported births near 1500g, showing pronounced mass points at whole ounces and smaller mass points at 100g intervals. Motivated by this histogram, Barreca, Guldi, Lindo, and Waddell (BGLW) examine a series of alternative specifications. Initially, they exclude newborns with reported birth weights at exactly 1500g. The estimated mortality discontinuity on this sample is smaller in magnitude, although it continues to be both statistically and economically significant. Subsequently, they exclude successively larger sets of newborns, up to a maximum of 1497g to 1503g (inclusive)—a sample restriction that removes 25% of the control observations and 0.38% of the treatment observations, with the asymmetry due to a mass point at 1503g (3lbs 5oz). The estimated mortality discontinuity on this sample declines somewhat and is no longer statistically significant.

Although there is no general economic or statistical case for exclusion of observations at or around the threshold in a regression discontinuity (RD) design, given the specific details of our

* We are grateful to a number of colleagues for helpful comments, and to the National Institute on Aging for financial support (Williams) through Grant T32-AG000186 to the NBER.

application we agree that the exclusion of newborns at 1500g is a useful robustness check that we should have included in our original paper. In contrast, we see no clear case for excluding the larger set of newborns from 1497g to 1503g, and we find that doing so changes the sample composition such that we would in fact expect smaller discontinuity estimates. Regardless, the two welfare-relevant results from ADKW are robust to the inclusion or exclusion of newborns at and around 1500g: we continue to find that discontinuities in both medical care and mortality are concentrated in low-quality hospitals, and our two-sample, two-stage least squares estimate continues to suggest large returns to medical care for these newborns. Below, we discuss these issues in more detail.

First, consider BGLW's exclusion of newborns reported to weigh exactly 1500g. The main empirical concern is that less healthy newborns may be disproportionately likely to have their birth weight rounded to 1500g.¹ Supporting this hypothesis is the fact that newborns at exactly 1500g are anomalous based on *ex ante* fixed characteristics such as race and mother's education. Less supportive of this hypothesis is that although we would expect less healthy newborns to receive correspondingly higher levels of medical care, this is not observed empirically: mean hospital charges for newborns at 1500g are \$83,000, which is in line with mean hospital charges in the one-ounce bin above the threshold (\$85,000), and is \$11,000 less than mean hospital charges in the one-ounce bin below the threshold.²

Given this concern, we agree that the exclusion of newborns at 1500g from the analysis is a useful robustness check. When these newborns are excluded, the mortality discontinuity is smaller in magnitude, although we continue to estimate statistically and economically significant discontinuities in both mortality and medical care.³ The estimates including or excluding newborns at 1500g are not statistically distinguishable.⁴ Our preferred estimate includes newborns at 1500g, especially because

¹ BGLW offer a different motivation based on the mortality rate of these newborns. However, mortality is an endogenous outcome. Neonatology manuals and diagnosis codes define the VLBW threshold as strictly less than 1500g, implying newborns at 1500g are "untreated" in our RD design, and may have higher mortality in part because they receive less medical care. BGLW also discuss two alternative hypotheses for the abnormally high mortality rates among newborns at 1500g: that low-quality hospitals may be more likely to report birth weights at 1500g, or that agents may manipulate reported birth weight to 1500g in order to receive additional medical care. On the first hypothesis, we find no evidence that hospitals with high-level NICUs are differentially less likely to report birth weight at 1500g relative to hospitals with low-level NICUs (2.3% vs. 1.8%, $p > 0.3$). On the second hypothesis, there appears to be no incentive for such manipulation, as newborns at 1500g generally receive less medical care (as described in the text).

² In theory, this observed distribution of medical care could be explained by providers choosing to allocate less medical care to newborns at exactly 1500g if mortality is imminent. If this were the case we would expect particularly large short-term mortality rates for these newborns compared to newborns at other birth weights, with a convergence in longer-term mortality. However, we do not observe this pattern; instead, newborns exactly at 1500g relative to newborns at nearby birth weights have similar differences in mortality at time horizons up to one year.

³ The 1-year mortality discontinuities including and excluding newborns at 1500g are -0.0072 (*s.e.* 0.0040) and -0.0034 (*s.e.* 0.0013), respectively. The analogous hospital charges discontinuities are \$9,022 (*s.e.* 3,538) and \$10,003 (*s.e.* 5,455).

⁴ Formally testing equality of the coefficients, neither the mortality estimates ($p = 0.33$) nor the hospital charges estimates ($p = 0.35$) are statistically distinguishable across specifications that include or exclude newborns at 1500g.

the medical care received by these newborns is in line with that of other non-VLBW newborns above 1500g—suggesting that these observations have not been misclassified.

Second, we find that the hospital heterogeneity results presented in our original paper are robust to BGLW's proposed sample restrictions. In Section VII of ADKW, we use quality measures of Neonatal Intensive Care Units (NICUs) available for California to investigate heterogeneity in treatment effects across hospitals. At hospitals with high-level NICUs (3a/3b/3c/3d), the 1500g threshold does not appear to determine medical care, and no mortality discontinuity is observed. At hospitals with low-level NICUs (0/1/2 or no NICU), we find substantial discontinuities in both medical care and mortality. Although in theory these results could have been driven by differential propensities to report newborns at 1500g across hospitals, Table 1 shows this is not the case. For example, at hospitals with low-level NICUs, the mortality discontinuity is 3.7 percentage points; when observations at 1500g and 1497-1503g are excluded, the estimates are 3.3 percentage points and 4.0 percentage points, respectively. The medical care discontinuity is relatively stable, although is smaller in magnitude and less precise as additional newborns are excluded.

Third, we find that the main welfare-relevant estimates in ADKW—the estimated returns to medical care—are also robust to BGLW's proposed sample restrictions. BGLW confine their analysis to reduced form mortality effects. We extend their analysis by combining discontinuities in medical care and mortality to estimate the returns to medical care. Our original point estimates imply that the cost of saving a statistical life is \$527,000 (using the full sample of mortality data) or \$615,000 (using mortality data from the five states for which we observe hospital discharge data). Dropping the observations at 1500g changes these estimates, respectively, to \$1.32 million and \$1.05 million. Dropping the observations from 1497g to 1503g changes these estimates, respectively, to \$1.53 million and \$584,000.⁵ While the estimates excluding newborns at and around 1500g are less cost-effective than our original estimates, these estimates all fall near or within the asymptotic 95% confidence interval of our original five-state estimate, which was \$30,000 to \$1.2 million. In addition, all of the estimates are well below conventional value of life estimates for this population, which are on the order of \$3 million.

Finally, consider BGLW's exclusion of the larger set of newborns from 1497g to 1503g. The pivotal aspect of this sample restriction is that newborns on the mass point at 1503g (3lbs 5oz) are

⁵ While BGLW focus on the fact that the estimated mortality discontinuity declines in magnitude and is no longer statistically significant when newborns from 1497g-1503g are excluded, as suggested by the relative robustness of the two sample two stage least squares (TS2SLS) estimates the estimated hospital costs discontinuity also declines and is no longer statistically significant on this restricted sample. Smaller mortality and hospital-cost discontinuity estimates on this restricted sample are consistent with a change in hospital composition, as described later in the text.

excluded. There appears to be no clear *a priori* case for excluding these newborns. Much of our data is found at ounce intervals, and there is no visible or statistical evidence of discontinuities in *ex ante* fixed characteristics across the threshold (Figure V and Appendix Table A2 of ADKW).

In extending BGLW's analysis, we find that the exclusion of newborns on this mass point at 1503g induces a sample-selection bias by changing the hospital composition such that the smaller observed mortality and medical care discontinuities on this restricted sample are not surprising. The propensity to report at ounce mass points differs across hospitals: hospitals with low-level NICUs are 60% more likely to report birth weight on whole ounces relative to hospitals with high-level NICUs. This means that dropping the mass point at 1503g changes the sample composition of hospitals just above the threshold, differentially excluding newborns at hospitals with low-level NICUs.⁶ That is, among newborns just above the threshold, higher-quality hospitals are overrepresented in this restricted sample. Because higher-quality hospitals have lower mortality rates and higher costs (Table 1), in this restricted sample newborns just above the threshold have higher costs and lower mortality relative to newborns just above the threshold in the full sample. This induced over-representation of newborns with lower mortality just above the threshold implies we should expect a smaller discontinuous increase in mortality at the threshold in the restricted sample relative to the full sample. Similarly, this induced over-representation means we should also expect a smaller discontinuous decrease in hospital costs at the threshold. In summary, the smaller discontinuities in mortality and hospitals costs in the sample excluding newborns from 1497g-1503g are both expected and consistent with our original hospital-heterogeneity findings.

To conclude, BGLW question whether the observed variation in medical care and mortality across the 1500g threshold is informative for estimating the marginal returns to medical care for these newborns. Our reading of the evidence suggests this variation is informative. In the end, regardless of whether the observations at and around 1500g are retained, the evidence continues to suggest large returns to medical care for these at-risk newborns, with effects concentrated in low-quality hospitals.

⁶ The fraction of births reported at 1503g is 5.0% at hospitals with high-level NICUs and 7.6% at hospitals with low-level NICUs ($p < 0.05$). Within our analysis dataset encompassing 3 ounces on either side of 1500g, the fraction reported at whole ounces is 25% at hospitals with high-level NICUs and 40% at hospitals with low-level NICUs ($p < 0.001$).

References

- Almond, Douglas, Joseph J. Doyle Jr., Amanda Kowalski, and Heidi Williams (2010) "Estimating marginal returns to medical care: Evidence from at-risk newborns," *Quarterly Journal of Economics* 125(2): 591-634.
- Barreca, Alan, Melanie Guldi, Jason Lindo, and Glen Waddell (2010) "Saving babies? Revisiting the effect of the very low birth weight classification," available online at: http://pages.uoregon.edu/jlindo/SavingBabies_BGLW10_QJE.pdf

Table 1. Robustness of ADKW (2010) hospital heterogeneity results

	<u>Low-level NICU hospitals</u> (Levels 0/1/2 or no NICU)			<u>High-level NICU hospitals</u> (Levels 3A/3B/3C/3D)		
	<i>N</i>	First stage estimate: Hospital costs	Reduced form estimate: One-year mortality	<i>N</i>	First stage estimate: Hospital costs	Reduced form estimate: One-year mortality
Control mean (full sample)		13,323	0.0562		46,642	0.0433
Full sample	3,732	7,060 (1,882)** [3,164]*	-0.0374 (0.0144)** [0.0166]*	12,796	2,666 (1,947) [1,778]	0.0073 (0.0076) [0.0083]
Excluding 1500g	3,664	7,190 (1,921)** [3,273]*	-0.0325 (0.0145)* [0.0158]*	12,506	3,036 (1,993) [1,874]	0.0100 (0.0077) [0.0080]
Excluding 1499g-1501g	3,655	6,939 (1,922)** [3,281]*	-0.0381 (0.0140)** [0.0141]**	12,463	3,229 (2,009) [1,899]	0.0100 (0.0078) [0.0081]
Excluding 1498g-1502g	3,640	6,901 (1,937)** [3,344]*	-0.0377 (0.0141)** [0.0142]**	12,360	2,812 (2,054) [1,844]	0.0115 (0.0080) [0.0083]
Excluding 1497g-1503g	3,338	2,987 (2,597) [3,364]	-0.0395 (0.0173)* [0.0176]*	11,670	3,482 (2,341) [2,349]	0.0060 (0.0089) [0.0086]

Notes: OLS models estimated on a sample within 3 ounces above and below the VLBW threshold. Sample is California hospitals, for which these detailed NICU quality data are available. Controls include those listed in Online Appendix Table A5 of ADKW as well as indicators for each year. As in our original analysis in Section VII of ADKW, the total number of observations in this analysis is 16,528, although some observations have missing data on hospital charges as described in the text of ADKW. Hospital charges are deflated by cost-to-charge ratios. Heteroscedastic-robust standard errors in parentheses and standard errors clustered at the gram level in brackets. * Statistically significant at 5%; ** statistically significant at 1%.