Schumpeter School
of Business and Economics

SCHUMPETER DISCUSSION PAPERS

# First do no harm. Then do not cheat: DRG upcoding in German neonatology

Hendrik Jürges
Juliane Köberlein

The Schumpeter Discussion Papers are a publication of the Schumpeter School of Business and Economics, University of Wuppertal, Germany
For editorial correspondence please contact
SSBEEditor@wiwi.uni-wuppertal.de

SDP 2014-001
ISSN 1867-5352

Impressum
Bergische Universiät Wuppertal
Gaußstraße 20
42119 Wuppertal
www.uni-wuppertal.de

BERGISCHE
UNIVERSITÄT
WUPPERTAL

# First do no harm. Then do not cheat: DRG upcoding in German neonatology

Hendrik Jürges & Juliane Köberlein [*]

July 16, 2013

**Abstract**

Since 2003 German hospitals are reimbursed according to diagnosis related groups (DRGs). Patient classification in neonatology is based inter alia on birth weight, with substantial discontinuities in reimbursement at eight different thresholds. These discontinuities create strong incentives to upcode preterm infants into classes of lower birth weight. Using data from the German birth statistics 1996 to 2010 and German hospital data from 2006 to 2011, we estimate that since the introduction of DRGs, hospitals have upcoded at least 12,000 preterm infants and gained additional reimbursement in excess of 100 million Euro. The scale of upcoding in German neonatology enables us to study the anatomy of cheating in a profession that otherwise claims to have high ethical standards. We show that upcoding is not only positively linked with the strength of financial incentives but also with expected treatment costs measured by poor newborn health conditional on weight. This suggests that doctors and midwives do not indiscriminately upcode any potential preterm infant as a rational model of crime would predict. Rather, they may find it easier to cheat when this helps aligning the lump-sum reimbursement with the expected actual treatment costs.

**JEL classification:** I11, I18, D20

**Keywords:** Neonatal care, DRG upcoding

[*] We are grateful to participants at the conference on "Empirical Health Economics" at Ifo/mea/University of Munich, and seminar participants at the Mannheim and Kassel University for helpful comments and suggestions. Authors' contact information: juerges@uni-wuppertal.de, koeberlein@wiwi.uni-wuppertal.de; Schumpeter School of Business and Economics, University of Wuppertal; Rainer-Gruenter-Str. 21 (FN); 42119 Wuppertal, Germany; Fax: +49 (0) 202-439-1384.

# 1 Introduction and Background

In the last two decades many industrialized countries have introduced prospective payments systems for the reimbursement of hospital inpatient care based on so-called diagnosis-related groups or DRGs. By paying a flat amount conditional on patient characteristics to health care providers, such payment systems generally aim at a more efficient allocation of resources in health care (Ellis and McGuire, 1993). Payment according to DRGs limits hospitals' incentives to provide unnecessary treatment and reduces average length of stay. Further, DRGs foster internal transparency, allowing hospitals to specialize in areas where they are relatively efficient, i.e. where actual treatment costs are below the flat reimbursement rates. They also foster external transparency, allowing comparisons across hospitals in terms of quality and efficiency, conditional on morbidity (measured by the hospital's "case-mix index").

However, payment by DRGs may also have a number of unintended consequences. For instance, providers may have an incentive for an inappropriately early discharge of patients, thereby shifting costs to other sectors (e.g. rehabilitation or long-term care). Further, medically necessary diagnostics and therapies may be withheld to save costs. Conditional on DRGs, hospitals may have an incentive to select patients with expected costs lower than the DRG payment and to turn down patients with higher expected costs. Finally, payment by DRGs invites the coding of patients into groups with a higher reimbursement, so-called upcoding.

DRG upcoding can take various forms. Patients are usually coded into their DRGs by specialist coders based on medical charts and using special coding software. At the coder level, there are legal, semi-legal and illegal types of upcoding. Legal types include just better coding (i.e. less downcoding), such as adding existing co-morbidities that raise treatment costs. Semi-legal types include changing the primary and secondary diagnosis in case of co-morbidities. Illegal types include adding co-morbidities that are not documented in the medical charts. The semi-legal and illegal types can in principle be detected in audits.

Yet another variant is false documentation, i.e. the manipulation of patient charts so that patients appear to be sicker than they really are. This type of upcoding is particularly interesting. First, it is not done by some revenue-maximizing manager in the hospital administration who usually has no contact with patients but by doctors or nurses. Second, manipulation of patient charts can hardly be detected by audits, especially the type of upcoding studied in this paper, where we estimate the extent and the determinants of of DRG upcoding in German neonatal care. Birth weight is one of the primary classification criteria in the German DRG system, with lower birth weights yielding substantially larger payments particularly for the care of infants with very low and extremely low birth weight. Differences in birth weight of a few grams can induce additional payments of more than 15,000 Euro (20,000 USD). Thus the financial incentives to manipulate documented birth weight by a few grams can be strong. At the same time, false documentation of birth weights is practically non-verifiable ex post because infants generally lose some 5% of their initial weight in the first few days after birth. Thus the only restraint to unfettered upcoding are the personal or professional ethics of the health care workers in charge of measuring newborn weight.

In our paper, we document widespread upcoding in German neonatology in the form of shifting birth weights from just above DRG-relevant thresholds to just below these thresholds. Such upcoding has already been reported by Abler et al. (2011), who compare birth weight distributions in the German federal state of North Rhine-Westphalia before and after the introduction of DRGs in 2003. We extend this analysis in several dimensions. First, we use data from all births in Germany since 1996. We show that, in the first eight years after the inception of DRGs, an estimated 12,000 newborns have been upcoded into lower birth weight categories. Second, we estimate that excess reimbursement due to upcoding in that period was nearly 115m Euro (150m USD). Third, we analyze economic determinants of upcoding, such as the strength of financial incentives, regional market conditions, or hospital ownership. In international comparison, Germany has a very high number of neonatal care units (so-called perinatal centers) and the average number of patients per unit is comparatively small (Gerber et al., 2008). Thus competition among providers may be strong. Considering the high fixed costs of setting up a neonatal care unit, average costs may be high and some hospitals may suffer from substantial underfunding in neonatal care (Hoehn et al., 2008; Müller et al., 2007). This underfunding can be compensated by coding newborns into higher paying categories. A detailed description of the DRGs that apply to German neonatology, their incentives for upcoding, and the structure of the market for neonatal care in Germany is given below.

Finally, we are the first to exploit the scale of upcoding in German neonatology to study fraudulent behavior in a profession that otherwise have high ethical standards and in an environment where the chances of getting caught by auditors are practically zero. We show that upcoding is not only positively linked with the strength of financial incentives but also with expected treatment costs measured by poor newborn health conditional on weight. This suggests that doctors and midwives do not indiscriminately upcode any potential preterm infant as a rational model of crime would predict. If upcoding was purely driven by the opportunity to increase revenue, newborn health status (and hence expected treatment costs) should not matter. However, if recording a wrong birth weight induces cognitive dissonance between the self-image of an ethical health care worker and factual fraudulent behavior, individuals may selectively upcode children with relatively high expected treatment costs (see e.g. Mazar et al. (2008) for a general discussion and experimental evidence). Doctors and midwives may find it easier to cheat when this helps aligning the lump-sum reimbursement with the expected actual treatment costs.

Our paper is structured as follows. In the remainder of this introductory section we describe DRGs in German neonatology and the market for neonatal care in Germany. Section 2 describes our data sources. Section 3 documents the pervasiveness of upcoding and provides an estimate of the excess reimbursement hospitals have received by upcoding preterm newborns. In Section 4 we explore how economic incentives to upcode and upcoding rates are related, and in Section 5 we explore the relationship between newborn health and upcoding. Finally, Section 6 summarizes the paper and draws both methodological and policy conclusions.

### 1.1 DRGs in German neonatology

Before 2003, neonatal care in German hospitals was reimbursed on a per diem basis. Since the introduction of the DRG system, reimbursement is based on the following case characteristics: birth weight (or admission weight), surgical (OR-) procedures, long-term artificial respiration, (severe) complications, and 28-day mortality. Birth weights are classified along eight threshold values: 600g, 750g, 875g, 1,000g, 1,250g, 1,500g, 2,000g, and 2,500g. Reimbursement changes suddenly and dramatically at these thresholds, so that very small differences in birth/admission weight of a few grams can result in reimbursement differences of more than 15,000 Euro. The relationship between birth weight and average reimbursement per birth weight category is shown in Table 1. It illustrates the substantial hikes in reimbursement at the threshold values. For instance, whether a newborn weights 1510g or 1490g makes a 13,500 Euro reimbursement difference to the neonatal care unit. Actual cost of care differences should be positive but definitely smaller.

**Table 1:** Expected reimbursement for neonatal care in 2010 depending on birth weight category.

| Birth weight (g) | < 600 | 600-<br>749 | 750-<br>874 | 875-<br>999 | 1,000-<br>1,249 | 1,250-<br>1,499 | 1,500-<br>1,999 | 2,000-<br>2,499 | ≥2,500 |
|---|---|---|---|---|---|---|---|---|---|
| Avg. Reimbursement (Euro) | 80,083 | 79,260 | 62,540 | 45,985 | 34,075 | 27,205 | 13,848 | 4,080 | 1,108 |

Source: Fee Schedule 2010 (InEK, 2009); Base rates published by the Federal Association of the AOK (AOK-Bundesverband, 2012)

The incentives created by DRGs based on birth weight are illustrated in Figure 1. It shows the true average treatment costs as a function of birth weight (solid line) and the reimbursement received by the hospital, conditional on birth weight across two birth weight thresholds. The reimbursement for the 1,500g to 2,000g group is determined so that it exactly covers the true treatment costs in that group. All newborns with birth weight between 1,500g and $x$ create financial losses to the hospital, and all newborns with birth weight between $x$ and 1,999g create financial gains. If hospitals were able to select newborns on the basis of their birth weight, they would select to treat those with a weight between $x$ and 1,999g and not treat those with a weight between 1,500g and $x$. But of course, this is not possible. Instead, it is possible to re-classify newborns by manipulating the birth weight. While the true treatment costs of each newborn remain the same, reimbursement will increase from less than expected treatment costs to more than expected treatment costs.

In the German DRG system, the reimbursement per case is obtained by multiplying the *relative DRG cost weight* of that case by a *base rate*. The relative cost weights represent the ratio of resource intensity between different DRG groups (InEK, 2007). They are determined annually by the Institute for the Hospital Reimbursement System (Institut für das Entgeltsystem

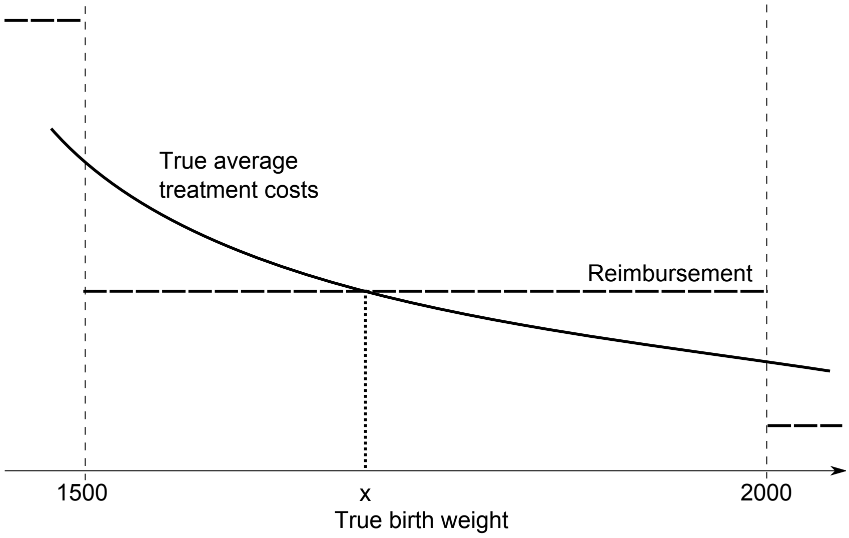

**Figure 1:** Relationship between average treatment costs and reimbursement (stylized graph)

im Krankenhaus, InEK). Calculations are based on actual cost data of a (voluntary) sample of about 250 hospitals (about 13% of German hospitals) covering some 4m individual cases, and using an exact full cost approach for the complete treatment process of a medical condition (InEK, 2011).[1] The relative cost weight of a DRG is computed as the average within-DRG treatment costs (after the elimination of outliers) divided by the average treatment costs across all cases. Thus relative cost weights larger than one represent higher than average costs and relative cost weights lower than one represent lower than average costs.

By multiplying these calculated cost weight by the base rate, one obtains the actual reimbursement per DRG. The base rate is currently determined by negotiation, separately in each federal state, between hospitals and health insurers. Thus base rates slightly depend on regional factors such as regional price or wage differentials (Vogl, 2012). Nation-wide base rates are planned to become operational in 2014.

As noted above, birth weight manipulation is easy and almost impossible to detect ex post in an individual case. It is thus no surprise that such manipulation can be found in other DRG systems as well. For instance, the Japanese partial DRG system does reimburse care

[1] According to §21 Hospital Remuneration Act (KHEntG), all German hospitals are obliged to provide annual hospital-related structural (e.g. ownership, number of beds, labor cost) and case-related performance data (e.g. diagnosis, procedures). Hospitals voluntarily participating as calculation institutions additionally provide patient-level cost data. Their effort is reimbursed by an additional fee consisting of a lump sum and a variable rate related to the number of transferred cases as well as data quality (Geissler et al., 2011).

in neonatal intensive care per diem but the number of days that can be claimed have caps depending on birth weight categories. Thus there is an indirect incentive to manipulate birth weight because it allows hospitals to extend actual treatment. Shigeoka and Fushimi (2013) provide evidence for such manipulation at the relevant thresholds of 1,000g and 1,500g.

Finally, another upcoding margin in German neonatology, which we cannot analyze with our data, is the number of hours of artificial respiration, with 120 hours being the threshold at which remuneration increases substantially. For instance, crossing that threshold in case of a newborn with birth weight 1,500 to 1,999 grams and significant complications, prolonging artificial respiration from 120 to 121 hours can results in fee increase of about 20,000 Euro. Whether this additional hour was medically necessary cannot be verified ex post. In contrast to manipulating recorded birth weights, prolonging artificial respiration might even be harmful as the risk of serious infections increases.

### 1.2 The market for neonatal intensive care

Since 2006, neonatal care hospitals in Germany are assigned to four different levels of treatment capacity. The assignment of a hospital to a certain level is based on the following structural characteristics: professional standards for medical and nursing care, hours of shift work and standby duty, number of normal, respiration and intensive care beds, and cooperation with pediatric hospitals and specialists (e.g. children's cardiology). Further, a minimum number of treated cases per year is required. These assignments are periodically reviewed. In contrast to international conventions (see e.g. Committee on Fetus and Newborn (2012)), levels of increasing capabilities are given lower numbers. Thus Level 4 denotes regular maternity clinics. Mothers at risk of giving birth prematurely are admitted to Level 3, 2 or 1 "perinatal centers", depending on predicted gestational age or birth weight. Table 2 summarizes the characteristics of newborns and the level of perinatal care centers capable of treating those children. In total, there are about 270 Level 1, 2 and 3 centers in Germany.

Level 1 perinatal centers provide the most advanced intensive care. They must have at least six intensive care beds, the neonatal intensive care unit has to be located close to the delivery room, and a specialized newborn emergency physician is required. About 60% of all

**Table 2:** Characteristics of newborns and their assignment to perinatal care center level

| Category | Level 1 | Level 2 | Level 3 |
|---|---|---|---|
| Birth weight in gram | <1,250 | 1,250-1,499 | ≥1,500 |
| Week+days of gestation | <29+0 | 29+0 to 32+0 | 32+1 to 36 |

NOTE: Classification criteria according to the agreement of the Joint Federal Committee, 2005

perinatal care centers in Germany are Level 1, mostly in bigger cities.[2] Typically, mothers at risk of giving birth prematurely choose Level 1 perinatal centers for childbirth, so that infants need not be transferred after birth. Analyses of German birth records 2003 to 2010 (see below for a general description of the data) reveal that of about 72,000 births with very low birth weight, 48% percent took place outside the county of residence of the mother. Neonatal care for low weight infants is concentrated in a few counties: 50% of live births <1,500g were registered in less than 10 percent of all German counties. The five counties with the largest number of very low weight births are Berlin, Munich, Hamburg, Cologne, and Bonn, covering 17.6% of all such births (but only 10.8% of mothers). An ongoing discussion relates to the minimum number of cases per annum a perinatal care center should care for to be assigned to Level 1. The idea is that a certain minimum number of cases is necessary for a hospital to have the relevant experience and provide sufficient quality of care. The current threshold is 14 cases with birth weight <1,250g per year.

Level 2 centers account for roughly 25% of all perinatal centers. They need to have at least four intensive care beds. Until 2010, they also had to care for at least 14 cases per year, but this requirement was recently abolished by the Federal Joint Committee (G-BA), the highest decision-making body in the German health system. Thus in contrast to Level 1 perinatal care centers, no minimum number of cases is necessary for a hospital to obtain Level 2 status.

Hospitals with an insufficient number of intensive care beds but with mechanical ventilation facilities are assigned to Level 3. They represent about 15% of all perinatal hospitals.

[2] Usually, hospital supply is determined in hospital plans that are drawn up at the federal state level. However, as information about the presence of perinatal care centers were not available for the majority of federal states' hospital plans, the proportion of care levels was estimated based on results of a review of hospitals' individual websites, supplemented by information from the German hospital search website `www.deutsches-krankenhaus-verzeichnis.de`.

Level 3 hospitals should cooperate with neighboring children's hospitals but do not have not have a special newborn emergency physician or children's specialists.

As mentioned above, Germany has a comparatively high density of neonatal care facilities (Gerber et al., 2008). There are roughly two Level 1 neonatal care centers per 10,000 births. In contrast, in Nordic countries, France and the Netherlands, the number of centers per 10,000 births is less than one. Between 2003 and 2011, the number of beds in neonatal care has increased by more than 10%, while the total number of beds in German hospitals and the number of beds in general pediatrics has decreased by about 10%. The strong increase in the number of neonatal care beds seems to have been supply rather than demand driven. Between 2000 and 2010, the absolute number of live births below 1,500g has increased by just about 1.5% (from 8,315 to 8,443, also see Figure 2).

## 2 Data

We use several data sources in this study. The main data source are official German birth statistics 1996 to 2010, covering both the pre-DRG- and the DRG-period in German hospitals. These data include about 10 million births, of which some 688,000 or roughly 7% were of low birth weight (<2,500g). To illustrate, Figure 2 shows the number and percentage of live births, by birth weight category and period (before/after introduction of DRGs). Note that data from the state of Bavaria in 1996 to 1999 have birth weights only in brackets of 100 grams. These were generally excluded from our analyses, hence the break in our time series. Figure 2 indicates a general trend towards a larger proportion of live births with low birth weight. While the absolute number of births with extremely low or very low birth weight has been fairly constant, their proportion has continuously increased. In 1996 for instance, 0.39% of all live births were below 1,000g and another 0.62% were between 1,000g and 1,499g. Until 2010, these proportions have increased to 0.51% and 0.73%, respectively. Detailed numbers can be found in Table A.1 in Appendix A. There are several reasons for this general trend, for instance the increase in the number of multiple births after in vitro fertilization, or better medical care for preterm babies allowing more very light babies to be born alive.

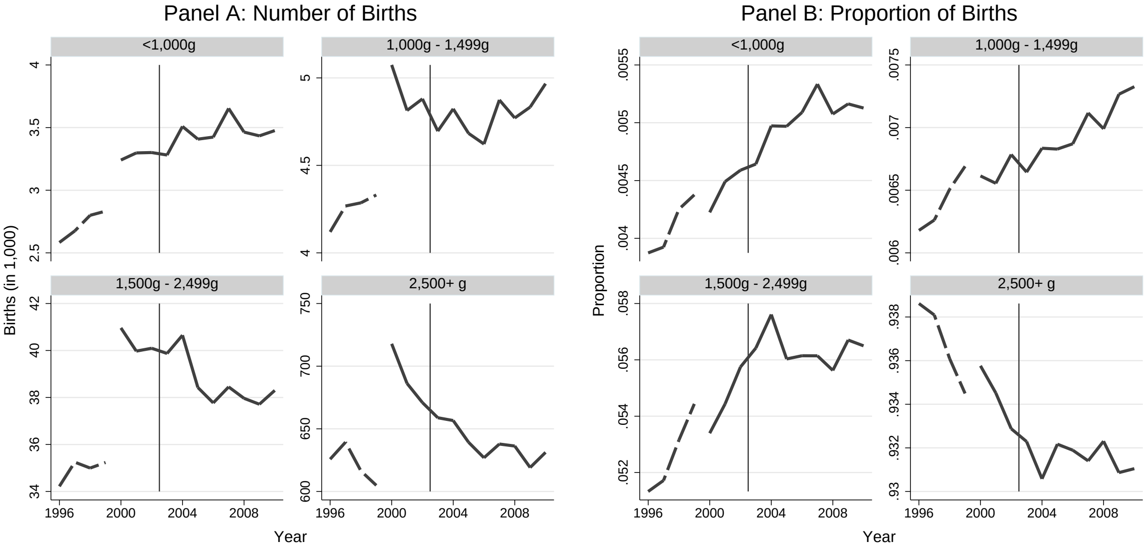

**Figure 2:** Number of births in, by birth weight category. Data before 2000 (dashed lines) do not include births in Bavaria. The vertical line denotes the introduction of DRG reimbursement. Source: Own calculations based on German birth records.

The second data source is German hospital data collected for the purpose of external quality control. Data from 2006 onward are available from the Institute for Applied Quality Improvement and Research in Health Care (AQUA), which is currently responsible for the external quality control. The data contain indicators of infant health taken from the medical records, such as birth weight, gestational age, APGAR scores, or early neonatal mortality. We use these data to study if upcoding is systematically related to infant health conditional on recorded birth weight. The data contain only information on births in hospitals. This will certainly not imply any selectivity problem for the present paper because more than 98% of all births and practically 100% of all preterm births in Germany take place in a hospital.

Third, we use the fee schedules and base rates of the German DRG system to estimate reimbursement differentials per DRG between 2003 and 2010. The fee schedule applicable in the following year is published annually by the Institute for the Hospital Remuneration System (InEK) and contains relative cost weights for each DRG that are computed on the basis of a retrospective full cost approach. Base rates are negotiated between health insurers and hospitals and are regularly published by the Federal Association of the AOK, the largest statutory health insurance in Germany.

Further, we use regional (county-level) data on the number of hospital beds in various departments, including general pediatrics and neonatology, and hospital ownership to describe regional supply-side characteristics in neonatology. These data are derived from the annual hospital directory of the German Federal Statistics Office. Finally, we use county-level data on the number of women aged 15-45, also published by the Federal Statistics Office, to describe the regional demand for neonatal care.

## 3 Upcoding and excess reimbursement over time

### 3.1 Changes in the birth weight distribution over time

We begin by describing trends in the distribution of recorded birth weights from 1996 to 2010, and how changes in the distribution might be related to the introduction of DRGs. Basically, we show that since the introduction of DRGs, recorded birth weights that have made their way into official statistics have been systematically bent below birth weight thresholds that are relevant for reimbursement. If such thresholds are irrelevant for reimbursement, however, there is no such change in the distribution of birth weights after the introduction of DRGs.

The German birth statistics contain the recorded birth weights to the exact gram. For our analyses, we usually recode them into brackets of e.g. 10g or 50g (depending on the analysis) and consider only brackets just below and above the eight DRG-relevant and a few non-relevant birth weight thresholds. The threshold value itself is included in the bracket above the threshold. Let $x$ denote some threshold in grams and $d$ the width of the bracket, then a rough measure of upcoding around $x$ is the ratio of the number of observations in the bracket below the threshold $[x - d, x - 1]$ divided by the number of observations in the brackets above and below the threshold $[x - d, x + d - 1]$. For future reference, we denote this measure as $R_d$:

$$R_d = \frac{\sum 1_{[x-d,x-1]}}{\sum 1_{[x-d,x+d-1]}} \tag{1}$$

Below, we often use $2R_d - 1$ to approximate the proportion of cases that were upcoded from $[x, x + d - 1]$ to $[x - d, x - 1]$, but note that this measure is likely to be an underestimate. It does not account for the fact that all relevant thresholds are in the left hand tail of the birth weight distribution. That means the *true* number of births should be slightly larger right than left of the threshold, whereas $2R_d - 1$ assumes the same true number left and right of the threshold (and thus underestimates the extent of upcoding). However, this does not matter so much for the comparison across years (shown in the present section) as long as the shape of the true distribution does not change over time. But it will of course make a difference when we estimate the absolute number of upcoded cases in each year. We thus calculate an alternative estimate later on.

Figure 3 shows the development of $R_{25}$ at the eight DRG-relevant birth weight thresholds and—for comparison—at four non-relevant thresholds from 1996 to 2010. For instance, the top left figure shows that before the introduction of DRGs, around 40 percent of the babies born with a recorded birth weight between 575 and 624 grams were recorded with less than 600 grams. The proportion lower than 50% must be largely attributed to the fact that birth weights are very often rounded, for instance to multiples of 100. These rounded values are included in the bracket above the threshold. After the introduction of DRGs, $R_{25}$ has increased to about 50 percent. This increase can partly be explained by more precise measurement (through the use of digital scales) or better documentation of birth weights that would formerly have been rounded, which is clearly legitimate.

Increasing trends in $R_{25}$ can be found for almost all DRG-relevant thresholds. The most striking developments can be found at 1,000g, 1,250g and 1,500g, where since the introduction of DRGs, 80% to 90% of recorded birth weights are below the threshold. This translates into a lower bound for the upcoding rates of 60% to 80%. To illustrate the pervasiveness of upcoding: in 2008, we find 742 children recorded with birth weight between 1,480g and 1,499g, but only 65 children with birth weight recorded between 1,500g and 1,519g. A staggering 92% percent of all births were coded below the 1,500g threshold. Possibly, 1,000g and 1,500g figure so prominently because they can be easily remembered as relevant thresholds. Both are also used to categorize birth weight into the familiar "extremely low" and "very low" categories. Another salient threshold, at least since 2006, is at 1,250g, since minimum volume requirements for Level 1 and 2 centers are or were defined with reference to the number of infants born below that weight. In contrast to all other thresholds, we find stable $R_{25}$ at the 875g threshold. Possibly

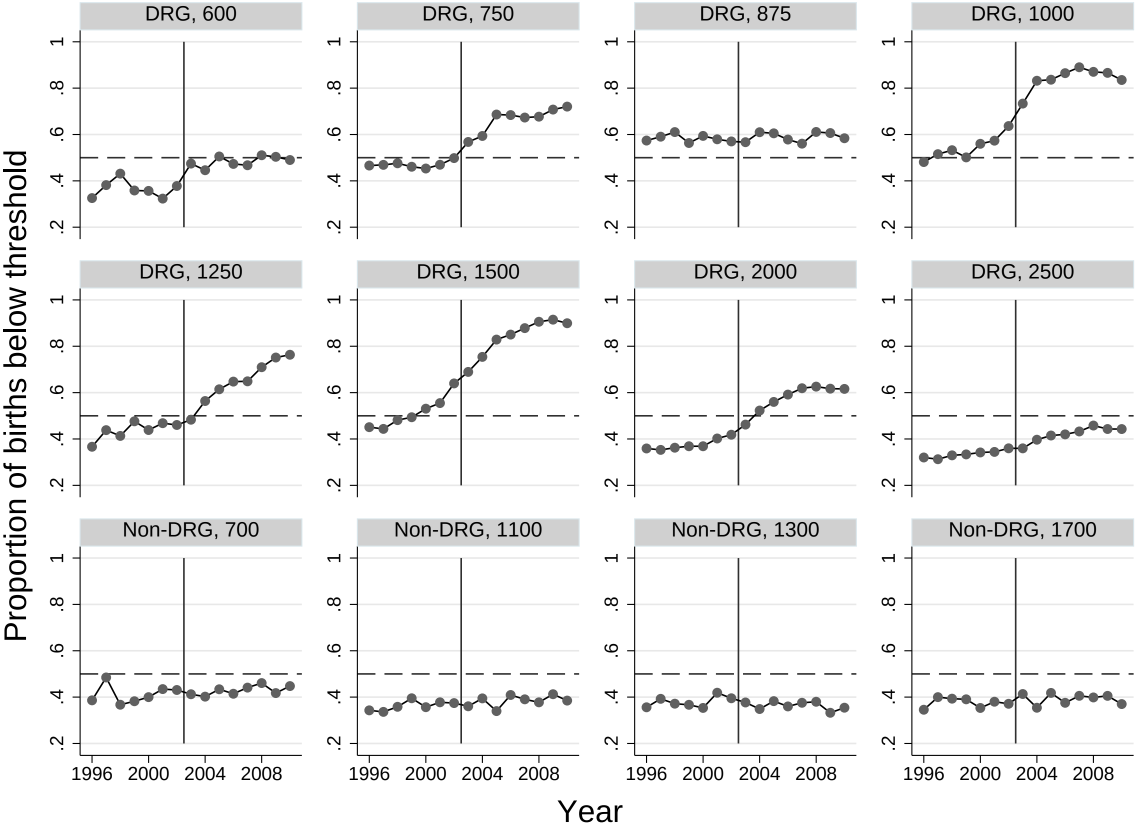

**Figure 3:** Proportion of *live* births with weight recorded below threshold ($d = 25$) at eight DRG-relevant thresholds and four non-relevant thresholds, 1996 to 2010. The vertical line indicates the introduction of DRGs. The horizontal line indicates the "no upcoding" value of 50%. Source: Own calculations based on German birth records.

this is the least memorable of all relevant thresholds. Moreover, $R_{25}$ is consistently at about 60%. This can be explained by the fact that the lower 25g bracket includes a multiple of 50 (850) whereas the upper bracket does not.

Now consider the third row in Figure 3. It shows the percentage of births below 700g, 1,100g, 1,300g, and 1,700g grams, respectively, around these thresholds. These are arbitrarily chosen examples for DRG-irrelevant thresholds, hence no economic incentives apply. Quite consistently, one finds around 40% of recorded birth weights below the threshold and there clearly is no visible change in this proportion over time or when DRGs were introduced.

Comparing the difference in $R_{25}$ for DRG-relevant thresholds with the same difference for non-relevant thresholds mimics a difference-in-difference regression strategy to identify the causal effect of DRGs on recorded birth weights. Our comparison thus provides quite convincing evidence that the introduction of DRGs has affected distribution of documented birth weights only at those thresholds that matter financially. However, it should be noted that the trend towards systematically recording birth weights below the threshold at 1,000g and 1,500g appears to have started a few years *before* the introduction of DRGs and that the DRGs have strengthened this trend. Despite our best efforts, we were not able to identify a single reason for this pre-DRG trend. Before 2003, neonatal intensive care was reimbursed per diem, so that direct financial incentives are not at work. Thus we have checked for non-financial reasons such as the introduction of minimum volume requirements or the introduction of external quality control requirements, but none of these matches the slight increase in birth weight below 1,000g and 1,500g.

Another possibility for comparison is provided by the birth weights recorded for still births. Since still births do not require neonatal care, hospital reimbursement is not affected by the weight of a stillborn child. We have thus reproduced Figure 3 using data on the weight of stillborn infants (details can be found in Figure B.1 in Appendix B). Due to the small number of cases (less than 0.4 percent), there is a lot more annual variation in the data, but the general picture that emerges from our analysis is quite clear: the introduction of DRGs had no effect on the distribution of birth weights among stillborn children. At most threshold values, there is neither a sudden jump nor a secular positive trend in $R_{25}$. A slight upward trend in $R_{25}$

can be seem at some thresholds, both this is both at thresholds relevant and not relevant for neonatal care DRGs. The secular trend might for instance be explained by increasingly precise measurements, i.e. less values recorded exactly at thresholds.

At this point, it seems appropriate to comment on the role of rounding and how it might affect our results. The birth statistics contain the recorded birth weights to the exact gram. Due to lack of precision of scales used in hospitals and due to midwives rounding birth weights, the actual distribution shows substantial heaping at multiples of 10, 50, or 100. Further, we find a secular trend towards more precision in documented birth weights. For instance, between 1996 and 2010, the proportion of birth weights rounded to multiples of 100 fell from 17 to 13 percent, the proportion rounded to multiples of 50 fell from 31 to 23 percent, and the proportion rounded to multiples of 10 fell from 97 to 89 percent (also see Table A.2 in Appendix A).

In our computation of trends in upcoding, rounded values appear in the bracket above the threshold. Thus more precise measurement and less rounding will naturally increase the proportion of cases below DRG thresholds over time. These are clearly legitimate cases of *less downcoding* rather than more upcoding, and the trend in upcoding shown in the preceding section might thus overestimate the true trend. We have therefore repeated our analyses excluding all birth weights exactly equal to the threshold values to see how much this changes the general increase in the number of upcoded values after the introduction of DRGs (see Figure B.2 in Appendix B for details). Mainly, this modification increases the ratio of birth weights below to birth weights above the thresholds. If we assume that rounding occurs randomly, i.e. approximately the same number of cases are rounded up from below the threshold value and rounded down from above the threshold, eliminating all observations exactly at the threshold should result in values of $R_{25}$ in the vicinity of 50% in the years before DRGs were introduced or for non-DRG thresholds. This is exactly what happens, with two exceptions discussed before (1,000g and 1,500g), where for example already in 2000, around 65% of all non-rounded birth weights were below the threshold.

## 3.2 Excess reimbursement due to upcoding 2003-2010

We estimate the total amount $A$ that has been received unfairly by German hospitals, and perinatal centers in particular, by first estimating the absolute number of cases that have been upcoded at each relevant threshold $k$ in each year $y$, $\hat{U}_{yk}$. Then we multiply these estimates by the average additional reimbursement obtained by the hospital for that upcoded case, $\bar{\Delta}_{yk}$, and sum up across all years and thresholds:

$$A = \sum_{y=2003}^{2010} \sum_{k=1}^{8} \hat{U}_{yk} \bar{\Delta}_{yk} \tag{2}$$

In the following, we describe our calculation of $\hat{U}_{yk}$ and $\bar{\Delta}_{yk}$.

### 3.2.1 Estimating the number of upcoded cases

Our estimate of the number of upcoded cases in year $y$ at threshold $k$ is given by

$$\hat{U}^d_{yk} = N^d_{yk} - \hat{N}^d_{yk} \tag{3}$$

where $d$ indicates a bracket of width $d$ *below* each threshold, $N^d_{yk}$ is the *observed* number of births within that bracket, and $\hat{N}^d_{yk}$ is the *counterfactual* number of births in the same bracket that would have been recorded if there was no DRG-based reimbursement. To compute this counterfactual, we estimate the distribution around each threshold in a way that is reasonably independent of the actual distortions introduced by upcoding. This can be done parametrically, for instance by specifying a higher order polynomial, or it can be done non-parametrically, for instance by local polynomial regression with sufficient bandwidth. We chose local polynomial regression to estimate the log number of births (with degree=2; bandwidth=200g, Epanechnikov kernel). The result is depicted in Figure 4. It shows, separately for the pre-DRG-period and the DRG-period, the estimated distribution of birth weights between 500 and 2,750 grams as a smooth dark line. Light dots represent the actual number of births in each 10g interval in each year. The right panel of Figure 4, which shows the distribution of birth weighs since the introduction of DRGs, reveals a conspicuous excess number of births just below most

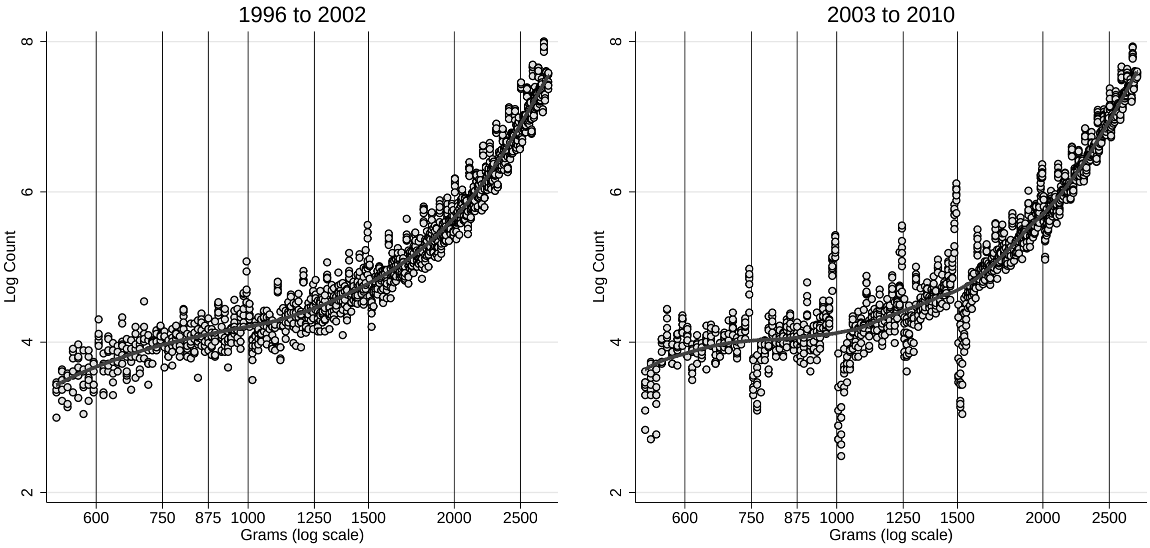

**Figure 4:** Distribution of birth weights before and after the introduction of DRGs. The smooth lines show the results of a local polynomial regression (degree 2, bw=200, Epanechnikov kernel). The light dots show the observed number of births in 10g brackets in each year. Source: Own calculations based on German birth records.

DRG-relevant thresholds and a corresponding shortfall in the number of births just above the thresholds. The excess is particularly large at 750g, 1,000g, 1,250g and 1,500g.

The next decision is to choose the width of the bracket left of the threshold *into which* cases are upcoded. This is a priori unclear. By inspection of Figure 4, it seems that most upcoded cases can be found either 10g or 20g below the threshold. For comparison, we have computed $\hat{U}^d_{yk}$ for $d = 10, 20, \ldots, 50$. The results are shown in Table 3. Some 7,000 cases are upcoded into the bracket 10 grams below the thresholds, and an additional 5,000 upcoded cases can be found in the 20g bracket. Moving further to the 30g bracket adds just 400 more upcoded cases. Then the total number decreases slightly and substantially increases when moving down to 50g below the threshold. The two latter findings are likely due to some rounding of birth weights to multiples of 50g. For our following calculations, we use a bracket width of 30g. Thus overall, we estimate that between 2003 and 2010, 12,133 births have been upcoded to a lower birth weight DRG.

Table 4 shows detailed results by year and threshold. First, it shows an increasing trend in the number of upcoded cases. In 2003, just after the introduction of DRGs, only 562 new-

**Table 3:** Estimated total number of upcoded cases, 2003 to 2010, by bracket width

| Bracket width in grams ($d$) | 10 | 20 | 30 | 40 | 50 |
|---|---|---|---|---|---|
| Estimated total number ($\hat{U}^d_{yk}$) | 6,970 | 11,758 | 12,133 | 11,971 | 15,086 |

borns were upcoded overall.[3] The number of annual upcodes has risen to more than 2,000 in 2010. Similar to the analysis above, the most important thresholds are at 1,000g, 1,500g, and 2,000g. These are salient numbers that can be easily memorized. Moreover, as noted before, 1,000g is the threshold for "extremely low birth weight" and 1,500g is the threshold for "very low birth weight".

**Table 4:** Estimated number of upcoded cases, by year and threshold.

| | Year | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Threshold | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | Total |
| 600 | 17 | 29 | 44 | 17 | 37 | 17 | 39 | 47 | 250 |
| 750 | 39 | 73 | 118 | 88 | 93 | 75 | 99 | 80 | 665 |
| 875 | 6 | 9 | -7 | 12 | 20 | 17 | 22 | 31 | 114 |
| 1000 | 178 | 277 | 220 | 231 | 275 | 245 | 259 | 259 | 1,941 |
| 1250 | 49 | 76 | 94 | 159 | 132 | 155 | 209 | 285 | 1,160 |
| 1500 | 290 | 362 | 477 | 437 | 559 | 569 | 628 | 583 | 3,908 |
| 2000 | 160 | 262 | 330 | 405 | 501 | 458 | 462 | 471 | 3,051 |
| 2500 | -179 | 55 | 182 | 35 | 245 | 211 | 238 | 254 | 1,044 |
| Total | 562 | 1,145 | 1,460 | 1,386 | 1,864 | 1,749 | 1,958 | 2,012 | 12,133 |

Source: Own calculations based on German birth records.

### 3.2.2 Estimating the excess reimbursement due to upcoding

Based on the DRG fee schedules for each respective year, we first calculate the average revenue per birth weight category as the weighted average of all DRG cost weights pertaining to a birth weight, with weights being given by the relative annual frequency of each DRG. In order to cal-

[3] This relatively low number is due to the fact that participation in the DRG system was voluntary in 2003 and became mandatory only in 2004. In the course of 2003, some two thirds of all acute care hospitals began to participate. Further, we obtain a substantial *negative* estimate for the 2,500g threshold in 2003. This happens primarily when cases are rounded to the threshold value itself, which clearly counteracts the effect of upcoding on reimbursement. We think that the amount "wasted" by hospitals because of downcoding should be subtracted from the amount received due to upcoding, although it is reasonable to assume that in 2003, mostly non-participating hospitals are among those who downcode.

culate reimbursement per birth weight category, we then multiply these average cost weights by the national average base rate. The national average is computed as the average of state-specific base rates weighted by the number of hospitals in each state. Reimbursement differentials and hence the average additional reimbursement for a specific upcoded case, $\bar{\Delta}_{yk}$, are obtained by subtracting the reimbursement for cases below a threshold value from the reimbursement for cases above.

**Table 5:** Estimated excess reimbursement in 1,000 Euro, by year and threshold

| | Year | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Threshold | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | Total |
| 600 | 158 | 210 | -227 | -23 | 662 | 216 | 32 | -155 | 872 |
| 750 | 553 | 524 | 882 | 421 | 617 | 1,338 | 1,631 | 847 | 6,813 |
| 875 | 55 | 162 | -71 | 193 | 297 | 178 | 367 | 319 | 1,500 |
| 1000 | 1,333 | 1,049 | 2,145 | 1,862 | 3,192 | 3,303 | 3,035 | 4,915 | 20,836 |
| 1250 | 237 | 639 | 1,206 | 1,152 | 1,055 | 1,083 | 1,416 | 2,221 | 9,009 |
| 1500 | 2,647 | 3,725 | 2,981 | 5,585 | 6,364 | 7,563 | 8,270 | 8,560 | 45,694 |
| 2000 | 1,100 | 1,910 | 2,460 | 3,044 | 4,480 | 3,890 | 4,440 | 4,741 | 26,065 |
| 2500 | -486 | 155 | 615 | 129 | 694 | 632 | 670 | 818 | 3,226 |
| Total | 5,598 | 8,374 | 9,990 | 12,363 | 17,361 | 18,203 | 19,861 | 22,265 | 114,015 |

Table 5 shows the total excess reimbursement due to neonatal upcoding. More than 114m Euro have been unfairly received by perinatal centers and other hospitals. The amount has continuously increased from 5.6m Euro in 2003 to 22.3m Euro in 2010. At the most salient threshold (1,500g) alone, some 45m Euro or 40% of the total additional reimbursement were obtained by deducting a few grams from the weight of some 4,000 premature newborns.

## 4 Economic incentives and upcoding

### 4.1 Reimbursement differentials and upcoding

We now examine the hypothesis that upcoding is particularly prevalent at financially salient thresholds, i.e. where the financial benefits of upcoding are particularly large in absolute terms. To that end, we have computed the *expected* reimbursement difference between adjacent DRGs in terms of birth weight at each relevant threshold and for each year. This is the payment

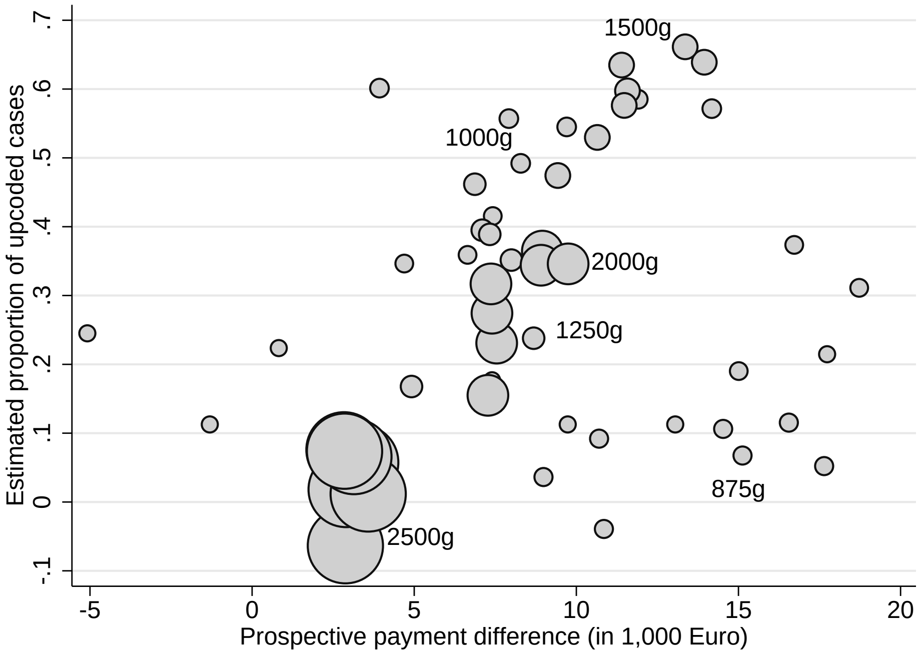

**Figure 5:** Proportion of upcoded cases around each threshold plotted against the expected payment difference in 1,000 Euro. The bubble size indicates the number of observations around each threshold.

difference that would be expected if one made a naive forecast of the relative weights in the current year based on the relative weights of the past year.

With seven years and eight thresholds, we have 56 data points overall.[4] Figure 5 plots the proportion of upcoded cases around each threshold against the expected payment difference in 1,000 Euro. The size of the bubbles indicates the absolute number of observations around each threshold. Overall, we find a positive relationship between financial incentives and upcoding in this graph. This relationship is driven by the five largest and quantitatively most important thresholds between 1,000g and 2,500g. A regression line drawn through the respective data points would yield a clear positive trend. At thresholds below 1,000g, the proportion of upcodes appears to be unrelated to the financial gain.

To quantify the relationship shown in Figure 5, we estimated the linear relationship using various specifications. The results are shown in Table 6. Column (1) shows the estimates for a regression line drawn through all data points by OLS and indicates a weakly significant 1 percentage point increase in the proportion of upcoded cases when the payment difference

[4] Prospective data for 2003 are not available.

rises by 1,000 Euro. When the data points are weighted by the number of observations around the threshold to reflect their quantitative importance, the slope increases to a highly significant 3.4 percentage points. As could be seen in Figure 5, the relationship is entirely driven by the thresholds above 1,000g. This becomes clearer when we split the sample and estimate separate regressions for thresholds below 1,000g (column (3)) and thresholds 1,000g and above (column (4)). While there is barely any relationship between the reimbursement differential and upcoding rates below 1,000g, the estimated increase at 1,000g and above is 5.4 percentage points per 1,000 Euro reimbursement differential.

**Table 6:** Relationship between the proportion of upcoded cases and expected payment difference

| | (1)<br>all thresholds | (2)<br>all thresholds | (3)<br>below 1,000g | (4)<br>1,000g and above |
|---|---|---|---|---|
| Payment difference | 0.010* | 0.034*** | -0.003 | 0.054*** |
| | (0.005) | (0.006) | (0.003) | (0.006) |
| Constant | 0.210*** | -0.031 | 0.221*** | -0.067 |
| | (0.047) | (0.035) | (0.042) | (0.063) |
| Weighted | no | yes | no | no |
| Observations | 56 | 56 | 21 | 35 |
| R-squared | 0.054 | 0.475 | 0.030 | 0.701 |

Note: Robust standard errors in parentheses; ***p<0.01, **p<0.05, *p<0.1

### 4.2 Hospital specialization and upcoding

Next, we analyze the relationship between hospital specialization and upcoding. Pregnant women at risk of delivering prematurely are routinely treated in perinatal centers with neonatal ICUs. If a child is accidentally born in a "lower level" hospital without specialized ward, it will be transferred as soon as possible to a neonatal care unit. In contrast to hospitals that actually care for low and very low birth weight infants, the non-specialized hospitals have no financial incentive to manipulate birth weights of babies born just above DRG-thresholds. Hence the proportion of upcoding should be higher in perinatal centers, because these hospitals have a stronger incentive to do so. Further, since low birth weights are rare events, midwives in regular birth clinics might also have no experience with the relevant DRGs and their substantial reimbursement differences at certain threshold. In other words, specialization might also entail a learning effect on the part of the hospital employees.

The birth record data do not allow us to link births directly to hospitals. But since we know the city or county of the hospital of birth, we use data on the existence of hospitals with neonatal intensive care units by county and year and link these with county-by-year estimates of upcoding rates. The latter are estimated as $2R_{50} - 1$ (see equation 1). This aggregate analysis allows us to draw some indirect conclusions about the link between specialization and upcoding simply because not all counties have perinatal centers with neonatal ICUs. In fact, there are about 220 Level 1 or Level 2 centers in Germany. In large cities such as Berlin or Munich, there are several of them. That means, on the other hand, that the majority of counties have no neonatal intensive care unit and mothers at risk of delivering prematurely who reside in one of those counties are usually treated in another county.

We demonstrate the link between specialization and upcoding graphically. The left panel of Figure 6 shows the proportion of upcoded newborns (aggregated across all eight DRG-thresholds) against the decile of the number of infants born with low birth weight (<2,500g). Small counties with very few such births are found on the left and counties with many such births are on the right of the graph. We further distinguish counties by the presence of a perinatal center. Vertical bars indicate 95% confidence intervals based on cluster-corrected standard errors.

Figure 6 clearly shows a strong positive association between the size of a county in terms of the number of underweight births and the proportion of upcoding. For the smallest counties we find on average *negative* upcoding rates of around 10%. Such downcoding can be explained by rounding to the nearest multiple of 50, which often is the DRG-threshold. In contrast, large counties have overall upcoding rates of more than 10%. As expected, there is not much overlap between counties with and without perinatal centers in terms of the total number of underweight infants. At two of the three deciles where we have overlap (5 to 7), we find significantly larger upcoding rates in counties with than in counties without perinatal centers. In sum, the findings shown in the left panel of Figure 6 support both the notion that

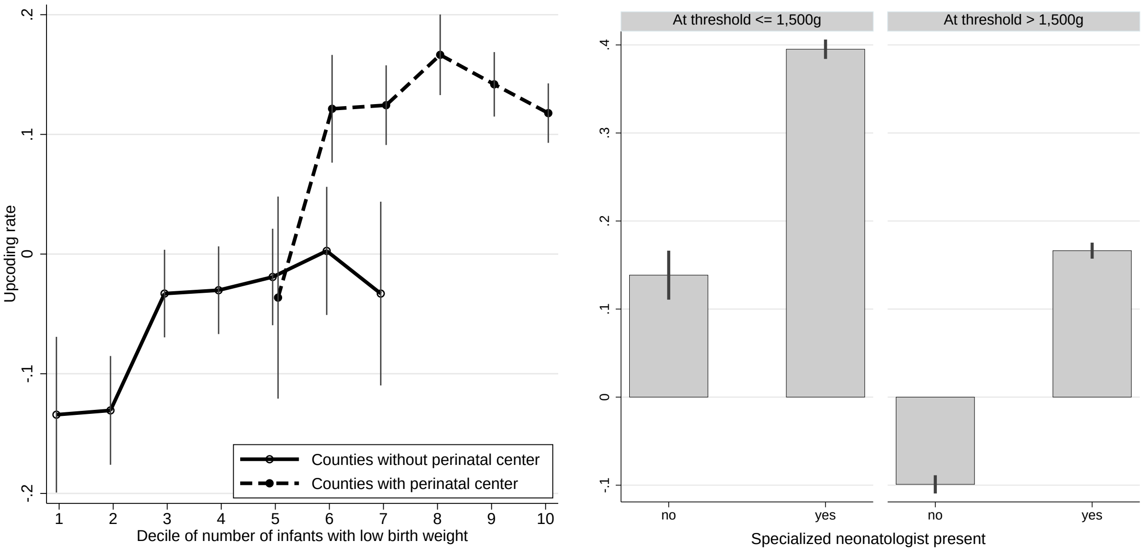

**Figure 6:** Proportion of upcoded newborns and 95% confidence intervals, by total number of babies born with low birth weight and county type, 2003-2010 (left panel) and by whether a neonatologist was present before birth and DRG-threshold, 2006-2011 (right panel).

county-level upcoding rates increase somewhat with the number of preterm births (and thus with experience) and largely with the presence of perinatal centers.

More direct evidence on higher upcoding rates in perinatal centers can be found in the right panel of Figure 6. Here, we show the proportion of upcoded cases up to the 1,500g threshold (left) and at the 2,000 and 2,500g thresholds (right) by whether a specialized neonatologist was present *before* the birth of the child. Having specialized neonatologists is one of the requirements to become a Level 1 or Level 2 perinatal center (see Section 1.2). The numbers shown are derived from individual-level records in the hospital quality data and are aggregated across all counties and all years from 2006 to 2011.[5] Consistent with findings in the preceding sections, we find larger overall upcoding rates at lower DRG thresholds. We also find consistent differences in upcoding rates related to the presence of a specialized neonatologist. Independent of the threshold, the estimated upcoding rates are 25 percentage points higher. This difference is substantial and clearly indicates that upcoding is less prevalent in non-specialized hospitals.

[5] In these data we are also not able to link births to individual hospitals.

### 4.3 Hospital ownership and upcoding

Casual reasoning may predict that for-profit hospitals should engage more in upcoding practices because managers have stronger incentives to maximize cash-flow (Sloan et al., 2001). Indeed, using Medicare claims data from the 1990s, Silverman and Skinner (2004) estimate that upcoding rates for DRGs related to respiratory infections were 60% larger in for-profit than in not-for-profit or government hospitals. Hospitals converting to for-profit status showed particular large increases in upcoding rates.

However, the relationship between hospital ownership and upcoding is theoretically ambiguous. Despite fundamentally different objectives of the management, it is a priori unclear whether for-profit (private) and not-for-profit hospitals behave differently. Managers in for-profit hospitals might support upcoding in order to increase the hospital owners' and their own income. Managers in not-for-profit hospitals might support upcoding to increase revenue that can be distributed according to the owners' or their own preferences (Duggan, 2002). Further, independent of ownership, midwives and doctors may have an incentive to upcode—also without active encouragement by the management—to increase the revenue of their unit and to secure their own future employment and income.

Since our data do not allow us to link births directly to hospitals, we use county-by-year data on the ownership of perinatal centers to analyze the relationship between ownership (rates) and upcoding rates. Since the former can be computed for counties with perinatal centers only, the analysis is restricted to those counties. Upcoding rates are again computed as $2R_{50} - 1$. We focus on upcoding rates for births up to the 1,500g threshold since most of these births take place in perinatal centers. In fact, the selected counties cover 98.3% of all births up to that threshold.[6]

We use two versions of the ownership variable. First we include all counties with perinatal centers; ownership status is then measured as the respective *proportion* of perinatal centers in each county. Second, we restrict the sample to counties with only one type of ownership in one year. Here ownership is measured as a binary variable. Control variables include year

[6] Results for the 2,000g and 2,500g thresholds are qualitatively similar to those shown here. See Table A.3 in Appendix A.

dummies interacted with dummies indicating the (decile of the) total number of births below 1,500g and federal state dummies. For each version of the ownership variable, we estimate pooled OLS models (with cluster-corrected standard errors) and fixed effects models.

**Table 7:** Linear regression estimates of relationship between hospital ownership and upcoding rates (DRG thresholds up to 1,500g, only counties with perinatal centers), county-by-year data, 2003 to 2010

| | OLS | | Fixed Effects | |
|---|---|---|---|---|
| | Prop.<br>(1) | Binary<br>(2) | Prop.<br>(3) | Binary<br>(4) |
| Not-for-profit perinatal centers | 0.065** | 0.068** | 0.089 | 0.079 |
| | (0.028) | (0.029) | (0.104) | (0.119) |
| For-profit perinatal centers | -0.010 | -0.007 | 0.013 | 0.064 |
| | (0.053) | (0.054) | (0.076) | (0.081) |
| Year × n of births dummies | x | x | x | x |
| State dummies | x | x | | |
| County dummies | | | x | x |
| Mean dep. variable | 0.378 | 0.386 | 0.378 | 0.386 |
| Observations | 1,418 | 1,270 | 1,418 | 1,270 |
| Number of counties | 195 | 180 | 195 | 180 |

Notes: Standard errors for OLS models are cluster-corrected (by county), *p<0.1, **p<0.05, ***p<0.01. In models (1) and (3), the sample is restricted to counties with *at least one* perinatal center; ownership status is measured as the respective *proportion* of perinatal centers in each county. In models (2) and (4), the sample is restricted to counties with perinatal centers of the same ownership status. Hence ownership is measured as a binary variable.

Results are shown in Table 7. We find that upcoding rates are largest in not-for-profit hospitals. They are 6 to 9 percentage points larger than in public hospitals (the reference group). The point estimates for the difference between not-for-profit and public hospitals are reasonably consistent across estimation methods and definitions of the ownership variable, yet standard errors are substantially larger when including county-fixed effects. This was to be expected because there is only little variation in ownership within counties across time. The point estimates for for-profit hospitals are close to zero in the OLS models and the first fixed effects model (where we use the proportion of ownership as explanatory variable). The difference between for-profit and public hospitals jumps to 6 percentage points when we restrict the sample to counties where ownership status is homogeneous. However, due to the large standard errors, this difference is not significant and we are not able to claim without reasonable doubt that for-profit hospitals are more likely to upcode than public hospitals. What seems fairly robust is that they are not more likely to upcode than not-for-profit hospitals. Thus turning not-for-profit into for-profit perinatal centers has most likely not increased upcoding rates on these hospitals.

### 4.4 Regional market conditions and upcoding

The early literature on supplier-induced demand (SID) has found strong links between per capita health care utilization and the regional supply of physicians measured by physician density (e.g. Fuchs (1978)). The idea behind this literature is that conditional on population size and population health, the demand for medical services per doctor decreases with the number of doctors. Doctors react by inducing demand for their services—exploiting the information asymmetry between themselves and their patients—even beyond what is medically necessary.

Supplier-induced demand in neonatal intensive care has recently been studied by Freedman (2012), who shows that an exogenous increase in the number of available beds in neonatal intensive care units (NICU) increases a baby's likelihood of being admitted to such a unit. This finding is clearly reminiscent of the formulation of supplier-induced demand known as Roemer's law: "a built bed is a filled bed" (Roemer, 1961). Notably, Freedman's results are almost exclusively driven by infants with birth weights between 1,500g and 2,500g, whereas empty NICU beds have much smaller effects on the admission of infants with birth weights below 1,500g. Thus demand inducement appears to be more prevalent when doctors have some leeway in their decision-making.

Deliberately filling empty beds means that a hospital increases its revenue by raising quantity. But one of the main rationales of DRGs is to make this strategy unprofitable. With DRGs, upcoding newborns has the same ultimate goal of raising revenue, but it is reached by increasing the reimbursement for treating a given newborn. Since there are considerable fixed costs of setting up a NICU, a large number of empty beds drives up average costs and thus may induce hospitals to engage in upcoding.

To study this hypothesis, we provide evidence on the relationship between regional supply of beds in neonatology and upcoding rates. We start out by defining regional markets for neonatal intensive care as neonatal care referral regions (NCRRs). The level of analysis in the preceding subsections was the county of birth. However, since the number of Level 1 perinatal centers is smaller than the number of counties, regional markets for newborn intensive care are clearly wider than single counties.

We define NCRRs as follows: first, we calculate empirical "migration matrices" relating the county of maternal residence and the county of birth of all preterm infants (lighter than 1,500g) in the period 2003-2010. A county is defined as the center of a NCRR if the majority of preterm infants born to mothers living in this county are born in this county. If the majority of preterm infants are born in another county then the county of maternal residence belongs to the NCRR defined by the county of birth. To illustrate, consider Table 8. Mothers residing in some county A may have given birth to 100 preterm infants, of which 70 were born in county A, 20 in county B, and 10 in county C. Then county A belongs to the NCRR A. Mothers residing in county B have given birth to 50 preterm infants, of which 30 were born in county A, 15 in county B, and 5 in county C. Then county B belongs to NCRR A. Now, Table 8 shows a third county C which sent the majority of its mothers at risk of preterm delivery to county B and not A. In this case, it would be unclear if county C belonged to NCRR A or some "new" NCRR B. Fortunately, we did not have to decide this problem in our data. All counties either were at the center of an NCRR or they belonged to an NCRR defined by such a county. Overall, we have arrived at a total of 159 NCRRs.

**Table 8:** Illustration of a migration matrix to define neonatal care referral regions (NCRRs)

| | | County of birth | | | |
|---|---|---|---|---|---|
| | | A | B | C | Total |
| County of residence | A | 70 | 20 | 10 | 100 |
| | B | 30 | 15 | 5 | 50 |
| | C | 20 | 30 | 20 | 70 |

For each NCRR in each year from 2003 to 2010, we have computed the total number of preterm births, the proportion of upcoded cases, the number of beds in neonatology, and the number of women aged 15-45. Table 9 shows the coefficient of interest of twelve different regression specifications. First, we use two definitions of the supply variable: the number of neonatology beds per 1,000 women aged 15-45 and the number of neonatology beds per birth <1,500g. Note that we do not have information specifically on NICU beds but only on the sum of all neonatology beds. Second, based on the findings in Freedman (2012), we distinguish between upcoding at lower thresholds and higher thresholds and also estimate joint models for all thresholds. Third, we estimate OLS and Fixed Effects models. In each model, we include year dummies, regional ownership rates of neonatal care centers, and a third-order polynomial of the total number of births <1,500g to account for size effects.

The results shown in Table 9 do not confirm our expectations. If anything, the relative supply of beds in neonatal care seems to be associated with a lower upcoding rate. We have only one possible but not very satisfying explanation for this counter-intuitive result: the number of neonatology beds reported in the hospital directory is a severely biased measure of the number of intensive care beds. Perhaps due to the general discussion on oversupply of perinatal centers in Germany, the hospital directory data as delivered by the German Federal Statistics Bureau are partly incomplete. Especially large (public) university hospitals have stopped reporting the number of beds in neonatology at about the same time that discussion about minimum quantities started, and we had to fill in the gaps manually on the basis of hospital websites or information from the last available systematic report on neonatal care facilities in Germany (Gesellschaft für Neonatologie und Pädiatrische Intensivmedizin, 2001), which describes the situation in 1999. It is conceivable that hospitals in counties with considerable oversupply were most hesitant to correctly report the number of beds in neonatology, thereby inducing systematic measurement error in our main explanatory variable, so that our results are affected in unpredictable ways.

## 5 Newborn health status and upcoding

Even conditional on DRG classification criteria, newborns are in different health states. Worse health usually means higher expected treatment costs incurred by the hospital. It is not possible to select newborns based on their health status. Instead, hospital staff might selectively upcode

**Table 9:** Association between supply of neonatal care beds and upcoding, by neonatal care referral region (NCRR) and year (2003-2010).

| | Upcoding rate ≤ 1,500g | | Upcoding rate > 1,500g | | Total upcoding rate | |
|---|---|---|---|---|---|---|
| | OLS | FE | OLS | FE | OLS | FE |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Neonatology beds per 1,000 women | -0.024 | -0.444** | 0.015 | -0.201 | 0.027 | -0.274** |
| | (0.079) | (0.208) | (0.042) | (0.147) | (0.046) | (0.117) |
| Neonatology beds per birth <1,500g | -0.030 | -0.027 | -0.012 | -0.004 | -0.034* | -0.033 |
| | (0.036) | (0.076) | (0.020) | (0.034) | (0.020) | (0.032) |
| Observations | 1,272 | 1,272 | 1,272 | 1,272 | 1,272 | 1,272 |
| Number of NCRRs | 159 | 159 | 159 | 159 | 159 | 159 |

Notes: Robust standard errors in parentheses; *p<0.1, **p<0.05, ***p<0.01; Control variables: Year dummies, hospital ownership rates, third-order polynomial of total number of births <1,500g per NCRR and year .

newborns who are likely to cost more - based on observable characteristics that are not part of the DRG classification such as being small, having low gestational age, or having a low APGAR score. Such systematic behavior can be expected for instance if doctors or midwives believe that upcoding is unethical but that it is necessary to align reimbursement and expected treatment costs. If this conjecture is true, we would expect that infants with a birth weight recorded just above a DRG-threshold should be on average healthier than those with a birth weight recorded just below the threshold. More precisely, health measures should exhibit a discontinuous jump at the DRG-threshold.[7]

The birth register data only include birth length as measure of infant health other than birth weight. Thus we now also use complementary data, namely information collected for the purpose of external quality control in hospitals. These data cover all births in German hospitals from 2006 to 2011, including information on birth weight, gestational age, APGAR scores, and early neonatal (≤7 days) mortality taken from medical charts.

To study whether infants just above DRG birth weight thresholds are healthier than those below, we estimated local polynomial regressions of gestational age (in days) and length (in mm) on recorded birth weight and analyzed the residuals from these regressions.[8] Figure

[7] On motivation to upcode could arise from medical guidelines if recommended treatment intensity was conditional on birth weight classifications. In that case, doctors could “upgrade” the treatment of newborns by recording birth weights below thresholds. However, to the best of our knowledge, clinical guidelines in Germany neonatology do not condition on birth weight.

[8] Scatterplots and local polynomial regression lines of the original relationship between birth weight and gestational age, etc., are shown in Figures B.3 to B.6 in Appendix B).

7 shows these residuals by 25g-birth weight bracket (top panel) together with the (square root of the) underlying number of observations (bottom panel). Positive residuals indicate "excess" age or length, meaning that infants in this birth weight bracket are older or longer than expected. Gestational age and birth length are measures of maturity and proxies for health independent of birth weight. Neither is relevant for reimbursement, so there is no incentive to manipulate these data. Further, gestational age is already determined at the time of birth and cannot be manipulated conditional on observed birth weight.[9]

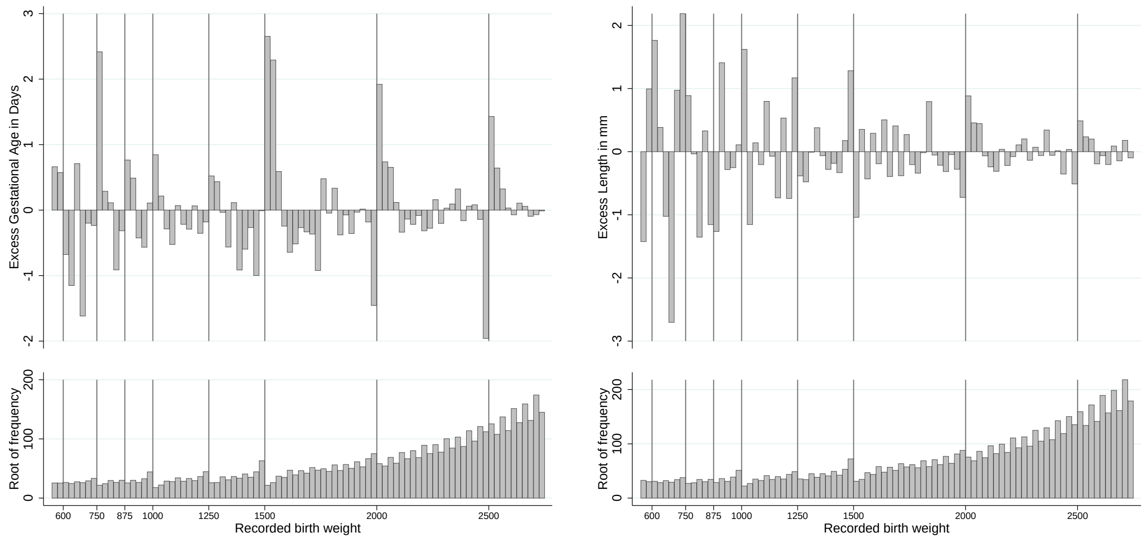

**Figure 7:** Excess gestational age and birth length (top panel) and number of observations (bottom panel) by 25g birth weight bracket. Data sources: German hospital data 2006-2011 (gestational age) and birth records 2003-2010 (length).

Especially the left graph paints a fairly consistent picture of systematic upcoding of children with relatively high expected treatment costs (and perhaps worse unobserved health). Gestational age is negatively linked to expected costs because younger infants should have a higher average length of stay. At each DRG-threshold except 600g, we find that those in the two 25g-brackets above the threshold (i.e. with a recorded birth weight up to 49g higher than threshold) have higher than expected gestational age; the excess is statistically significant at 750, 1500, 2000 and 2500g. Put differently, infants with low gestational age are systematically

[9] Gestational age can also be determined after childbirth using a set of clinical criteria (Finnström, 1977). However, in our data, it is calculated as the difference between date of birth and the beginning of the last menstrual period.

shifted below the DRG-birth weight threshold. For birth length, we find systematic upcoding of “short” infants only at the 2000 and 2500g thresholds.

Next, we look at the APGAR scores. APGAR is the most common measure of neonatal health (Casey et al., 2001). Newborns are scored 0 to 2 on five subscales for appearance, pulse, grimace, activity, and respiration one, five and ten minutes after birth. The sum of the subscores yields the total APGAR score with range from 0 to 10. Scores below 4 are considered critically low. Similar to the analysis above, we estimated local polynomial regressions of the proportion of infants with a critically low (0-3) or normal (7-10) one minute APGAR score on birth weight (the APGAR score arguably being measured before birth weight), and we computed the excess risk as the ratio of actual and expected proportion within each birth weight bracket (see Figure 8). Stillborn infants, who have an APGAR score of zero, are excluded. Let us first look at the 2,000g and 2500g thresholds. Here we find results that are consistent with the finding above: the proportion of infants with a critically low APGAR score is relatively low right of the threshold and relatively large left of the threshold. This can be explained by low APGAR infants being shifted below the DRG-threshold. At the same time, there are systematically more (healthy) infants with a normal one minute APGAR score right of the thresholds.

At thresholds below 2,000g, Figure 8 seems to suggest an opposite tendency. There are systematically less sick children left of the threshold and more sick children right of the threshold. In other words, at birth weights up to the 1,500g threshold, it is healthier infants who appear to be systematically upcoded, and unhealthy infants with critical APGAR scores tend not to be upcoded.

We can use our data to roughly estimate the effect of having a low versus a normal APGAR score on being upcoded by calculating $2R_{25} - 1$ separately for low and normal APGAR score infants and taking differences. Table 10 shows that low APGAR infants are about 10 percentage points *less* likely to be upcoded than normal APGAR infants at all threshold below 2,000g except 875g. This difference is statistically significant at the 1,000g, 1,250g, and 1,500g thresholds. In contrast, at the 2,000g and 2,500g thresholds, low APGAR infants are significantly *more* likely to be upcoded. The difference between DRG-thresholds below and above 1,500g thus mirrors the results we found for birth length.

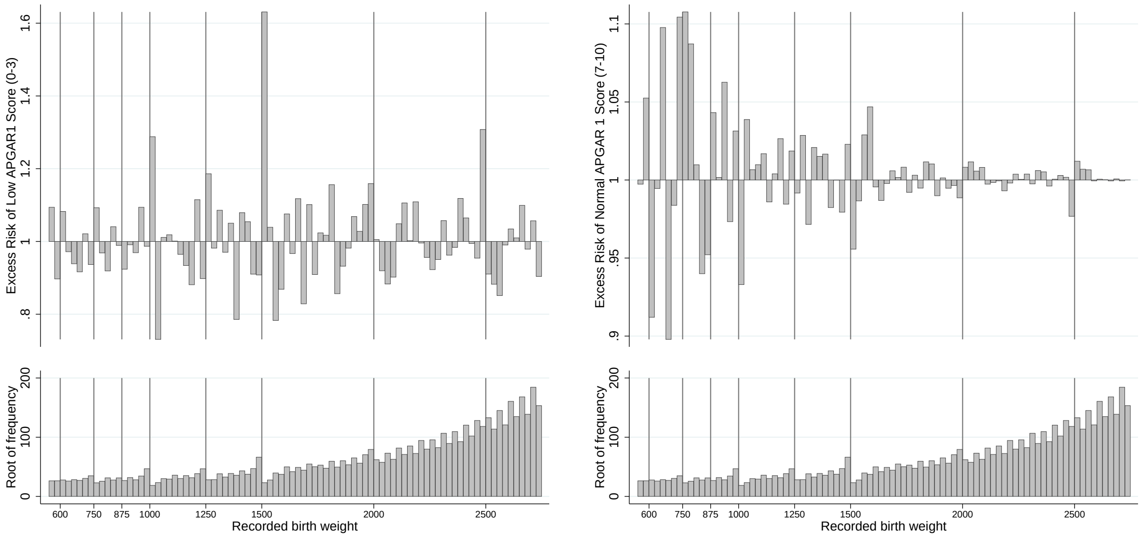

**Figure 8:** Excess risk of having a low or high one minute APGAR-Score by birth weight bracket. German hospital data 2006-2011.

**Table 10:** Number of infants with birth weight around DRG-thresholds, by APGAR score. German hospital data 2006-2011.

| | Low APGAR score | | | Normal APGAR score | | | Difference | |
|---|---|---|---|---|---|---|---|---|
| Threshold | *N* below | *N* above | % upcoded | *N* below | *N* above | % upcoded | % | *p*-value |
| 600 | 237 | 295 | -11 | 455 | 476 | -2 | -9 | 0.114 |
| 750 | 285 | 136 | 35 | 924 | 389 | 41 | -5 | 0.300 |
| 875 | 184 | 121 | 21 | 791 | 601 | 14 | 7 | 0.277 |
| 1,000 | 320 | 63 | 67 | 1 860 | 281 | 74 | -7 | 0.089 |
| 1,250 | 196 | 91 | 37 | 1 976 | 695 | 48 | -11 | 0.041 |
| 1,500 | 290 | 61 | 65 | 4 103 | 472 | 79 | -14 | <0.001 |
| 2,000 | 279 | 143 | 32 | 6 012 | 3 699 | 24 | 8 | 0.082 |
| 2,500 | 325 | 274 | 9 | 13 645 | 17 384 | -12 | 21 | <0.001 |

Notes: *N* calculated from 25g brackets; *p*-values based on Fisher's exact test.

We conclude from this analysis that upcoding is systematically related to non-DRG relevant health indicators of the newborn. However, this relationship is complex. At higher birth weights, unhealthy infants are systematically upcoded. To the extent that these indicators reflect expected treatment costs not accounted for in DRGs, this can be interpreted as hospitals upcoding in particular newborns for which they expect high treatment costs. At lower birth weights, where infants' health status is more critical, unhealthy infants are systematically less often upcoded (with the notable exception of infants with low gestational age, which appear to be upcoded at all thresholds). One explanation for systematically not upcoding sicker infants could be that sicker children have a smaller survival probability and thus invoke both lower costs and lower reimbursements. If a newborn dies within 28 days, relative DRG-weights are reduced by 85%. The expected financial gain from upcoding is thus comparatively small when the probability of dying is large. Another explanation for differential upcoding by health could be the perinatal centers' mandate to publish standardized quality reports. These quality reports contain outcomes by birth weight categories defined by thresholds similar to (but slightly different from) those used for reimbursement: 500g, 750g, 1,000g, 1,250g, and 1,500g. Shifting comparatively healthy newborns below those threshold may lead to better average reported outcomes conditional on birth weight categories.

**Table 11:** Number of infants with birth weight around DRG-thresholds, by early neonatal mortality. German hospital data 2006-2011.

| | Died within 7 days | | | Survived 7 days | | | Difference | |
|---|---|---|---|---|---|---|---|---|
| Threshold | *N* below | *N* above | % upcoded | *N* below | *N* above | % upcoded | % | *p*-value |
| 600 | 128 | 149 | -8 | 527 | 539 | -1 | -6 | 0.338 |
| 750 | 81 | 47 | 27 | 1056 | 428 | 42 | -16 | 0.061 |
| 875 | 33 | 20 | 25 | 870 | 652 | 14 | 10 | 0.460 |
| 1,000 | 56 | 17 | 53 | 1955 | 306 | 73 | -20 | 0.018 |
| 1,250 | 28 | 18 | 22 | 2013 | 721 | 47 | -26 | 0.052 |
| 1,500 | 31 | 20 | 22 | 4069 | 494 | 78 | -57 | <0.001 |
| 2,000 | 12 | 29 | -41 | 5816 | 3453 | 25 | -67 | <0.001 |
| 2,500 | 11 | 15 | -15 | 11749 | 14178 | -9 | -6 | 0.758 |

NOTES: *N* calculated from 25g brackets; *p*-values based on Fisher's exact test.

We can address the question how upcoding is related to subsequent mortality by looking at early neonatal mortality. Table 11 shows the same information as for APGAR scores in Table 10. An infant who dies within 7 days after birth is systematically less likely to be upcoded. The only exception can be found yet again at the 875g threshold. Percentage point differences are substantial (up to 67 percentage points) and almost always statistically significant. Again, this supports the notion that very sick children with low expected treatment costs are not as often upcoded as healthier children. Moreover, infant mortality is regularly reported in quality reports on neonatal intensive care.

There is one caveat here, however, namely that the relationship between upcoding and early neonatal mortality might actually be due to births in low volume, low quality hospitals where infant mortality is higher and upcoding far less common than in large perinatal centers (see Section 4.2). Since the scientific use files of the hospital quality data do not allow us to identify individual hospitals, we cannot test this explanation.

## 6 Summary and conclusion

In this paper we show that since the introduction of DRGs in German neonatal care, birth weights around thresholds relevant for reimbursement are increasingly manipulated, i.e. they are systematically shifted below the thresholds. We estimate that, between 2003 and 2010, about 12,000 newborn with birth weights above the DRG thresholds have been recorded as having a birth weight below the threshold. At some particularly salient DRG thresholds the number of infants with birth weights coded just below the threshold exceeds the number of infants just above the threshold by a factor of ten. As a result, hospitals have received excess reimbursement of 114m Euro in our study period. The current annual numbers are around 2000 upcodes that yield an additional 20m Euro.

We demonstrate that birth weight manipulation is systematically related to economic incentives. First, the proportion of upcoded cases is related to the reimbursement differential at the respective threshold. Second, systematic upcoding can almost exclusively be found in counties that have specialized perinatal care centers. In counties without such centers and where adequate care is not available so that very low birth weight infants are transferred to the perinatal care centers, we find no evidence for systematic upcoding. This is because hospitals that are not perinatal centers have no financial incentive to manipulate birth weights. Third, we examined whether public, not-for-profit and for-profit hospitals differ in the extent of upcoding.

We find significantly higher than average rates in counties with not-for-profit perinatal centers. Depending on specification, counties with for-profit hospitals also have higher upcoding rates than public hospitals, but that difference is not statistically significant. Finally, we examine the relationship between upcoding and measures of regional supply (or regional competition), i.e. the density of hospital beds in neonatology. We hypothesized that dense markets (where supply is large relative to demand) would also exhibit higher upcoding rates, but we rather find the opposite relationship. One reason for this counter-intuitive finding may be that the data on the number of neonatal intensive care beds in Germany have quality issues.

Further, we have studied whether hospitals are more likely to upcode newborns that appear to be less healthy and thus more costly to care for, conditional on recorded weight. We do in fact find that low gestational age systematically increases the chances of being upcoded. However, for alternative health measures such as APGAR scores and early neonatal mortality, the relationship is more complex. We find that at DRG thresholds up to 1,500g, newborns with critically low APGAR scores are significantly less often upcoded, whereas at the two largest DRG thresholds of 2,000g and 2,500g, they are more often upcoded. One explanation for this finding is that a critically low APGAR score is a significant predictor or neonatal death. Since reimbursement is substantially reduced if a child dies within 28 days, incentives to upcode children with low survival chances are comparatively small. Findings related to early neonatal mortality confirm this explanation: neonatal death within 7 days is related to significantly lower upcoding rates.

The results in our paper have several implications. The first is methodological. In a recent paper, Almond et al. (2010) examine the effect of medical care on the health of low birth weight infants in the U.S. exploiting a discontinuity in treatment provision around the 1,500g threshold. They also find a corresponding discontinuity in outcomes (28-day mortality), with babies below the threshold (and thus more intense treatment) having higher survival changes than babies above the threshold. Although the results of this specific application of a regression discontinuity design around birth weight threshold has been shown to critically depend on the inclusion of rounded values at exactly 1,500g (Barreca et al., 2011), the general approach to study causal effects of medical care on infant health seems useful. However, the relevance of our findings for any study in Germany that aims at exploiting regression discontinuity designs along the lines of Almond et al. is obvious. Since the “treatment determining variable” is clearly manipulated systematically along infant health such a design is unfeasible.

A second implication relates to the nature of cheating in an environment where people hold specific ethical standards but cheating is virtually undetectable. A purely rational model of crime would predict that it does not matter which infant is upcoded. From the perspective of the hospital the treatment costs of an admitted newborn are sunk and the best strategy is to upcode indiscriminately. Our data are clearly at odds with this simple model. Indicators of neonatal health that are related to expected treatment costs are significant predictors of upcoding. We interpret this finding as evidence that doctors and nurses find it easier to manipulate birth weight if they can justify their actions by aligning expected treatment costs and reimbursement.

Finally, our paper holds a message concerning optimal reimbursement in neonatal care. Obviously, there are large differences between actual treatment costs and reimbursement for birth weights close to the thresholds simply by construction of the DRGs. Immediately above each threshold, actual average treatment costs are much larger than reimbursements, and immediately below each threshold, they are much smaller. It is thus no surprise that hospitals try to align expected costs and reimbursements. The conclusion from our findings is that the current German DRG-classification relying to a large extent on birth weight is inappropriate because it is too easy to manipulate. But what are the alternatives? First, one could revert to the former per diem reimbursement for each day a newborn receives intensive care, although this creates well-known incentives to increase length of stay in case of a uniform per diem rate. Second, one could reimburse hospitals on the basis of gestational age. Clearly, the day of conception can be calculated and documented long before childbirth and the related records cannot be manipulated easily. But this criterion is also problematic. If substantial decreases in reimbursement are related to merely one additional day of pregnancy, this creates incentives to hospitals and doctors *not to* arrest labor, which—in contrast to recording the wrong birth weight—may be harmful to the newborn’s health. As a third alternative, one could reimburse hospitals based on a smooth function of birth weight which could be easily estimated using the data routinely collected to compute relative DRG weights. Unfortunately, even this scheme would not be perfect. There would still be small marginal gains in manipulating birth weights along the entire birth weight distribution, and it is a priori unclear if in the end the sum of many small gains does or does not outweigh the sum of few large gains generated in the present system. Ultimately one has to conclude that it is hard to design the perfect reimbursement system. But of course, this is old news to health economists.

# References

Abler, S., Verde, P., Stannigel, H., Mayatepek, E., and Hoehn, T. (2011). Effect of the introduction of diagnosis related group systems on the distribution of admission weights in very low birthweight infants. *Arch Dis Child Fetal Neonatal Ed*, 96:F186–F189.

Almond, D., Doyle, J. J., Kowalski, A. E., and Williams, H. (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *The Quarterly Journal of Economics*, 125(2):591–634.

AOK-Bundesverband (2012). *Basisfallwerte (Zahlbetrag) aller DRG-Krankenhäuser 2003-2012.* Federal Association of the AOK.

Barreca, A. I., Guldi, M., Lindo, J. M., and Waddell, G. R. (2011). Saving babies? Revisiting the effect of very low birth weight classification. *The Quarterly Journal of Economics*, 126(4):2117–2123.

Casey, B. M., McIntire, D. D., and Leveno, K. J. (2001). The continuing value of the APGAR score for the assessment of newborn infants. *The New England Journal of Medicine*, 344(7):467–471.

Committee on Fetus and Newborn (2012). Levels of neonatal care. *Pedriatrics*, 120(3):587–597.

Duggan, M. (2002). Hospital market structure and the behavior of not-for-profit hospitals. *RAND Journal of Economics*, 33(3):433–446.

Ellis, R. P. and McGuire, T. G. (1993). Supply-side and demand-side cost sharing in health care. *Journal of Economic Perspectives*, 7(4):135–151.

Finnström, O. (1977). Studies on maturity in newborn infants. IX. Further observations on the use of external characteristics in estimating gestational age. *Acta Paediatr Scand*, 66(5):601–604.

Freedman, S. (2012). Capacity and utilization in health care: The effect of empty beds on neonatal intensive care admission. Unpublished working paper, Indiana University.

Fuchs, V. (1978). The supply of surgeons and the demand for operations. *Journal of Human Resources*, 13(Suppl.):35–56.

Geissler, A., Scheller-Kreinsen, D., Quentin, W., and Busse, R. (2011). Germany: Understanding G-DRGs. In Busse, R., Geissler, A., Quentin, W., and Wiley, M., editors, *Diagnosis-Related Groups in Europe*, pages 243–271. World Health Organization.

Gerber, A., Lauterbach, K., and Lüngen, M. (2008). Manchmal ist weniger mehr – Wie viele Perinatalzentren der Level 1 und 2 sind in Deutschland gesundheitspolitisch notwendig und finanzierbar? *Deutsches Ärzteblatt*, 105(26):A1439–A1441.

Gesellschaft für Neonatologie und Pädiatrische Intensivmedizin (2001). *Versorgung von Frühgeborenen in Deutschland: Zentren, Frauenkliniken und Kinderkliniken.*

Hoehn, T., Drabik, A., Lehmann, C., Stannigel, H., and Mayatepek, E. (2008). Correlation between severity of disease and reimbursement of costs in neonatal and pediatric intensive care patients. *Acta Paediatrica*, 97:1438–1442.

InEK (2007). *Kalkulation von Fallkosten. Handbuch zur Anwendung in Krankenhäusern. Version 3.0.* Institut für das Entgeltsystem im Krankenhaus gGmbH.

InEK (2009). *G-DRG Fallpauschalen Katalog 2010.* Institut für das Entgeltsystem im Krankenhaus gGmbH.

InEK (2011). *Abschlussbericht zur Weiterentwicklung des DRG-Systems für das Jahr 2012.* Institut für das Entgeltsystem im Krankenhaus gGmbH.

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45(6):633–644.

Müller, C., Weiß, U., von Schnakenburg, C., and Hentschel, R. (2007). The significance of DRG refunds in neonatology. Examples of calculations for a perinatal centre based on case numbers for 2009. *Monatsschr Kinderheilkd*, 10:947–953.

Roemer, M. I. (1961). Bed supply and hospital utilization: a natural experiment. *Hospitals*, 35:36–42.

Shigeoka, H. and Fushimi, K. (2013). Supply induced demand in newborn treatment: Evidence from Japan. Unpublished working paper, Columbia University.

Silverman, E. and Skinner, J. (2004). Medicare upcoding and hospital ownership. *Journal of Health Economics*, 23:369–389.

Sloan, F. A., Picone, G. A., Jr., D. H. T., and Chou, S.-Y. (2001). Hospital ownership and cost and quality of care: is there a dime?s worth of difference? *Journal of Health Economics*, 20(1):1 – 21.

Vogl, M. (2012). Assessing DRG cost accounting with respect to resource allocation and tariff calculation: the case of Germany. *Health Economics Review*, 2:15.

# A Supplementary Tables

**Table A.1:** Number of births 1996-2010, by year and birth weight category. Data before 2000 do not include births in the state of Bavaria. Source: Own calculations based on German birth records.

| Year | <1,000g | 1,000g-1,499g | 1,500g-2,499g | >2,500+ | Total |
|---|---|---|---|---|---|
| 1996 | 2,583 | 4,120 | 34,220 | 625,780 | 666,703 |
| 1997 | 2,675 | 4,268 | 35,250 | 639,368 | 681,561 |
| 1998 | 2,800 | 4,286 | 34,999 | 616,391 | 658,476 |
| 1999 | 2,833 | 4,331 | 35,236 | 604,884 | 647,284 |
| 2000 | 3,241 | 5,074 | 40,962 | 717,854 | 767,131 |
| 2001 | 3,298 | 4,814 | 39,976 | 686,255 | 734,343 |
| 2002 | 3,301 | 4,880 | 40,097 | 670,972 | 719,250 |
| 2003 | 3,281 | 4,696 | 39,874 | 658,870 | 706,721 |
| 2004 | 3,509 | 4,823 | 40,651 | 656,639 | 705,622 |
| 2005 | 3,408 | 4,683 | 38,427 | 639,279 | 685,797 |
| 2006 | 3,424 | 4,622 | 37,771 | 626,905 | 672,722 |
| 2007 | 3,653 | 4,874 | 38,451 | 637,892 | 684,870 |
| 2008 | 3,464 | 4,771 | 37,966 | 636,305 | 682,506 |
| 2009 | 3,434 | 4,833 | 37,715 | 619,142 | 665,124 |
| 2010 | 3,476 | 4,967 | 38,303 | 631,201 | 677,947 |

**Table A.2:** Proportion of rounded birth weights, 1996 to 2010 (excludes Bavaria). Multiples of 100, 50, and 10. Source: Own calculations based on German birth records.

| Year | 100s | 50s | 10s |
|---|---|---|---|
| 1996 | 0.172 | 0.311 | 0.974 |
| 1997 | 0.168 | 0.303 | 0.970 |
| 1998 | 0.163 | 0.296 | 0.964 |
| 1999 | 0.159 | 0.288 | 0.955 |
| 2000 | 0.156 | 0.283 | 0.946 |
| 2001 | 0.154 | 0.278 | 0.938 |
| 2002 | 0.148 | 0.269 | 0.929 |
| 2003 | 0.145 | 0.264 | 0.922 |
| 2004 | 0.142 | 0.258 | 0.915 |
| 2005 | 0.139 | 0.253 | 0.909 |
| 2006 | 0.137 | 0.249 | 0.904 |
| 2007 | 0.134 | 0.245 | 0.900 |
| 2008 | 0.132 | 0.242 | 0.898 |
| 2009 | 0.131 | 0.240 | 0.891 |
| 2010 | 0.128 | 0.235 | 0.888 |

**Table A.3:** Linear regression estimates of relationship between hospital ownership and upcoding rates (DRG thresholds up to 2,000g and 2,500g, only counties with perinatal centers), county-by-year data, 2003 to 2010

| | OLS | | Fixed Effects | |
|---|---|---|---|---|
| | Prop. | Binary | Prop. | Binary |
| | (1) | (2) | (3) | (4) |
| Not-for-profit perinatal centers | 0.0464** | 0.0507*** | 0.0258 | 0.0421 |
| | (0.0182) | (0.0187) | (0.0549) | (0.0636) |
| For-profit perinatal centers | -0.0354 | -0.0333 | 0.0382 | 0.0454 |
| | (0.0313) | (0.0299) | (0.0400) | (0.0430) |
| Year × n of births dummies | x | x | x | x |
| State dummies | x | x | | |
| County dummies | | | x | x |
| Mean dep. variable | 0.139 | 0.141 | 0.139 | 0.141 |
| Observations | 1,418 | 1,270 | 1,418 | 1,270 |
| Number of counties | 195 | 180 | 195 | 180 |

NOTES: Standard errors for OLS models are cluster-corrected (by county), *p<0.1, **p<0.05, ***p<0.01. In models (1) and (3), the sample is restricted to counties with *at least one* perinatal center; ownership status is measured as the respective *proportion* of perinatal centers in each county. In models (2) and (4), the sample is restricted to counties with perinatal centers of the same ownership status. Hence ownership is measured as a binary variable.

# B Supplementary Figures

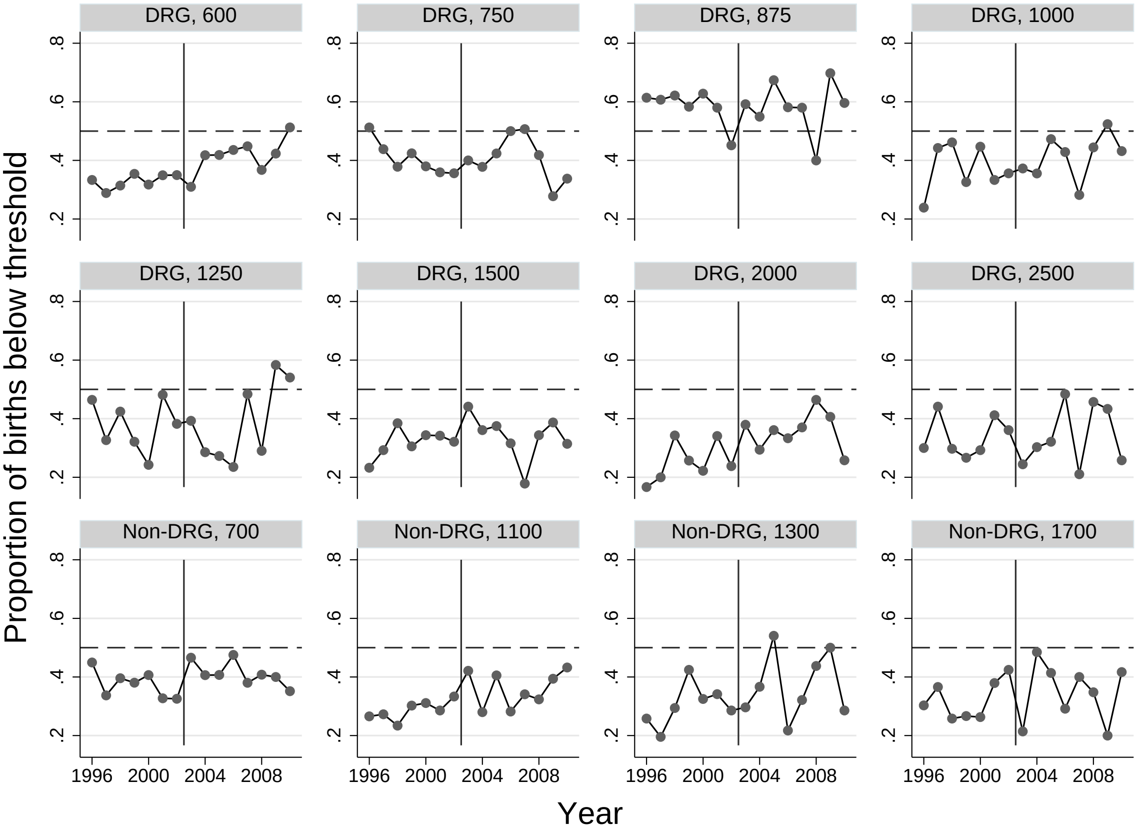

**Figure B.1:** Proportion of *still* births with weight recorded below threshold ($d = 25$) at eight DRG-relevant thresholds and four non-relevant thresholds, 1996 to 2010. The vertical line indicates the introduction of DRGs. The horizontal line indicates the "no upcoding" value of 50%. Source: Own calculations based on German birth records.

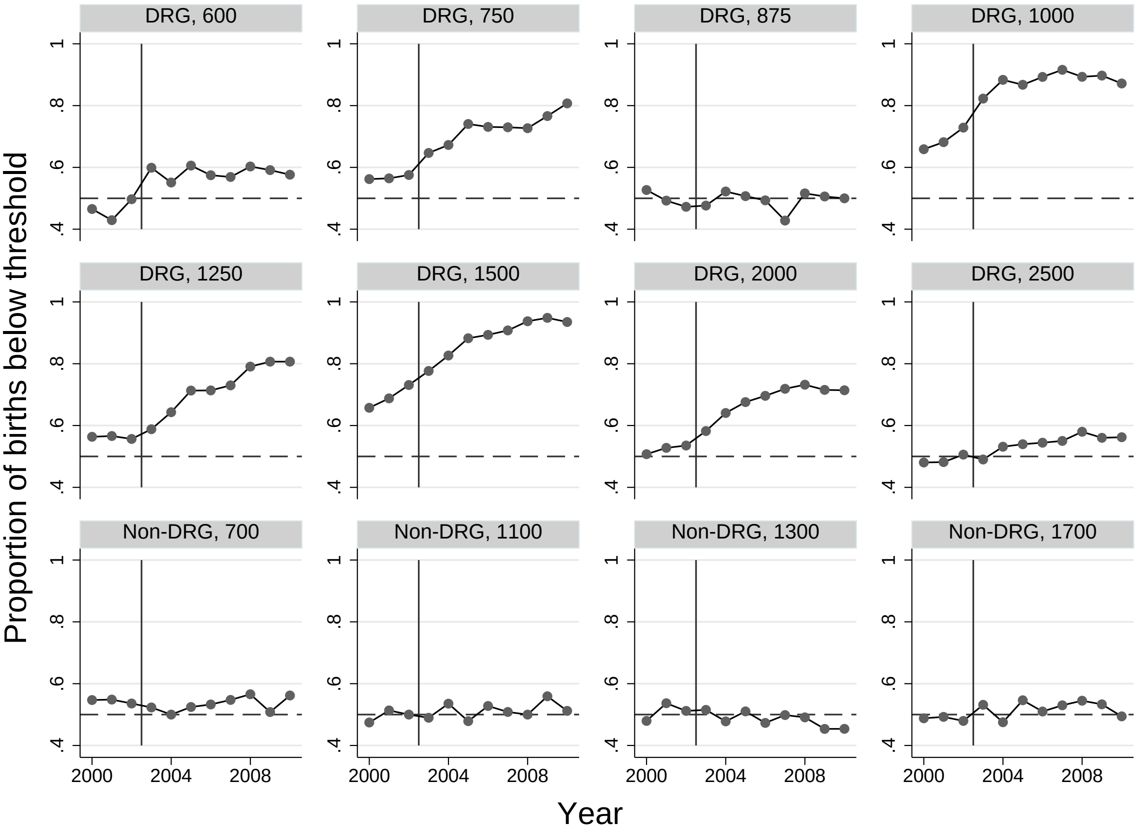

**Figure B.2:** Proportion of *live* births with weight recorded below threshold ($d$ = 25) at eight DRG-relevant thresholds and four non-relevant thresholds, 2000 to 2010. Birth weights exactly equal to threshold values are excluded. Source: Own calculations based on German birth records.

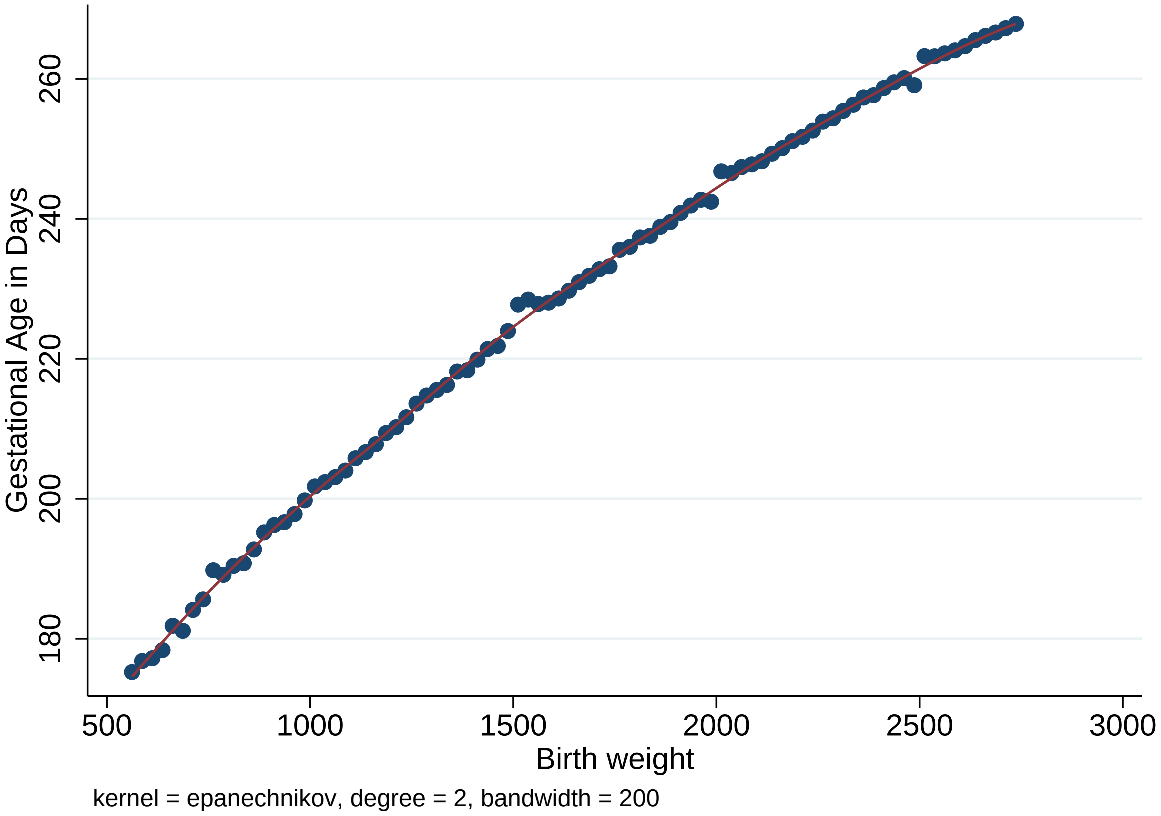

**Figure B.3:** Gestational age in days, by 25g birth weight bracket, 2006 to 2011. Source: Own calculations based on hospital quality data

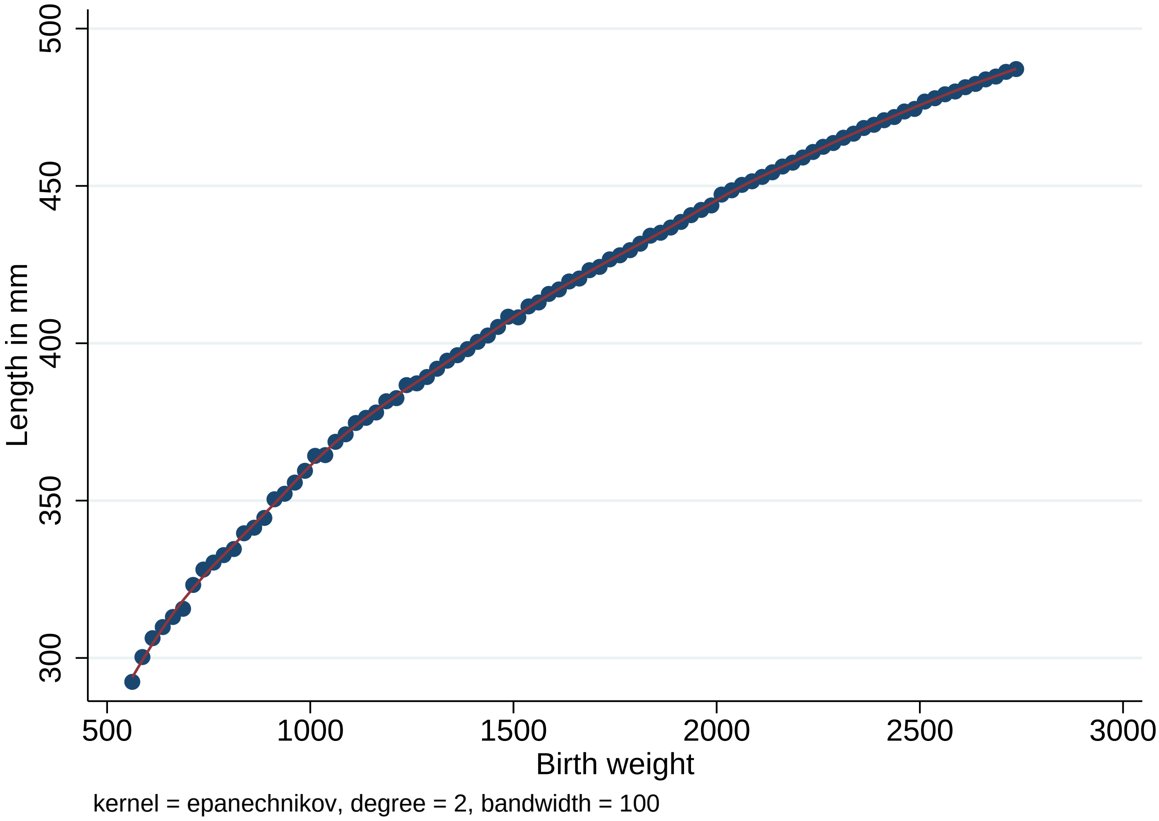

**Figure B.4:** Birth length in mm, by 25g birth weight bracket, 2003 to 2010. Source: Own calculations based on German birth records.

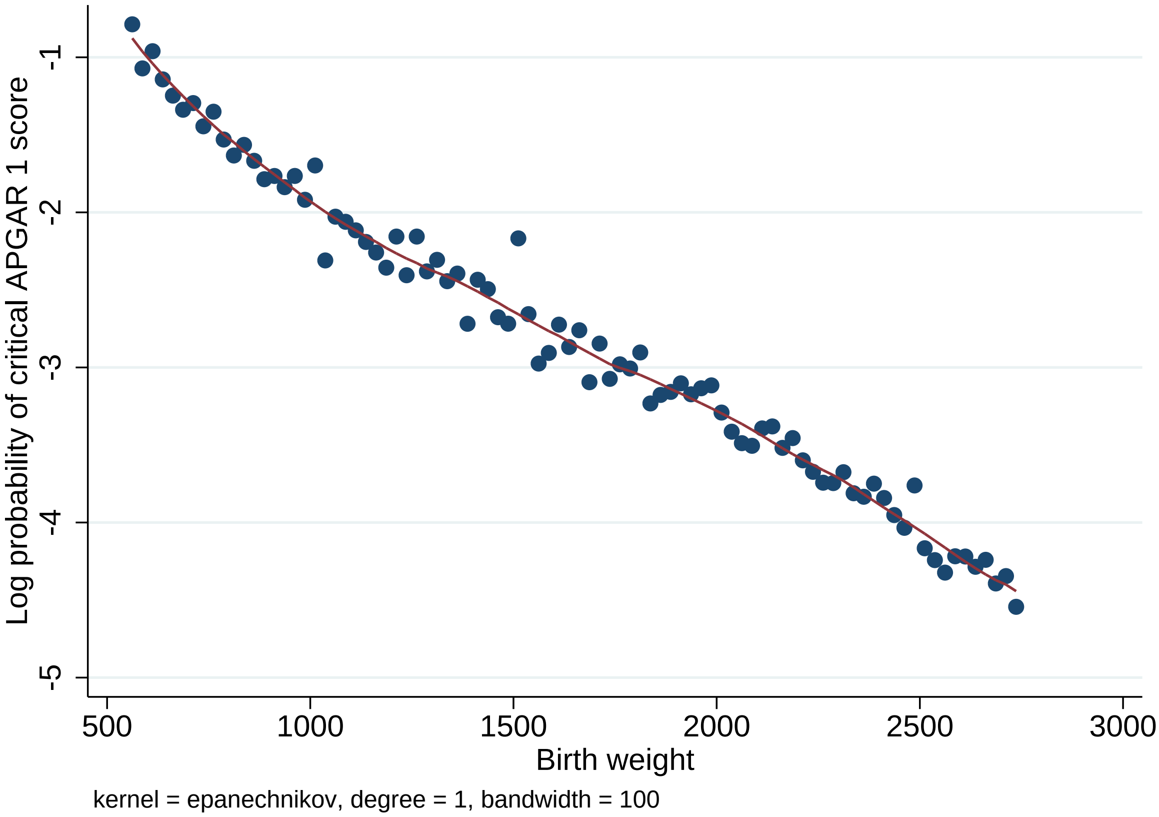

**Figure B.5:** Log proportion of infants born with critically low APGAR 1 score, by 25g birth weight bracket, 2006 to 2011. Source: Own calculations based on hospital quality data.

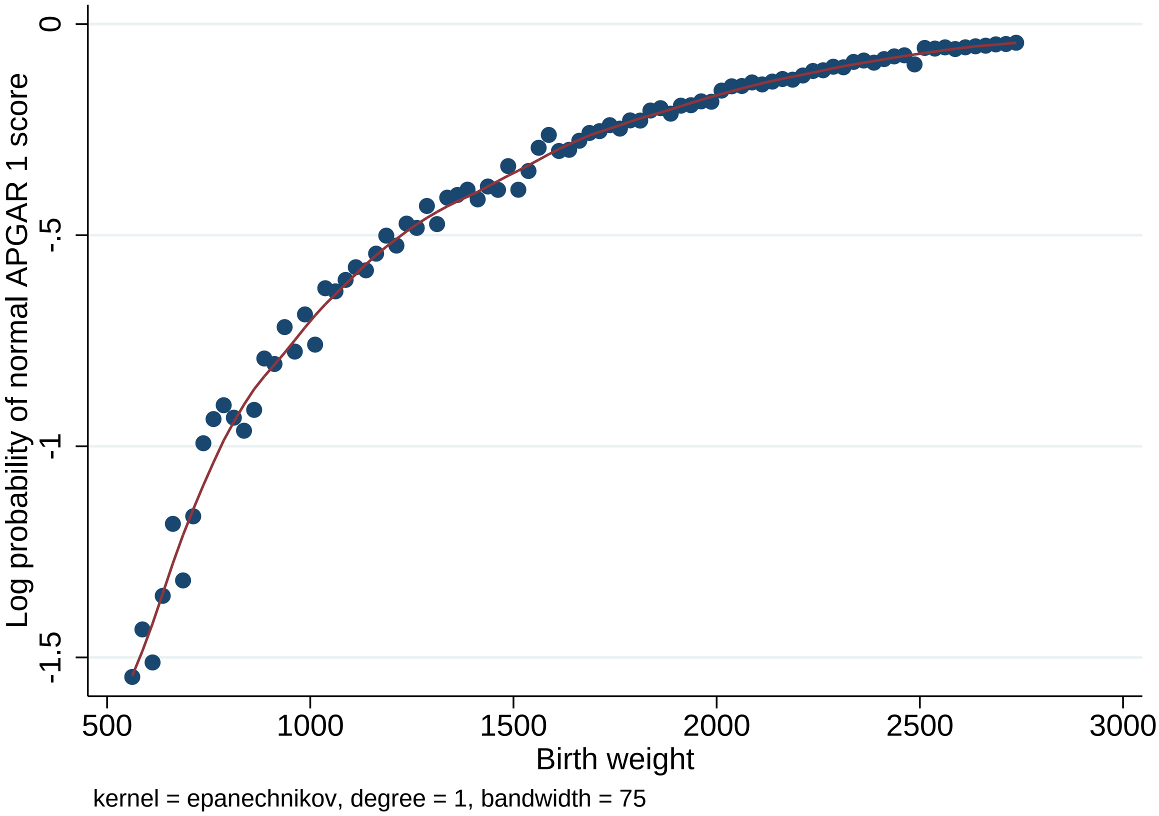

**Figure B.6:** Log proportion of infants born with normal APGAR 1 score, by 25g birth weight bracket, 2006 to 2011. Source: Own calculations based on hospital quality data.