

Jürges, Hendrik; Köberlein, Juliane

Conference Paper

DRG upcoding in German neonatology

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2013: Wettbewerbspolitik und Regulierung in einer globalen Wirtschaftsordnung - Session: Health Economics, No. G07-V3

Provided in Cooperation with:

Verein für Socialpolitik / German Economic Association

Suggested Citation: Jürges, Hendrik; Köberlein, Juliane (2013) : DRG upcoding in German neonatology, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2013: Wettbewerbspolitik und Regulierung in einer globalen Wirtschaftsordnung - Session: Health Economics, No. G07-V3, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel und Hamburg

This Version is available at:

<http://hdl.handle.net/10419/79866>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

DRG upcoding in German neonatology

Hendrik Jürges & Juliane Köberlein *

February 22, 2013

Preliminary and incomplete:
please do not cite without permission from the authors

Abstract

Since 2003/2004, German hospitals are reimbursed based on a prospective payment scheme (diagnosis related groups, DRGs). Patient classification in neonatology is based inter alia on birth weight, with substantial discontinuities in reimbursement at the relevant thresholds. These discontinuities create strong incentives to upcode especially newborns with very low or extremely low birth weight into classes of even lower birth weight. In this paper, we use data from the German birth statistics 1996 to 2010 to document the extent to which German hospitals have upcoded newborns by systematically documenting birth weights below the DRG thresholds. We estimate that between 2003 and 2010, hospitals have gained additional reimbursement in excess of 100 million Euro from the manipulation of birth weights alone. Further, we show that the likelihood of birth weight manipulation may be linked with the strength of financial incentives, regional competition in the hospital sector, ownership status, and newborn health. A corollary of our study is that, for cohorts born after 2002, regression discontinuity designs for effects of neonatal care on child health based on birth weight classifications are invalid.

JEL classification: I11, I18, D20

Keywords: Neonatal care, DRG upcoding

*We are grateful to participants at the conference on “Empirical Health Economics” at Ifo/University of Munich, and seminar participants at the University of Mannheim for helpful comments and suggestions. Authors’ contact information: juerges@uni-wuppertal.de, koerberlein@wiwi.uni-wuppertal.de; Schumpeter School of Business and Economics, University of Wuppertal; Rainer-Gruenter-Str. 21 (FN.01); 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 a more efficient allocation of resources in health care. 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, necessary services (diagnostics and therapies) may not be delivered 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 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 when grouping that are not documented in the medical charts. The semi-legal and illegal types can in principle be detected in audits.

Yet another type of upcoding is false documentation, i.e. manipulation patient charts so that patients appear to be sicker than they really are. This is done at the doctor or nurse level and can hardly be detected by audits. The aim of the present study is to estimate the extent and the determinants of this type of DRG upcoding in German neonatal care. As we will discuss in detail below, birth weight is one of the 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 15,000 Euro. Thus the financial incentives to manipulate documented birth weight by a few grams are strong. At the same time, false documentation of birth weights is virtually non-verifiable ex post, simply because of the initial weight loss of 5% of birth weight in the first few days after birth.

In the following, we document widespread upcoding in German neonatology – in the form of shifting birth weights from just above DRG-relevant threshold 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 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 10m German births 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.

Third, an important part of our paper is to analyze the 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 (Perinatalzentren) (Gerber et al., 2008). Thus competition among providers may be strong and the average number of patients per unit comparatively small. Considering the high fixed costs of setting up a neonatal care unit, some hospitals may suffer from substantial underfunding in neonatal care (Hoehn et al., 2008; Müller et al., 2007) that is compensated by coding newborns into higher paying categories. In the remainder of this section, we give a more detailed description of the DRGs that apply to German neonatology and their incentives for upcoding. We also describe the structure of the market for neonatal care in Germany.

1.1 DRGs in German neonatology

The DRG system in German neonatology is generally based on the following case characteristics: birth weight (or admission weight), OR-Procedures, long-term artificial respiration (>120h), (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. These thresholds have become very important because 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 up to 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

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

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 (i.e. the integral under the treatment cost curve is equal to the integral under the reimbursement curve). 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.

In the German DRG system, the actual reimbursement per case is obtained by multiplying a *relative cost weight* of that case according to its DRG 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 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, 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 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 in each individual case. Other DRG-systems are based on alternative classification criteria. For instance, the Japanese partial DRG system does reimburse care 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 (2011) provide evidence for such manipulation at the relevant thresholds of 1,000g and 1,500g.

The Nordic patient classification system (NordDRG) or the English Healthcare Resource Groups (HRG) are based on gestational age. Clearly, the day of conception is calculated and documented long before the birth and cannot be manipulated. But this criterion is also prob-

¹According to §21 Hospital Remuneration Act (KHEntG), all German hospitals are obliged to provide annual hospital-related structural (e.g. ownership, number of beds, labour 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).

lematic. If substantial reimbursement differences are related to merely one day of pregnancy, this creates incentives to hospitals and doctors not to arrest labor, which – in contrast to recording the wrong birth weight – may even be harmful to the newborn’s health. While an incentive not to prolong a pregnancy is generally also present if birth weight is used a criterion for reimbursement, it should be very hard to target delivery so that birth weight thresholds are not crossed.

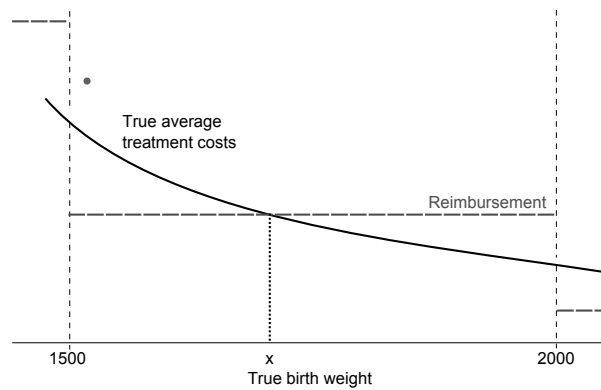


Figure 1: Relationship between average treatment costs and reimbursement (stylized graph)

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 aspiration from 120 to 121 hours can result in fee increase of about 20,000 Euro. Whether this additional hour was medically necessary cannot be verified ex post.

1.2 The market for neonatal intensive care

We now give a description of the market for neonatal intensive care in Germany. We are currently collecting ownership and structural data from all German perinatal centers and the section will be expanded.

Generally, neonatal care hospitals are assigned one of four different levels. Levels 1 to 3 each have a specific target population of mothers at risk of giving birth prematurely. Table 2 shows the characteristics of newborns and the level of perinatal care centers allowed to treat those children. Level 4 denotes regular maternity clinics. The assignment of a hospital to a certain level is based on the following structural criteria: professional standards for medical and nursing care, hours of shift work and standby duty, minimum number of normal, respiration and intensive care beds, and cooperation with pediatric hospitals and specialists (e.g. children’s cardiology). Assignments are periodically reviewed. 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 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 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 (36 of 409). 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 annual cases 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 abolished by the Federal Joint Committee (G-BA). 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 artificial ventilation facilities are assigned to Level 3. They represent about 15% of all perinatal hospitals. 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.

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: According to the agreement of the Joint Federal Committee, 2005

²Usually, hospital supply is determined in hospital plans that are drawn up at 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. Information on the number of perinatal care centers was supplemented by information from the German hospital search website www.deutsches-krankenhaus-verzeichnis.de.

2 Data

The birth weight data used in this study are derived from official German birth statistics 1996 to 2010, covering both the pre-DRG- and the DRG-period in German hospitals. The legal basis for data collection are §2 of the Population Statistics Law (Bevölkerungstatistikgesetz), which determines which kind of variables are collected and §20 of the Civil Status Law (Personenstandsgesetz), which stipulates that all hospitals are obliged to report the birth weight and length to the local civil registry office.

Our data include about 10 million births, of which some 688,000 or roughly 7% were of low birth weight (<2,500g). Table 3 shows the number of live births, by birth weight category and period (before/after introduction of DRGs). Note that data from Bavaria in 1996 to 1999 includes birth weights only in brackets of 100 grams. These were generally excluded in our analyses. Table 3 indicates a general trend towards live births with low birth weight. The proportion of babies born with extremely low, very low, and low birth weight has increased from 6.3% of all births in 1996-1999 to 6.8% of all births in 2003-2010, that is a relative increase of about 8%. The relative increase is even more pronounced with respect to very low and extremely low birth weights. The proportion of births with very low birth weight (below 1,500 grams) has increased from 1.05% to 1.20% (relative increase 14%) and the proportion of births with extremely low birth weight (below 1,000 grams) has increased from 0.41% to 0.50% (relative increase 23%). 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 babies to survive.

Table 3: Number of live births, by birth weight category and period

Category	Weight in grams	1996-1999		2000-2002		2003-2010	
		N	Cumulative Percent	N	Cumulative Percent	N	Cumulative Percent
Extremely low	<1,000	10,891	0.41	9,840	0.44	27,649	0.50
Very low	1,000-1,499	17,005	1.05	14,768	1.11	38,269	1.20
Low	1,500-2,499	139,705	6.31	121,035	6.56	309,158	6.84
Normal	2,500+	2,486,423	100.00	2,075,081	100.00	5,106,233	100.00
Total		2,654,024		2,220,724		5,481,309	

Note: 1996-1999 does not include births in Bavaria.

3 Upcoding and excess reimbursement over time

3.1 Changes in the birth weight distribution over time

In this subsection, we describe 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, the claim we make here is that with 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. Where and when such thresholds were irrelevant for reimbursement, there is no change in the distribution of birth weights after the introductions of DRGs.

The German birth statistics contain the recorded birth weights to the exact gram. For the purpose of our analyses, we usually recode them into brackets of 10g, 25g, or 50g (depending on the analysis) and consider only brackets just below and above the eight DRG-relevant birth weight thresholds. The threshold value itself is included in the bracket above the threshold. Let x denote some threshold and d the width of the bracket, then a rough measure of upcoding around at 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 :

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

where $2R_d - 1$ approximates the proportion of cases that were upcoded from $[x, x + d - 1]$ to $[x - d, x + d - 1]$, but note that this measure is likely to be biased. 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 larger right than left of the threshold, whereas R_d 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 over time shown in this section as long as the shape of the true distribution does not change over time. But it will of course matter when we provide an estimate of the absolute number of upcoded cases in each year. We will therefore present an alternative measure.

Figure 2 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 of 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 a birth weights are very often rounded, for instance to multiples of 100. Such rounding occurs when the scales are imprecise or when the midwives recording the birth weights are just being sloppy. 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 could 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 after the introduction of DRGs, 80% to 90% of recorded birth weights are below the threshold, which translates into a lower bound for the upcoding rates of 60% to 80%. For instance, in 2008, a staggering

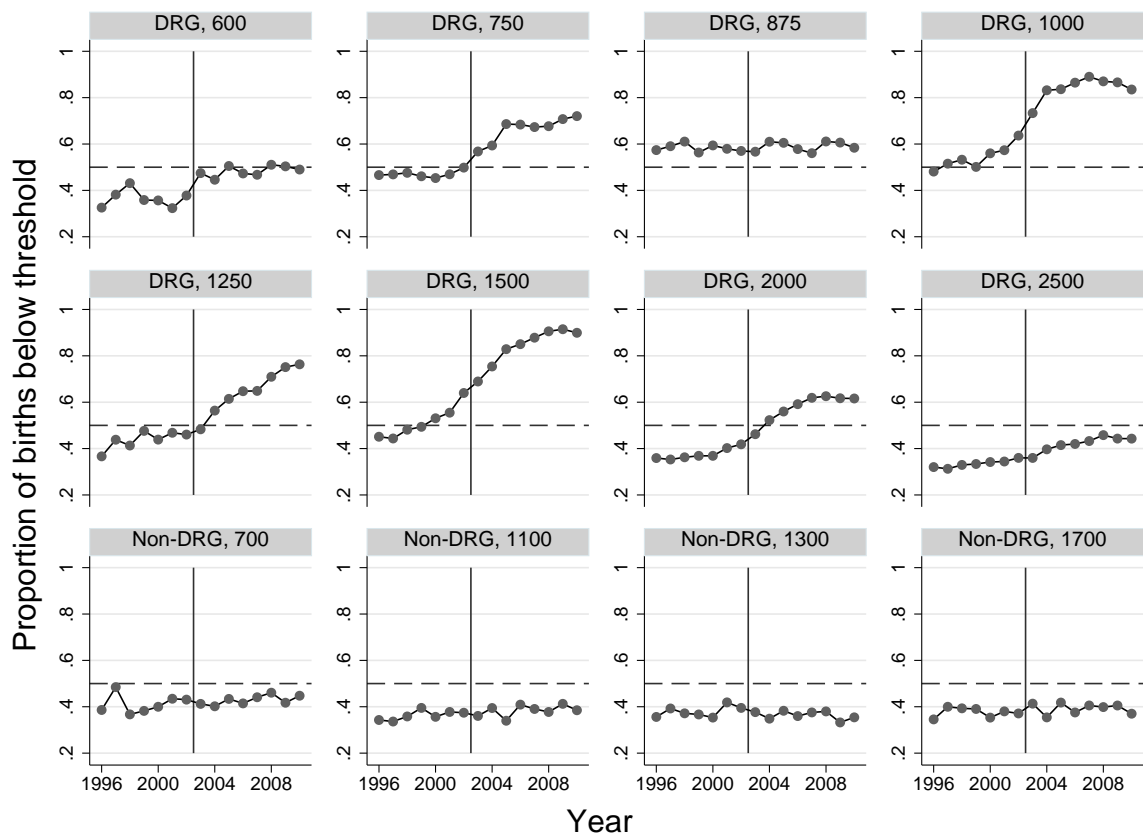


Figure 2: Proportion of *live* births with weight recorded below threshold ($d = 25$) at eight DRG-relevant thresholds and four non-relevant thresholds, 1996 to 2010. The vertical line indicates the introduction of DRGs. The horizontal line indicates the “no upcoding” value of 50%.

88.5% percent of all births in the 1,475 to 1,524 grams bracket were recorded below 1,500 grams. Possibly, 1,000g and 1,500g can be easily remembered as relevant threshold because they are also used to categorize birth weight into the familiar “extremely low” and “very low” categories. Also note that the trend towards recording more birth weights below the threshold already started one or two years before the introduction of DRGs but the introduction of DRG seems to have strengthened this trend. The 1,250g threshold has become salient as well as the threshold value for computing minimum quantities is neonatal intensive care.

In contrast to all other thresholds, we find stable R_{25} at the 875 grams threshold. Possibly 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 bracket includes a multiple of 50 (850) whereas the upper bracket does not.

Now consider the third row in Figure 2. It shows the percentage of births below 700, 1100, 1300, and 1700 grams, respectively, in the 50g brackets around these thresholds. These thresholds are irrelevant for reimbursement, hence no economic incentives apply for coding weights below these thresholds. 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. Still births provide another possibility for comparison, because hospital reimbursement is not affected by the weight of a stillborn child. Figure 3 displays the same information as in Figure 2, but now for still births. Due to the small number of stillbirths, there is a lot more variation in the data. However, the general picture is quite clear. There appears to be a slight upward trend in R_{25} , e.g. at 2,000g or 1,100g, but there is no visible jump after 2003, so that the secular trend might be explained by increasingly precise measurements. For most threshold values, the development seems rather flat. Note also the large number of values below the threshold at 875g which supports the finding for live births at that threshold.

3.2 A brief note on the role of rounding

As mentioned above, the German birth statistics contain the recorded birth weights to the exact gram. However, due to lack of precision of scales used in hospitals and the extent to which midwives round 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

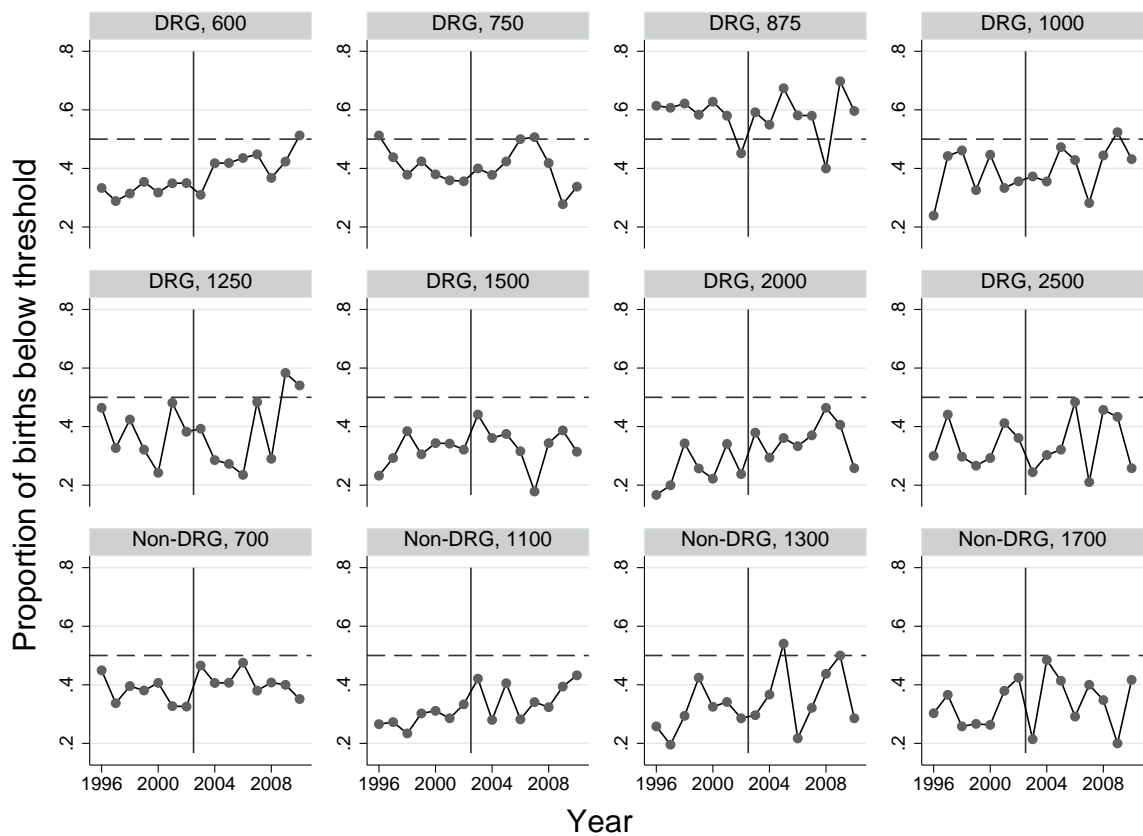


Figure 3: Proportion of *still* births with weight recorded below threshold ($d = 25$) at eight DRG-relevant thresholds and four non-relevant thresholds, 1996 to 2010. The vertical line indicates the introduction of DRGs. The horizontal line indicates the “no upcoding” value of 50%.

31 to 23 percent, and the proportion rounded to multiples of 10 fell from 97 to 89 percent.

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 the 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 analysis excluding all birth weights exactly at the threshold values to see how much this changes the general increase in upcoded values after the introduction of DRGs. Mainly, this 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 will be 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, namely at 1,000g and 1,500g, where already in 2000, 66% of all non-rounded birth weights were below the threshold. This means that already before 2000, there was some incentive to code birth weights below these thresholds. Unfortunately, as of now, we were not able to identify the source of this incentive.

3.3 Excess reimbursement due to upcoding 2003-2010

We now provide an estimate of the absolute number of cases that have been upcoded at each relevant threshold between 2003 and 2010. By multiplying the estimated number of upcoded cases in year y at threshold k , \hat{U}_{yk} , with the average additional reimbursement obtained by the hospital for that upcoded case, $\bar{\Delta}_{yk}$, and aggregating across all years and threshold, we obtain an estimate of the total amount A that has been received unfairly by German hospitals, and perinatal centers in particular:

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

In the following, we describe how we computed \hat{U}_{yk} and $\bar{\Delta}_{yk}$.

3.3.1 Estimating the number of upcoded cases

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

$$\hat{U}_{yk}^d = N_{yk}^d - \hat{N}_{yk}^d \quad (3)$$

where d indicates a bracket of width d below each threshold, N_{yk}^d is the *observed* number of births within that bracket and \hat{N}_{yk}^d is the *predicted* number of births in the same bracket that would have been recorded if there was no DRG-based reimbursement. To compute the latter prediction, we estimate the slope of the distribution function around each threshold in a way

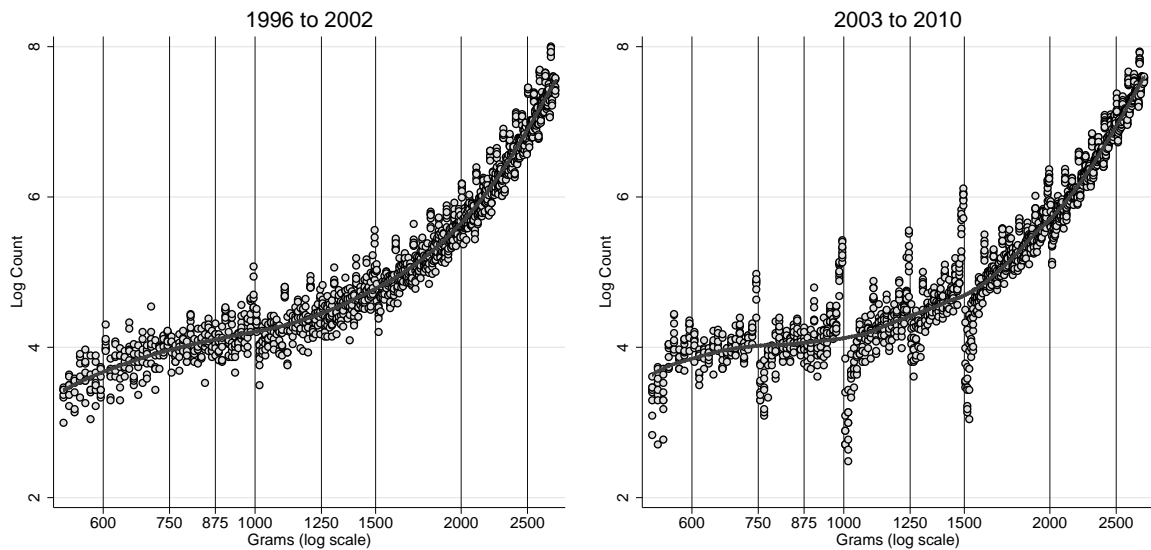


Figure 4: Distribution of birth weights before and after the introduction of DRGs. The smooth lines show the results of a local polynomial regression (degree 2, $bw=200$, Epanechnikov kernel). The light dots show the observed number of births in 10g brackets in each year.

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 can be seen 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. The smooth dark line shows the estimated distribution of birth weights between 500g and 2750g. The 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 weights since the introduction of DRGs, reveals a conspicuous excess number of births just below most 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}_{yk}^d for $d = 10, 20, \dots, 50$. The results are shown in Table 4. 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 up to 50g below the threshold. The two latter findings are probably due to some rounding of birth weights to multiples of 50g. For our following estimates, we use a bracket width of 30 grams. Thus overall, we find that between 2003 and 2010, 12,133 births have been upcoded to a lower birth

weight DRG.

Table 4: 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}_{yk}^d)	6,970	11,758	12,133	11,971	15,086

Table 5 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 newborns 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 1,000, 1,500, and 2,000 grams. These are salient numbers that can be easily memorized. Moreover, as noted before, 1,000 grams is the threshold for “extremely low birth weight” and 1500 grams is the threshold for “very low birth weight”.

Table 5: Estimated number of upcoded cases, by year and threshold

Threshold	Year								Total
	2003	2004	2005	2006	2007	2008	2009	2010	
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

3.3.2 Estimating the excess reimbursement due to upcoding

Based on the DRG reimbursement catalogs for each respective year, we first calculated 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 calculate reimbursement fees per birth weight category, we then multiplied these average cost weights by the national average base rate. The national average was 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 obtained by hospitals for a specific upcoded case, $\bar{\Delta}_{yk}$, were obtained by subtracting the reimbursement fee of cases below a threshold value from cases above.

³Note that we obtain a substantial *negative* estimate for the 2,500g threshold in 2003. This happens for instance if many 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 unintentional downcoding should be subtracted from the amount received due to upcoding.

Table 6: Estimated excess reimbursement in 1,000 Euro, by year and threshold

Threshold	Year								Total
	2003	2004	2005	2006	2007	2008	2009	2010	
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 6 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

In this section, we 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 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 the smaller 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 7. 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 rises

⁴Prospective data for 2003 are not available.

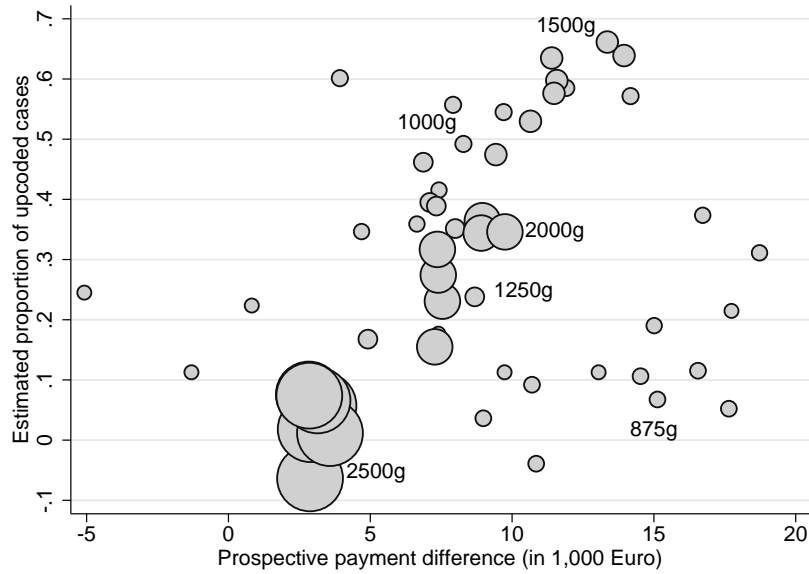


Figure 5: Proportion of upcoded cases around each threshold plotted against the expected payment difference in 1,000 Euro. The bubble size indicates the number of observations around each threshold.

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 (in column (3)) and thresholds 1,000g and above (in 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 7: Relationship between the proportion of upcoded cases and expected payment difference

	(1) all thresholds	(2) all thresholds	(3) below 1000g	(4) 1000g and above
Payment difference	0.010* (0.005)	0.034*** (0.006)	-0.003 (0.003)	0.054*** (0.006)
Constant	0.210*** (0.047)	-0.031 (0.035)	0.221*** (0.042)	-0.067 (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

We now turn to an analysis of the relationship between hospital specialization and the proportion of babies born with a birth weight below a DRG-threshold. 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. Thus in contrast to hospitals that actually care for low and very low birth weight infants, the non-specialized hospitals have no strong financial incentive to manipulate birth weights of babies born just above DRG-thresholds.

We are not able to link individual births to hospitals, but we are able to link individual births to the city or county in which the hospital is located. Thus we turn to a county-level analysis – which still allows us to draw some conclusions about the link between specialization and upcoding. The reason is that not all counties have Level 1 or Level 2 perinatal centers with neonatal ICUs. In fact, there are about 220 Level 1 or Level 2 hospitals 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 has no neonatal intensive care unit and mothers at risk of delivering prematurely who reside on those counties are usually treated in another county.

We expect the proportion of upcoding to be lower than average in counties with few low birth weight infants, just because hospitals have only little incentive to do so. In contrast, in counties with a high volume of underweight babies and one or more neonatal ICUs, we expect higher than average upcoding rates.

With more than 400 counties in Germany, the disaggregated number of births with low and very low weight becomes very small in less densely populated regions without perinatal care centers. Therefore, while we disaggregate our data by county, we aggregate them over time and also compute a somewhat simplified upcoding measure. In the following, the county specific extent of upcoding is captured by the number of births 50g below all relevant thresholds divided by the number of births in the interval 50g below to 49g around the relevant thresholds.

The left panel of Figure 6 shows a scatterplot of the proportion of newborns coded below DRG-thresholds against the total number of infants born with low birth weight (<2,500g). Each dot represents one county. The three largest counties are easily identifiable as Berlin, Munich, and Hamburg. The solid horizontal line at 50% represents a crude reference for “no upcoding”. Dots located above the line indicate a larger number of babies with birth weight recorded below than above DRG-relevant thresholds.

Notably, Figure 6 shows a positive association between the size of a county in terms of the number of underweight births and the proportion of upcoding. This association appears to be driven by two groups of counties. First, there are “small counties”, for which we set an arbitrary limit of up to 600 births in eight years. This group of counties exhibits a distribution of the percentage of newborns below DRG-threshold around an average of slightly more than 50%, seemingly indicating random variation around the “no upcoding” average (also see the

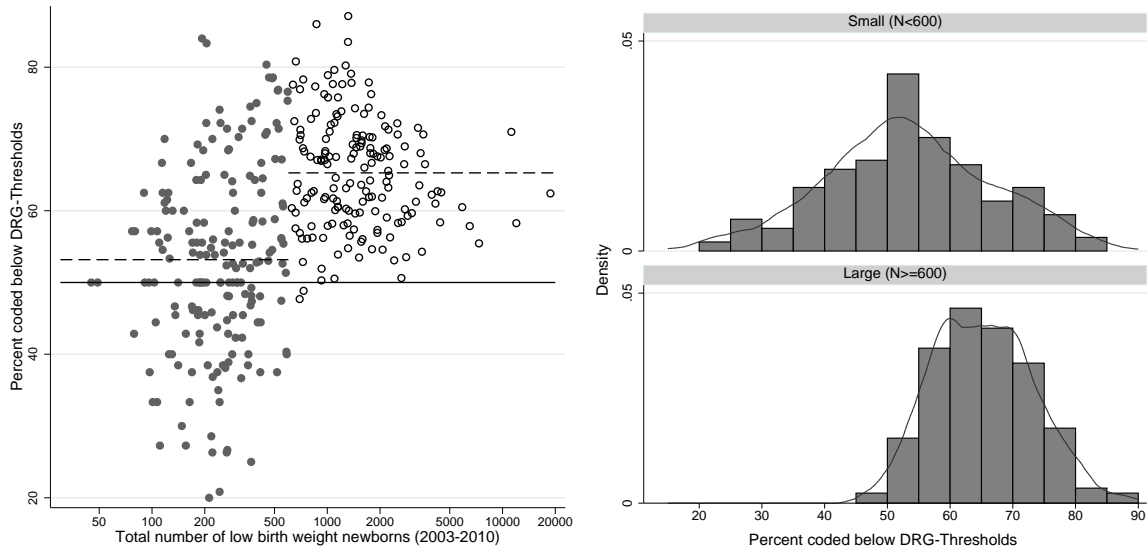


Figure 6: Proportion of newborns coded below thresholds by total number of babies born with low weight, by county

top part of the right panel). The majority of these counties has no perinatal centers, hence very low birth weights are rare accidents. These infants are very likely to be transferred to specialized perinatal centers, so that hospitals in those counties have no incentive to upcode babies. Hospital midwives might also have no experience with the relevant DRGs and their substantial reimbursement differences at certain threshold.

Second, there are “large counties” (above the arbitrarily chosen limit of 600 births in eight years). Almost all counties can be found above the 50% line and the unweighted average percentage of babies with birth weight recorded below the DRG-thresholds is 65. In these counties, we are more likely to find perinatal centers that have an incentive to upcode and where midwives should also have the relevant knowledge when this is profitable.

4.3 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 care has recently been studied by Freedman (2012), who convincingly 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 largely 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 raising the reimbursement for treating a given newborn. We therefore believe it is plausible that the regional supply of NICU beds will be positively related to the extent of upcoding. The more beds per newborn exist in a region, the larger the proportion of upcoded cases.

We are currently collecting information on the number of neonatal ICU beds, but in the following we test our hypothesis by using as proxy the total number of beds in obstetrics, gynecology and neonatal care per 1,000 women aged 15 to 45 in each county (taken from regional statistics). Further, we use German counties (Kreise) as the relevant regional market for newborn care. We will clearly improve upon this by defining regional markets as counties with perinatal centers plus neighboring counties, and by using the number of NICU beds when they become available.

Table 8 shows the results of a series of regressions of the number of proportion of cases below DRG-thresholds (the county-level upcoding measure used in the preceding section) on the number of beds in ob/gyn and neonatology per 1,000 women aged 15-45 and control variables. Data are from 2004 to 2010. Columns (1) and (2) show results of within county-regressions without year dummies and with year dummies. In the model without year dummies, the relationship between upcoding and supply is found to be positive and statistically significant. Each time the supply of beds per 1,000 women increases by one, the upcoding rate increases by 1.1 percentage points. Considering the fact that the average number of beds per 1,000 women is 2.4, this seems to be a modest effect in substantive terms. Also note that we are controlling for number of births <1,500g and its square to control for centrality – the fact that larger counties have both a larger number of hospital beds and a larger number of perinatal centers with a higher propensity to upcode. We also control for the total number of hospitals in a county but it has no significant effect on the proportion of upcoded cases.

Controlling for a general time trend diminishes the coefficient of number of beds to an insignificant 0.4 percentage points. This coefficient must be interpreted as showing the effect of a county-specific deviation of the trend in the number of beds from the a general trend, which obviously does take up a lot of the within-county variation in the number of beds.

Columns (3) and (4) show the results of between-county regressions (basically regressing county-averages on county-averages). Here there are no major differences between the models with and without time dummies. The coefficient of the number of beds in ob/gyn and neonatology shows a statistically significant increase of 1.4 to 1.5 percentage points in upcoding with

each additional bed per 1,000 women.

Table 8: County-level panel regressions predicting percentage of birth weights coded below threshold

VARIABLES	Fixed Effects		Between Effects	
	(1)	(2)	(3)	(4)
N of Ob/Gyn/Neo beds per 1,000 women	0.011*** (0.002)	0.004 (0.003)	0.015*** (0.003)	0.014*** (0.003)
N of births (in 1,000s) < 1,500g	0.242*** (0.065)	0.256*** (0.065)	0.163*** (0.024)	0.162*** (0.023)
N of births (in 1,000s) < 1,500g squared	-0.055* (0.032)	-0.066** (0.032)	-0.071*** (0.013)	-0.069*** (0.012)
N of hospitals	-0.002 (0.002)	-0.002 (0.002)	0.002 (0.001)	0.001 (0.001)
Year dummies	no	yes	no	yes
Constant	0.464*** (0.012)	0.450*** (0.013)	0.452*** (0.010)	-0.009 (0.067)
Observations	3,099	3,099	3,099	3,099
Number of counties	397	397	397	397

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

4.4 Hospital ownership and upcoding

While casual reasoning may predict that for-profit hospitals should engage more in upcoding practices, the relationship between hospital ownership and upcoding is theoretically ambiguous. Despite the 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 owner's and their own income. Managers in not-for-profit hospitals might support upcoding to increase revenue that can be distributed according to the owner's or their own preferences (Duggan, 2002). Further, independent of ownership, midwives have an incentive to upcode – also without intervention by the management – to increase the revenue of their unit and to secure their own future employment and income.

Our data do not allow us to link cases to individual hospitals. We only know the city or county of the hospital of birth, so that we will use county data on the number and ownership of perinatal centers to analyze relationship between ownership rates and upcoding rates. We currently only have data on hospital ownership by county from Northrhine-Westphalia, Lower Saxony, and Bavaria, thus our analysis is very preliminary and will be updated in future versions of the paper. In 2009, there were exactly 1,000 hospitals in 195 counties in those

three federal states, of which 331 were public, 418 were not-for-profit and 251 were private. 90 counties had no private hospital.

Estimates of differences in upcoding rates by ownership status were obtained by *random effects* regressions of county-level annual upcoding rates on the *proportion of hospitals* of each ownership type (see Table 9). We refrained from estimating county-fixed effects models because there is hardly any variation in ownership rates over time, leaving very little variation to estimate the ownership effect in such a model. We estimated three models overall, first without control variables, second controlling for the number of beds in ob/gyn and neonatology, the number of births with weights below 1,500g, and the number of hospitals, and third additionally controlling for a general time trend.

The model without controls predicts that in a county with public hospitals only, the upcoding rate would be 47.3% (the constant). In a county with not-for-profit hospitals only it would be 9.5 percentage points higher, and in a county with for-profit hospitals only, it would be 4.2 percentage points higher. Both differences are statistically significant. Accounting for county differences in the number of beds, low birth weight infants, and hospitals – which have partly been shown to be important in the preceding section – does reduce the predicted differences in upcoding rates to 6.2 percentage points and 5.4 percentage points, respectively. Additionally controlling for a time trend reduces the predicted differences by 0.3 to 0.4 percentage points. Although these findings must be interpreted with some caution, they provide a first indication that public hospitals generally have lower upcoding rates than both not-for-profit hospitals and for-profit hospitals.

4.5 Newborn health status and upcoding

Even conditional on DRG classification criteria, newborns have different health states. Worse health usually means higher costs for the hospital. It is not possible to select newborns based on their health status. Instead, hospital staff might selectively upcode newborns who are likely to cost more - based on observable characteristics that are not part of the DRG classification such as being small, having a small head circumference, having a low APGAR score, etc. Such behavior can be expected for instance if doctors or midwives think that upcoding is unethical but that it is worth doing if actual costs and reimbursement can be aligned.

Because the birth register data only include birth length, we now use an alternative data source, namely data collected for the cross-sectoral quality assurance in health care. They cover all births in German hospitals from 2006 to 2011, including birth weight and APGAR scores. APGAR is the most common measure of neonatal health. 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 as critically low.

Figure 7 shows the average 1 minute APGAR scores by birth weight (in 10g brackets) in

Table 9: Random effects estimates of hospital ownership effects on upcoding rates, county-level data, Northrhine-Westphalia, Lower Saxony and Bavaria, 2003 to 2010

	(1)	(2)	(3)
Proportion not-for-profit	0.094*** (0.020)	0.062*** (0.020)	0.058*** (0.018)
Proportion for-profit	0.042* (0.024)	0.054** (0.021)	0.051*** (0.019)
N of Ob/Gyn/Neo beds per 1,000 women		0.012*** (0.002)	0.011*** (0.003)
N of births (in 1,000s) < 1,500g		0.165*** (0.032)	0.170*** (0.030)
N of births (in 1,000s) < 1,500g squared		-0.059*** (0.015)	-0.063*** (0.013)
N of hospitals		-0.238 (1.831)	0.091 (0.678)
Time dummies	no	no	yes
Constant	0.473*** (0.012)	0.475*** (0.011)	0.492*** (0.014)
Observations	1,476	1,462	1,462
Counties	190	188	188

Standard errors in parentheses
 *** p<0.01, ** p<0.05, * p<0.1

the range between 550 and 1750 grams (left panel) and 1750 to 2750 grams (right panel). The distribution of birth weights is shown at the bottom of both panels. Average APGAR scores clearly rise with birth weight. If our hypothesis about doctor and midwife behavior is correct, we would expect that upcoded infants are in worse average health. In the Figure 7 this should show up as a discontinuity in APGAR scores across DRG-thresholds, with those on the left of the threshold having on average lower scores and those on the right having on average higher scores (because low APGAR scores are systematically shifted from right to left). However, what we see in the left panel is rather the opposite. Infants with birth weights just above the thresholds appear to have lower than expected APGAR scores and vice versa. In other words, at birth weights up to the 1,500g threshold, it is healthier infants who are systematically upcoded or put differently, unhealthy children with critical APGAR scores are not upcoded. But when we look at the 2,000g and 2,500g thresholds, we find out initial hypothesis confirmed. That is, infants with birth weights recorded just below the threshold had on average lower APGAR scores, or relatively unhealthy infants more often upcoded than health ones.

We have repeated the analysis shown here for 5 and 10 minute APGAR scores, with very similar results. Results for birth length and head circumference reveal that at the 2,000g and 2,500g thresholds, it is smaller infants and infants with smaller heads that are disproportionately often upcoded. No systematic patterns could be found at the lower DRG thresholds.

Our conclusion from this analysis is that upcoding is systematically related to non-DRG relevant health indicators of the newborn. The relationship is complex. At higher birth weights, unhealthy infants are systematically upcoded. Inasmuch 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 critical, unhealthy infants are systematically not upcoded. One explanation 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.

5 A digression on donuts

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.. In particular they exploit a discontinuity in treatment provision around the 1,500g threshold, measured by sudden increases in average medical charges and length of stay as one crosses the threshold from above. They also find a discontinuity in outcomes (in particular 28-day mortality) around the 1,500g threshold, with lighter babies below the threshold (and thus more intense treatment) having higher survival changes than heavier babies above the threshold. The regression discontinuity design used in this study has been criticized by Barreca et al. (2011) as invalid because of a great number of observations at ex-

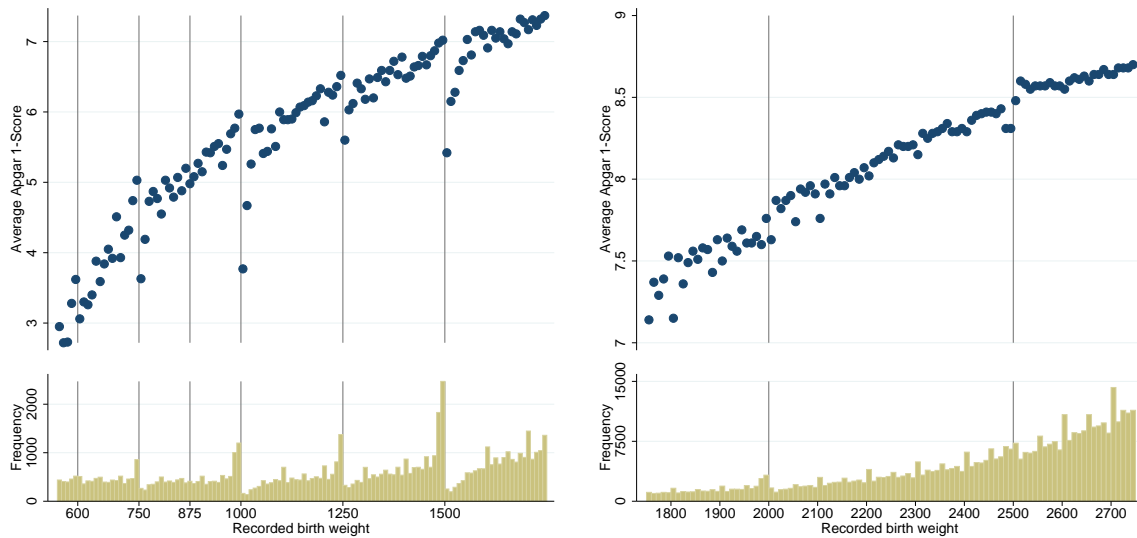


Figure 7: Average Apgar 1-Score by birth weight. Data source: Hospital Quality Assurance data, 2006-2011

actly 1,500g (due to rounding), which have higher than average mortality rates. Rounding thus artificially increases mortality rates above the 1,500g threshold. To overcome this problem, Barreca et al. (2011) propose to use a so-called “donut” regression discontinuity design, that is, to leave out the rounded values at the discontinuity. If this approach is applied to the data used in Almond et al. (2010), one obtains substantially smaller and statistically insignificant outcome differences across the threshold.

The relevance of our findings for any study in Germany that aims at exploiting regression discontinuity designs along the lines of Almond et al. or Barreca et al. is clear. Since the “treatment determining variable” is clearly manipulated – and probably even systematically manipulated along infant health, such a design is unfeasible. The donut hole needed to exclude manipulated cases would have to be unacceptably wide. Hence, regression discontinuity designs to estimate causal effects of neonatal intensive care in Germany are invalid.

6 Summary and conclusion

In this paper, we have shown that, since the introduction of DRGs in 2003, 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 slightly above the DRG thresholds have been recorded as having a birth weight slightly below the threshold. As a result, hospitals have received excess reimbursement of 114m Euro in that period.

We also show that birth weight manipulation is clearly 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 hospitals 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 can be explained the hospitals in the latter case having no financial incentive to manipulate birth weights. Third, we find a relationship between supply or regional competition, measured as the number of Ob/Gyn/Neonatology hospital beds per 1,000 women aged 15-45, and the extent of upcoding. Whereas filling beds makes no sense when hospitals are reimbursed prospectively, raising revenue per case by upcoding does. We also examined whether public, not-for-profit and for-profit hospitals differ in the extent of upcoding. We find significantly higher rates in counties with higher proportions of not-for-profit and for-profit hospitals, but since our data on ownership were limited to only three (albeit large) federal states in Germany, it is unclear whether it survives when we gather more data. Finally, we have also checked whether hospitals are more likely to upcode newborns that appear to be less healthy (and thus more costly) conditional on weight. We find that – for the DRG thresholds up to 1,500g – newborns with lower APGAR scores appear to be less often upcoded, which could be linked to the higher mortality rate and thus lower expected costs of and reimbursements for treating such infants. At the two largest DRG thresholds at 2,000g and 2,500g, newborns with lower APGAR appear to be more often upcoded. Here, mortality does not play a big role and our finding is compatible with the view that when doctors or midwives shift babies into lower birth weight categories, they do so more often if expected costs of treatment are higher.

Obviously, there must be 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. We conclude from our findings that the current German DRG-classification relying to a large extent on birth weight is inappropriate. In contrast to gestational age, which is used in alternative DRG systems such as the NordDRG, it is prone to easy manipulation. But gestational age creates incentive problems of its own – namely for doctors not to arrest labor although it would be medically justified. Thus, as an alternative to (arbitrary) birth weight categories or gestational age, one could also 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.

References

- Abler, Sylvie, Pablo Verde, Hans Stannigel, Ertan Mayatepek, and Thomas Hoehn (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, Douglas, Joseph J. Doyle, Amanda E. Kowalski, and Heidi Williams (2010) 'Estimating marginal returns to medical care: Evidence from at-risk newborns.' *The Quarterly Journal of Economics* 125(2), 591–634
- Barreca, Alan I., Melanie Guldi, Jason M. Lindo, and Glen R. Waddell (2011) 'Saving babies? Revisiting the effect of very low birth weight classification.' *The Quarterly Journal of Economics*. DOI: 10.1093/qje/qjr042
- Duggan, Mark (2002) 'Hospital market structure and the behavior of not-for-profit hospitals.' *RAND Journal of Economics* 33(3), 433–446
- Freedman, Seth (2012) 'Capacity and utilization in health care: The effect of empty beds on neonatal intensive care admission.' Unpublished working paper, University of Michigan.
- Fuchs, Victor (1978) 'The supply of surgeons and the demand for operations.' *Journal of Human Resources* 13(Suppl.), 35–56
- Geissler, Alexander, David Scheller-Kreinsen, Wilm Quentin, and Reinhard Busse (2011) 'Germany: Understanding g-drgs.' In *Diagnosis-Related Groups in Europe*, ed. Reinhard Busse, Alexander Geissler, Wilm Quentin, and Miriam Wiley (World Health Organization) pp. 243–271
- Gerber, Andreas, Karl Lauterbach, and Markus Lungen (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
- Hoehn, Thomas, Attyla Drabik, Christian Lehmann, Hans Stannigel, and Ertan Mayatepek (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* Institut für das Entgeltsystem im Krankenhaus gGmbH, version 3.0 ed.
- (2011) *Abschlussbericht zur Weiterentwicklung des DRG-Systems für das Jahr 2012* Institut für das Entgeltsystem im Krankenhaus gGmbH
- Müller, C, U Weiß, C von Schnakenburg, and R Hentschel (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, Milton I. (1961) 'Bed supply and hospital utilization: a natural experiment.' *Hospitals* 35, 36–42
- Shigeoka, Hitoshi, and Kiyohide Fushimi (2011) 'Supply induced demand in newborn treatment: Evidence from Japan.' Unpublished working paper, Columbia University.
- Vogl, Matthias (2012) 'Assessing DRG cost accounting with respect to resource allocation and tariff calculation: the case of Germany.' *Health Economics Review* 2, 15