Is it good to be too light?

The consequences of birth weight thresholds in hospital reimbursement systems

Simon Reif[*], Sebastian Wichert[†], Amelie Wuppermann[‡]

May, 28 2015

Preliminary and incomplete! Do not cite without permission!

Abstract

Birth weight manipulation is common in per-case hospital reimbursement systems, in which hospitals receive more money for otherwise equal newborns with birth weight just below compared to just above specific birth weight thresholds. These thresholds could further result in differential treatment of newborns as the hospitals receive more money for these cases and as medical guidelines suggesting differential treatment may coincide with the thresholds. We examine how birth weight thresholds in German hospitals are related to quantity and quality of care using an administrative claims dataset on the universe of hospital births in Germany for the years 2005-2011. Our results indicate that birth weights are manipulated more often for more fragile newborns. Furthermore, newborns just below the thresholds receive more care and are more likely to survive despite their higher fragility. This suggests that low birth weight newborns profit from having low birth weight (or from being fragile). Overall, a reconsideration of hospital reimbursement or treatment guidelines may be indicated.

JEL Codes: I110, I180,
Keywords: neonatal care, DRG upcoding, quanity & quality of care

[*]Friedrich-Alexander-University Erlangen-Nuremberg, Germany (simon.reif@fau.de)
[†]Ludwig-Maximilians-University Munich, Germany (sebastian.wichert@econ.lmu.de)
[‡]Ludwig-Maximilians-University Munich, Germany (amelie.wuppermann@econ.lmu.de)

1 Introduction

Small changes in birth weight can have important financial implications for hospitals in many of the widespread prospective payment schemes (PPS) which reimburse hospitals with a fixed rate for the treatment of strictly defined diagnosis related groups (DRGs). More specifically, hospitals receive a higher per-case reimbursement for newborns with birth weight just below certain thresholds than for newborns with weight above, leading to an incentive to underreport birth weight. The evidence is accruing that the introduction of birth weight thresholds has led to large underreporting - so called upcoding - of birth weight (e.g. Jürges and Köberlein 2014; Shigeoka and Fushimi 2014). At the same time, similar thresholds in birth weight are part of medical guidelines, resulting in discontinuities in the quantity and quality of care around birth weight thresholds (Almond et al. 2010).

While low birth weight is typically linked to worse health outcomes, having a (reported) birth weight below certain weight thresholds could be beneficial for newborns. First, the hospital gets higher reimbursement for the specific case and thus may be able to provide more (expensive) care. Second, medical guidelines may lead to different and more extensive treatment of newborns below certain thresholds because lower birth weight is known to increase risks. To our best knowledge, there is no empirical evidence up to date on how birth weight thresholds in DRG reimbursement systems affect the quantity and quality of care. In this paper, we fill this gap by analyzing the effect of birth weight thresholds in the German DRG system (G-DRG) on care provided and mortality using an administrative hospital claims dataset including information on the universe of hospital births in Germany for the years 2005-2011.

Earlier studies on the effect of thresholds in medical guidelines on quality and quantity of care (e.g. Almond et al. 2010) use the discontinuity in the probability of treatment around birth weight thresholds to estimate the effect of medical care on mortality under the assumption that newborns with birth weight just below and just above the threshold do not vary systematically in any observed or unobserved health-related characteristic other than birth weight. Observed differences in mortality can then be interpreted as causal effects of differences in treatment intensity induced by different medical guidelines around the thresholds. In systems, in which hospitals have an incentive to manipulate birth weights around the relevant thresholds, the assumption of equal underlying health around birth weight thresholds requires that birth weight manipulation is not related to the newborns' underlying health status.[1]

In this paper, we describe the health-related incentives for birth weight manipulation

[1]Barreca et al. (2011) challenge the validity of this assumption also in the setting without upcoding incentives.

in German hospitals. Furthermore we empirically test whether observable characteristics that are known right after birth, cannot be manipulated easily (such as newborn's sex, type of birth (single, multiple) and gestational age), and are known predictors of newborn mortality[2] are smooth around the birth weight thresholds. We do find some systematic differences that suggest that birth weight manipulation arises with higher probability for more fragile newborns. The paper proceeds by comparing quality and quantity of care around the birth weight thresholds, holding fixed observable health-related characteristics of the newborns. In the interpretation of the results we take into account that newborns with weight on the different sides of the thresholds might still differ in unobservable characteristics.

Our main results are based on the universe of births that took place in German hospitals between 2005-2011.[3] The data stem from the DRG reimbursement system and contain information on newborns' gender, birth weight in grams, hours of ventilation, length of the hospital stay, mortality in the hospital or within the first 28 days of the newborn's life, and specific DRGs coded as well as detailed information on diagnoses and procedure codes.

Consistent with the existing literature, we find large amounts of birth weight manipulation in German neonatology. Upcoding is more prevalent in specialized neonatal intensive care units (NICUs) but the share of upcoded newborns is only loosely linked to excess reimbursement. Furthermore, quantity of care is discontinuous at birth weight thresholds. Newborns with weight directly below certain thresholds receive significantly more procedures and stay longer in the hospital than newborns just above the thresholds. In addition, newborns below the thresholds survive with a higher probability.

Our results indicate that quantity of care is generally higher slightly below birth weight thresholds and that newborns with weight below the thresholds are also more likely to survive. This finding is striking, as the observed birth weight manipulation in Germany tends to increase the share of fragile newborns below the threshold. The additional care that these fragile newborns receive compared to newborns with reported weight just above the threshold thus seems to pay off.

The rest of the paper is structured as follows: In section 2, we give an overview on the institutional background in Germany, specifically the G-DRG system and German neonatology regulations. We introduce our data and empirical strategy in section 3. In section 4, we present our results. The paper closes with a short discussion and conclusion

[2]German Medical Guidelines recommend that physicians take these characteristics into account when counseling parents of at risk newborns (see "Frühgeborene an der Grenze der Lebensfähigkeit" AWMF-Leitlinien-Register Nr. 024/019)

[3]In Germany, only roughly 1% of births do not take place in hospitals (see Loytved (2014)).

in section 5.

2 Institutional Background

In this section we give a brief overview on the G-DRG system and the general effects of its introduction. We then outline the institutional background of neonatal care in Germany, explain in detail how DRGs operate in this environment, and why treatment may vary across birth weight thresholds.

2.1 The G-DRG reimbursement system

Until the year 2003, German acute care hospitals were reimbursed through a multiple-source-system consisting of a hospital specific patient/day base-rate, a specialty specific rate and case-based lump-sums. In order to increase efficiency and, in particular, reduce length of stay, the G-DRG system was introduced. Under the old and the new system, reimbursement was and is generally the same irrespective of individuals' insurance status (public or private). Based on the Australian DRG system, 664 DRGs defined by combinations of diagnoses, performed procedures, hours of ventilation, age and (for perinatal DRGs) birth weight were set up. To each DRG a case-weight which determined the final reimbursement was assigned. This case-weight multiplied by a base-rate gives the amount of money a hospital gets for treating the respective patient.[4] In a transition period from 2003 to 2010 hospital specific base-rates, which over the years advanced towards state specific rates, were used.[5] Since 2010 most hospitals are reimbursed according to this state base-rate (Busse and Riesberg 2005). From 2010 to 2014 these state specific base rates in turn narrowed down to a federal base rate interval.

Case-weights and group definitions vary from year to year. The German Institute for Hospital Reimbursement[6] decides on changes of case-weights and DRG definitions every year using information on actual costs occurring for different treatments in certain baseline-hospitals which voluntarily provide their (usually non-public and confidential) cost structure data to the institute. Since the G-DRG introduction, the number of groups almost doubled to 1200 in 2015. Many case-weights changed over time. Especially in the

[4] As long as the length of stay is within a specific interval for each DRG. If the patient stays longer, the extra costs are partially reimbursed by additions while shorter length of stay leads to partial deductions.

[5] The reason for this was that hospital cost structures were drastically different when the G-DRG system was introduced. Starting off directly with identical reimbursement in each state would have lead to major financial struggles for hospitals.

[6] Institut für das Entgeldsystem im Krankenhaus (InEK), financed by private and social health insurers as well as the German Hospital Association.

first years many complex DRGs were given a higher weight, while DRGs based on simpler procedures got lower case-weights after initial evaluation. These changes influenced coding behavior in hospitals. Schreyögg et al. (2014) for example show that a 1% increase in the case-weight on average leads to a 0.2% increased frequency of this DRG being used. The authors state that these increases are expectable market reactions. The results could indicate that upcoding takes place in German hospitals.

2.2 Neonatal Care in Germany

Incentives for upcoding are especially present in neonatal care. In order to understand the mechanisms behind physicians' coding behavior, some knowledge on the institutional background is vital. Depending on the expected complexity of treatment, mothers should deliver in more or less specialized hospitals or – if severity is diagnosed post-birth – newborns should be moved to an appropriate clinic. In Germany, four degrees of specialization are defined by the Gemeinsamer Bundesausschuss (2013)[7]: Level 4 indicates regular maternity clinics, Level 3 means that basic neonatal care can be provided, Level 2 clinics have some specialization in neonatal care and Level 1 clinics maintain neonatal intensive care units (NICUs). One crucial indicator for the assignment of newborns to clinics is birth weight. Pre-term births above 1500 grams should be treated in Level 3 clinics, birth weight between 1250 and 1500 grams induces Level 2 care, while newborns below 1250 grams should be treated in Level 1 clinics.[8]

An international comparison shows that Germany has a higher number of NICUs per capita than similar countries like Sweden, the United Kingdom or the United States (Zimmer 2012). Due to the high fixed costs of such clinics, high competition between hospitals could provide incentives to manipulate birth weight and hence increase the number of newborns "in need" of specialized neonatal care.[9]

2.3 DRGs in German Neonatology

When DRGs were first put into use in Germany in 2003, 38 different groups for the treatment of newborns had been defined. Until 2015 this number has slightly increased to

[7]The highest self administration unit in the German health sector.

[8]Interestingly, the *German Association for Perinatal Medicine* uses different Level labels (1, 2a, 2b and 3) and slightly different thresholds which would admit less newborns to NICUs (Bauer et al. 2006). Nevertheless, the Gemeinsamer Bundesausschuss (2013) regulations are binding.

[9]Theoretically this structure could offer an incentive for hospitals to increase birth weight of babies that would be too light for this type of clinic and would be transferred otherwise. Since case-weights are calculated for average newborns within a birth weight bracket, reimbursement would most likely not cover all the costs for newborns at lower end of a birth weight DRG, making such a behavior implausible (Jürges and Köberlein 2014).

42 groups. There are special DRGs for some newborns dying within 4 days after birth and also for cardiovascular interventions. The majority of DRGs, however, (37/42 DRGs in 2008) are defined along eight birth weight thresholds and within each threshold multiple DRGs cover different degrees of treatment complexity and complications. We identified six DRG clusters of varying complexities and complications wherein birth weight is the only grouping criterion. Table A.1 (in the Appendix) shows which DRGs are used for different birth weight intervals, holding everything other than birth weight constant within a cluster. Cluster 1 in table A.1 comprises the most severe cases, cluster 6 the least severe ones.

How hospital reimbursement is affected by changes in birth weight within clusters and across different birth weight thresholds can be inferred from table A.2 (in the Appendix). This table contains the case-weights for these DRGs for the years 2005-2011. Figure 1 shows graphically, how changes in case-weights within clusters occur with changes in birth weight for the year 2008. The figure shows that the case-weight decreases discontinuously with birth weight across the different thresholds in all clusters. At very low birth weights there are only two different groups of severity: Clusters 1-2 and clusters 3-6. Starting from a birth weight of 1000 grams and above, 3 severity clusters apply as the joint cluster 1 and 2 is split into two severity groups. At birth weights of 1500 grams and higher all 6 clusters have different DRGs and case-weights.

Total hospital reimbursement depends on the case-weight of a specific DRG and the hospital's base-rate. As an example for how small changes in birth weight can affect reimbursement let's consider a notional hospital in the German state of Hesse in the year 2010 which has a base-rate equal to the state base-rate of €2968.56. A generic newborn in this hospital may have had a birth weight of 1004 grams and was assigned DRG P63Z[10] which in this year had a case-weight of 8.776. Assuming treatment took place within the regular length of stay for this DRG (15-62 days) total reimbursement to the hospital would have been €26,052.08. Slight manipulation of birth weight to 999 grams had assigned DRG P62D with a case-weight of 15.51. Given length of stay was also in the regular interval for this DRG (21-80 days)[11], reimbursement would have increased to €46,042.37, a plus of almost €20,000 for a 5 gram manipulation.

Incentives to upcode are present irrespective of the newborns' health status. As can be seen in tables A.1, A.2, and figure 1, a decrease in weight from above to below the thresholds generally results in an increase in case-weights that translates into increased

[10]For example, because it was a girl, ventilation was in use for 48 hours, respiratory distress syndrome (ICD-10-GM: P28.5), systematic inflammatory response-syndrome (ICD-10-GM: R65.0), and some other infectious disease (ICD-10-GM: P37.9) were diagnosed and the girl was treated by monitoring cardiovascular levels (OPS: 8-930), was given a VR-infusion (OPS: 8-811.0), CAPA ventilation was performed (OPS: 8-711.00) and after this did not work out, endotracheal intubation (OPS: 8-701) came into use.

[11]This is not unrealistic as the mean length of stay for DRG P63Z in 2010 was 45.2 days.

reimbursement amounts. Exceptions are decreases in birth weight below 1250 grams in the most severe clusters (1 and 2) in all years, as well as decreases below 875 grams in the most severe clusters (1 and 2) in 2011. Decreasing birth weight in these cases does not change the DRG. In addition, decreasing birth weight below 600 grams leads to a slight decrease not an increase of the case weight but only for the most severe clusters (1 and 2) in 2011. In general, figure 1 suggests that the increases in case-weights are larger at the birth weight thresholds below 1500 grams compared to above, giving hospitals larger incentives to underreport birth weight around the low weight thresholds. Looking at each specific threshold, figure 1 suggests that for some thresholds upcoding incentives are higher for more severe cases, while for others incentives are larger among the less severe cases. Across all thresholds, there is thus no systematic rule that hospitals should be more likely to upcode more or less healthy newborns.

– FIGURE 1 ABOUT HERE –

A newborn's underlying health status could still be related to the probability to be upcoded for two different reasons. On the one hand, hospitals could expect higher costs for less healthy newborns holding actual birth weight constant. In order to cover these expected costs on a case-by-case basis, they could be more inclined to underreport birth weight for the more frail newborns. On the other hand, for newborns who are very frail and can be expected to die within the first 4 days of their life, the hospital has no incentive to upcode the birth weight as birth weight does not change the reimbursement for these cases. Since the incentives to underreport birth weight are not unambiguously related to the newborn's underlying health status, it remains an empirical question whether health is related to upcoding. We investigate this by looking at differences in observable health-related characteristics around the birth weight thresholds in the next section.

2.4 DRG Thresholds and Medical Guidelines

The German Association of the Scientific Medical Societies (AMWF) publishes a multitude of medical guidelines for neonatal care from the different medical societies, ranging from guidelines on treatment of very frail newborns to care for normal healthy newborns as well as after hospital care for specific groups of newborns. Many of these guidelines contain information on specific disease risks for different birth weight groups where the birth weight thresholds align with the thresholds that determine the DRGs. Examples are risks of complications like bronchopulmonary dysplasia, necrotizing enterocolitis, and patent ductus arteriosus. For the first, the guideline also gives different therapeutic recommendations depending on whether the birth weight is above or below 1000 grams,

which may directly induce differential treatment. Reporting of the risks by birth weight groups may also result in differential treatment, if physicians classify newborns in risk groups according to their birth weight and this changes awareness for risks of the different conditions.

– FIGURE 2 ABOUT HERE –

Differential treatment depending on birth weight might in addition be triggered by the design of the DRG system itself. Hospitals only receive the full reimbursement for a case if the patient stays in the hospital for at least a minimum number of days. If the patient leaves the hospital before the minimum length of stay the hospital only receives a reduced payment. Figure 2 indicates how the minimum length of stay varies with birth weight for the different DRG severity-cluster defined above in the year 2008. Overall, minimum length of stay varies from 39 days for very light newborns with complications (Clusters 1 and 2) to 2 days for relatively healthy newborns with birth weight above 2499 grams. Except for the birth weight threshold 750 grams, minimum length of stay decreases sharply when birth weight crosses the different DRG thresholds. These discontinuities in minimum length of stay introduce an incentive for hospitals to keep newborns with (reported) birth weight just below the specific threshold longer in the hospital than newborns with birth weight at the threshold or above – possibly leading to differences in treatment for otherwise very similar newborns.

3 Data and Empirical Strategy

Our data include in the universe of all hospital claims in Germany from the years 2005-2011. They are collected as part of the hospital reimbursement process. Mapping of diagnoses and procedures to DRGs takes place in the hospitals. This information is then sent to the patient's health insurance for reimbursement. The German Federal Statistical Office makes these data available to researchers under strict non-disclosure regulations. For our analysis, we restrict the sample to all births with a birth weight between 450 and 3000 gram that took place in 2005-2011, amounting to 985,875 cases.

– TABLE 1 ABOUT HERE –

The data include some basic information on the newborns, e.g. gender, birth weight, type of birth (single, twin or more than twin), and detailed medical information, such as exact diagnoses (ICD-10-GM) codes, procedure codes (the German version of ICPM codes, called OPS codes), hours of ventilation and length of the hospital stay. Table 1

displays descriptive statistics for all births with weight between 450 and 3000 grams. It also shows averages for newborns that fall between the different birth weight thresholds. On average, newborns have a birth weight of 2601 grams, have 2.6 reported procedures, and stay in the hospital for 8.5 days. They receive ventilation for an average of nearly 19 hours. Only 0.6% of newborns die during their hospital stay or within their first 28 days of life. 44% of all newborns are boys, 19% of births are premature, i.e. born with at least 28, but less than 37 completed weeks of gestation. Moreover 1.2% are extremely premature, i.e. have a gestational age of less than 28 weeks. 11% of newborns are twins.Moreover, in table 1 we look explicitly at two leading causes of infant death. The first condition is whether children had mild or severe neonatal asphyxia, a form of oxygen deprivation. The ICD-10-GM coding system contains two different codes for asphyxia, one for newborns with severe asphyxia and a 1-minute Apgar score[12] 0-3 (ICD-10-GM: P.21.0), and one for newborns with mild asphyxia and a 1-minute Apgar score of 4-7 (ICD-10-GM: P.21.1). Among all newborns in our sample, 1.6% had mild asphyxia and 0.5% severe asphyxia. The second common threat for a newborns' life, that we look at, is the infant respiratory distress syndrome (IRDS; ICD-10-GM: P22.0) which is a lung malfunction often caused by preterm birth. About 4.6% of all newborns in our data suffer from IRDS.

Unfortunatly, linking the data on newborns to their mothers – who have their own hospital case – is impossible. The data therefore contain almost no information on what happened during pregnancy or during the birth itself. Similar information, however, would have been useful to gain more insights on what triggers birth weight manipulation. Table 1 contains information on one variable that gives information on the newborn's mother and her behavior during pregnancy: whether the mother smoked. This is derived from diagnoses codes that newborns receive. On average, however, only 1.2% of newborns in our data have mothers who smoked during pregnancy.

Table 1 further provides information on how the different variables change across birth weight thresholds - looking at all newborns with weight in the respective birth weight brackets from below 600, 600–749, 750–874, 875–999, 1000–1249, 1250–1499, 1500-1999, 2000-2499, and 2500-3000 grams. Starting from the group with 600 grams or above, the development with increases in birth weight is as expected: treatment intensity (length of stay, number of procedures, hours of ventilation) increases, while mortality decreases. At the same time the general health indicators (gender, pre-maturity and extreme pre-maturity, whether birth is a twin birth, asphyxia, IRDS) indicate improvements in the newborns underlying health. The picture looks a little different when going from below

[12]The Apgar score was developped in the 1950s to summarize a newborn's health. It measures Appearance (skin color), Pulse (heart rate), Grimace (reflex irritability), Activity (muscle tone), and Respiration.

600 grams to the group with 600–749 grams. Here intensity of treatment increases with an increase in weight. This however, is likely driven by the higher mortality rate among the lighter newborns.

– FIGURE 3 ABOUT HERE –

Figure 3 presents evidence in line with the results of Jürges and Köberlein (2014) for underreporting of birth weight in German hospitals. It shows birth weight frequencies pooled across the years 2005-2011 for all births below 2700 grams and the eight different DRG birth weight thresholds. Statistically implausible jumps in frequencies slightly below thresholds, especially at 1250, 1500, 2000, and 2500 grams indicate that hospitals underreport birth weight. More thorough evidence on upcoding of birth weight can be found in Appendix B. The results presented there suggest that upcoding takes place at all thresholds and irrespective of the exact measure used to identify upcoding.

Since the earlier literature shows that slightly lower birth weight can be beneficial for newborns as it may trigger additional or different treatment (Almond et al. 2010), we focus on the question whether newborns in Germany with birth weights just below the relevant DRG thresholds receive different care than newborns just above the thresholds. To interpret differences around the thresholds as causal effects of having lower (reported) birth weight, we would have to assume that newborns with weight just below and just above the thresholds do not differ in other observed or unobserved characteristics. More specifically, we need to assume that reporting of birth weight is not related to the newborns' underlying health. To test whether reported birth weight depends on health, we investigate differences in observable health-related characteristics around the DRG thresholds.

Importantly, weighing should take place shortly after birth. Once determined, birth weight is reported to the registrar's office. No further adjustment to the reported weight is possible without committing an easily detectable fraud. If health related information on the newborn is taken into account in reporting birth weight, this information would have to be available to the midwife or nurse who performs the weighing right at birth. We therefore look at characteristics that are likely observable to nurses and midwives at birth and may be related to the newborns' health, such as the newborns' gender, type of birth (single, twin, or higher order), whether the birth is premature, and several diagnoses codes that either measure neonatal illnesses usually appearing directly after birth (like asphyxia and IRDS) or maternal behavior during pregnancy (such as smoking).

– TABLE 2 ABOUT HERE –

The mean differences in the latter observable characteristics around the different birth weight thresholds for 50-grams windows of birth weight below and above each threshold

are shown in table 2. Positive signs indicate that a higher fraction of newborns below the threshold has the respective condition. Except for the lowest threshold (600 grams), the observed differences point into the direction that more fragile newborns more likely have weights below the thresholds. Newborns below are more likely to be (extremely) pre-mature, for the highest thresholds more likely male and twins. Furthermore, for all thresholds more newborns below that above have IRDS. The last 4 columns of table 2 report differences across placebo thresholds. Some of these differences are also significantly different from zero – indicating that the differences observed for the actual thresholds may reflect that children within 100 gram birth weight intervals may truly differ in terms of health. However, the differences that we observe around the placebo thresholds are much smaller than the differences around the real thresholds. In future work, we will limit our results to 25 gram intervals to eliminate "true" birth weight related health differences around the thresholds.

Overall, table 2 suggests that - except for around the lowest threshold - birth weight manipulation may be more frequent for more fragile newborns. In addition, there could be characteristics unobservable to us as researchers - such as information on the birth itself and the pregnancy - that are related to the newborns' health and determine upcoding. These differences in observed characteristics as well as possible additional differences in unobserved characteristics should be borne in mind in the interpretation of our results.

We compare quantity and quality of care for newborns in windows of reported birth weight just below and just above the different thresholds in a regression discontinuity-like approach. We include the observable health-related variables as covariates to control for health-related upcoding of birth weight. For a causal interpretation of our results, we have to assume that no further information is available to nurses or midwifes that influences reporting of birth weight and the quantity and quality of treatment. Given that we find differences in observed characteristics, this assumption seems rather strong. We discuss the likely sign of the biases in our results that would arise if the assumption does not hold in the concluding section.

4 Quantity and Quality of Care

We use three different variables in our data to measure quantity of care that newborns receive. The first measure is the number of days that newborns stay in the hospital. As described above, the length of the hospital stay may vary around the DRG birth weight thresholds because the minimum number of required days for full reimbursement jumps at the birth weight thresholds. The second measure we analyze is the number of procedures

that newborns receive during their hospital stay. The third measure reflects the hours of ventilation that a newborn received. The latter three measures are of course related to each other as more procedures can be performed and longer ventilation can take place when the hospital stay increases.

To measure quality of care we use mortality in the hospital. Assuming that newborns with weight just below and just above the threshold have similar mortality risks a priori, differences in mortality should result from differential treatment. A disadvantage of using mortality as a measure for quality of care is that mortality is a very extreme health outcome that is very rare. We cannot use it to infer more general differences in newborns' health status that result from differences in treatment.

Table 2 reports simple mean differences between newborns within a 50-gram bandwidth around the coding-relevant and placebo weight thresholds for all outcome variables. The results suggest that quantity of care varies at almost all birth weight thresholds. The largest effects can be seen at the 1000 gram threshold. Newborns with birth weight below 1000 grams stay in the hospital for 5.9 additional days, receive 1.3 additional procedures and get 93.5 more hours of ventilation compared to newborns with weight of 1000 grams or above.

The mean differences in mortality are negative at most birth weight threshold, but this difference is only significant at the 1500 and 2000 gram thresholds. Simple mean comparisons thus suggest that newborns with weight below the thresholds have a lower probability to die, despite their lower weight (and more fragile health).

– TABLE 3 — 5 ABOUT HERE –

Tables 3 - 6 report regression discontinuity estimates of the differences in the outcomes around the thresholds. Each table reports results for one specific outcome for each threshold with and without control variables and using quadratic and cubic functions of birth weight to fit the data. Neither the choice of fit nor whether control variables are included or not has a large impact on the results. The results basically confirm the simple mean comparisons discussed above. Newborns with weight just below the thresholds generally stay longer in the hospital than newborns with weight above. This difference is, however, only significant at the 750, 1000, 1500, 2000 and 2500 gram thresholds. Similarly, the number of performed procedures is generally higher among newborns with weight below the thresholds, but the differences are only robustly significant for the highest threshold. Hours of ventilation does not vary robustly around thresholds. Some increases in hours of ventilation can be seen at 2000, and 2500 gram at the thresholds. The results for length of stay and number of procedures are also displayed graphically for each threshold in figures

4 and 5. While large discontinuities in length of stay and number of procedures occur also at the lower thresholds, the figures show large variation around the thresholds, which may be due to relatively small sample sizes, resulting in imprecise estimates. For the 2500 gram threshold, the results are precisely estimated and show small but significant discontinuities.

– FIGURE 4 and 5 ABOUT HERE –

Table 6 displays the results for mortality. It indicates that mortality is lower below the 600, 750, 1250, 1500, and 2000 thresholds.

– TABLE 6 ABOUT HERE –

Taken together, newborns below the thresholds are more likely more fragile and receive more intensive care, but are less likely to die. As newborns vary in observed and possibly unobserved health around the thresholds, we cannot conclude that newborns in German hospitals receive more care because their weight is below a certain threshold. It may well be the case that exactly the same care would have been provided to a newborn if she had a reported weight above the threshold. Instead it seems as if hospitals manipulate birth weight to cover the costs that they expect for fragile cases. They then get adequately reimbursed for the intensive treatment that the newborns need. This treatment then seems to lower these newborns mortality. This suggests that also the newborns with weight above the threshold may profit from additional treatment.

5 Discussion and Conclusion

In this paper, we investigate how birth weight thresholds in reimbursement systems affect the quantity and quality of health care delivered to newborns. The empirical challenge in our analysis is that due to underreporting of birth weight in hospital reimbursement systems, in which small changes in birth weight induce large changes in reimbursement, newborns with reported weight below the threshold may vary systematically from newborns above the threshold. This may occur if incentives to underreport birth weight vary by observed or unobserved health-related characteristics of the newborns.

In the German hospital reimbursement system, hospitals have incentives to underreport the weight of frail newborns, if they want to cover the higher expected costs of treating these newborns. Our results indeed suggest that hospitals seem to underreport birth weight more for more fragile newborns. This renders the usual assumptions

required for causal interpretation of regression-discontinuity analyses using discontinous changes around thresholds implausible. We therefore refrain from interpreting our results on quantity and quality of care causally.

Nevertheless, the results on quality and quantity of care are interesting – in particular as we see that fragile newborns are more likely to be upcoded. The results show that quantity of care and quality of care are higher for newborns with weight below the thresholds. That is, newborns below the thresholds stay longer in the hospital and are treated with a higher number of procedures. In addition, their mortality risk is lower. These results suggest that the additional care that the lighter newborns receive compared to the slightly heavier ones has relevant health impacts.

Given the observables differences between the newborns above and below the thresholds, we cannot conclude that the lighter newborns only received the additional care because their weight was just below the threshold. Instead, they may have received the additional care because they appeared more fragile and that they would have received this additional care independent on their exact birth weight. Thus, we cannot conclude that it is necessarily good for a newborn to have a reported birth weight below one of the relevant thresholds. Instead, it may be good for a newborn to appear more frail in terms of observable characteristics that are known determinants of newborn mortality.

Overall, however, the additional care that the lighter newborns receive - whatever may be the reason - seems to be effective, as their mortality risk is lower despite their higher fragility. This suggests that in general the discontinuities in quantity of care may harm newborns that are just not light (or not frail) enough to receive the additional care. Discontinuities - linked to birth weight or other values - in hospital reimbursement as well as in medical guidelines should therefore be reconsidered.

Tables and Figures

Table 1: Summary statistics, all hospital births, 2005-2011

	<600	600-749	750-874	875-999	1000-1249	1250-1499	1500-1999	2000-2499	2500-3000	All births
Birth weight (gram)	527.092	680.814	816.950	948.530	1,149.752	1,404.486	1,808.099	2,306.878	2,808.360	2,601.269
	(43.828)	(45.423)	(34.144)	(37.177)	(69.413)	(72.487)	(134.899)	(136.259)	(140.054)	(441.936)
Length of stay (days)	53.706	69.947	67.946	61.468	50.428	39.012	23.776	9.821	4.662	8.452
	(62.826)	(55.513)	(43.185)	(34.718)	(27.785)	(21.412)	(16.439)	(9.564)	(4.196)	(14.196)
Procedures (#)	11.729	14.354	13.638	12.192	10.203	8.296	5.942	3.208	1.724	2.559
	(11.393)	(10.689)	(9.407)	(7.569)	(6.662)	(5.129)	(4.154)	(2.988)	(1.601)	(3.275)
Ventilation (hours)	711.444	806.233	656.342	457.690	245.815	104.302	33.314	7.444	1.381	18.622
	(918.528)	(771.244)	(599.021)	(474.152)	(343.694)	(205.969)	(126.909)	(69.779)	(34.656)	(140.817)
Mortality	0.487	0.236	0.126	0.069	0.042	0.025	0.012	0.003	0.001	0.006
	(0.500)	(0.425)	(0.332)	(0.253)	(0.201)	(0.155)	(0.111)	(0.057)	(0.028)	(0.077)
Male births	0.474	0.502	0.521	0.529	0.518	0.501	0.484	0.453	0.432	0.442
	(0.499)	(0.500)	(0.500)	(0.499)	(0.500)	(0.500)	(0.500)	(0.498)	(0.495)	(0.497)
Extreme prematurity	0.590	0.581	0.495	0.369	0.146	0.037	0.010	0.003	0.001	0.012
	(0.492)	(0.493)	(0.500)	(0.483)	(0.353)	(0.188)	(0.102)	(0.055)	(0.025)	(0.107)
Prematurity	0.158	0.198	0.261	0.364	0.527	0.611	0.592	0.350	0.108	0.188
	(0.364)	(0.399)	(0.439)	(0.481)	(0.499)	(0.488)	(0.491)	(0.477)	(0.311)	(0.391)
Twin births	0.172	0.173	0.186	0.211	0.229	0.260	0.303	0.229	0.058	0.107
	(0.378)	(0.379)	(0.389)	(0.408)	(0.420)	(0.439)	(0.459)	(0.420)	(0.235)	(0.309)
Severe asphyxia - AGPAR score 0-3	0.065	0.055	0.051	0.036	0.029	0.022	0.013	0.006	0.003	0.005
	(0.247)	(0.228)	(0.220)	(0.187)	(0.168)	(0.145)	(0.115)	(0.079)	(0.050)	(0.071)
Moderate asphyxia - AGPAR score 4-7	0.059	0.074	0.072	0.065	0.065	0.057	0.041	0.022	0.010	0.016
	(0.235)	(0.262)	(0.258)	(0.246)	(0.246)	(0.233)	(0.198)	(0.147)	(0.098)	(0.124)
Infant respiratory distress syndrome	0.609	0.756	0.748	0.706	0.591	0.428	0.194	0.045	0.006	0.046
	(0.488)	(0.429)	(0.434)	(0.456)	(0.492)	(0.495)	(0.395)	(0.207)	(0.080)	(0.210)
Maternal smoking	0.013	0.019	0.023	0.029	0.027	0.030	0.037	0.027	0.006	0.012
	(0.112)	(0.136)	(0.150)	(0.167)	(0.161)	(0.170)	(0.189)	(0.163)	(0.076)	(0.108)
Number of births	3,059	4,588	3,461	5,135	9,698	15,044	49,707	167,621	727,562	985,875

Notes: Standard deviations in brackets below means. Procedures are the coded number of OPS (German modification of ICPM) codes. Mortality is the share of newborns died in hospital. Extreme prematurity is defined as birth with gestational age less than 28 weeks (ICD-10-GM: P07.2). Prematurity is defined with 28 completed weeks or more but less than 37 completed weeks of gestation (ICD-10-GM: P07.3). **Source:** Own calculations based on the DRG-Statistic.

Table 2: Differences above and below threshold, 2005-2011

	T = 600	T = 750	T = 875	T = 1000	T = 1250	T = 1500	T = 2000	T = 2500	T = 700	T = 1300	T = 2200	T = 2700
Health-related controls												
Male births	.004	-.029	-.016	.004	.012	.026	.017*	.019***	.003	.004	.006	.003
	(.02)	(.019)	(.018)	(.019)	(.016)	(.016)	(.008)	(.004)	(.018)	(.017)	(.006)	(.003)
Extreme Prematurity	-.022	.063**	.047**	.115***	.018*	-.002	.001	0	.023	.011	0	.001**
	(.02)	(.019)	(.018)	(.017)	(.008)	(.005)	(.001)	(0)	(.017)	(.008)	(.001)	(0)
Prematurity	.015	.006	-.013	-.058**	-.015	.012	.08***	.014***	-.03*	-.033*	.041***	.03***
	(.015)	(.016)	(.017)	(.019)	(.016)	(.015)	(.008)	(.003)	(.014)	(.017)	(.006)	(.002)
Twin births	-.014	-.013	-.026	-.023	-.025	.003	-.007	.031***	.005	.01	.021***	.019***
	(.015)	(.015)	(.015)	(.016)	(.014)	(.014)	(.007)	(.003)	(.013)	(.015)	(.005)	(.002)
Severe asphyxia - AGPAR score 0-3	.007	-.006	.008	0	.004	-.002	0	.001*	.018*	-.007	.002*	.001*
	(.009)	(.009)	(.007)	(.007)	(.005)	(.005)	(.002)	(.001)	(.008)	(.005)	(.001)	(0)
Moderate asphyxia - AGPAR score 4-7	-.004	.01	.004	.009	.019*	.008	.005	.006***	.005	-.016*	.001	0
	(.01)	(.01)	(.009)	(.009)	(.007)	(.007)	(.003)	(.001)	(.009)	(.008)	(.002)	(.001)
Infant respiratory distress syndrome	-.047*	.06***	.035*	.073***	.054***	.106***	.049***	.014***	-.048**	.014	.011***	.002**
	(.018)	(.017)	(.016)	(.019)	(.016)	(.014)	(.005)	(.001)	(.015)	(.017)	(.003)	(.001)
Maternal smoking	-.006	.002	.001	.009	.006	.003	.002	.008***	-.008	-.004	.003	.002***
	(.005)	(.005)	(.006)	(.006)	(.005)	(.005)	(.003)	(.001)	(.005)	(.005)	(.002)	(.001)
Outcomes												
Length of stay (days)	-5.318*	4.32*	2.608	5.87***	2.424**	4.228***	3.931***	1.316***	-1.798	2.898***	1.208***	.239***
	(2.432)	(1.811)	(1.411)	(1.255)	(.84)	(.723)	(.204)	(.052)	(1.944)	(.858)	(.131)	(.027)
Procedures (#)	-.74	.885*	.643*	1.292***	.767***	.856***	.973***	.542***	.109	.153	.28***	.082***
	(.469)	(.38)	(.293)	(.29)	(.192)	(.171)	(.057)	(.019)	(.374)	(.19)	(.04)	(.01)
Ventilation (hours)	-49.912	80.607**	67.835***	93.478***	46.457***	12.119	6.435***	1.793***	38.572	7.35	1.302	.02
	(34.467)	(25.866)	(19.528)	(15.922)	(8.715)	(8.399)	(1.695)	(.441)	(26.931)	(8.859)	(.946)	(.196)
Mortality	.07***	-.006	.02	0	-.006	-.024***	-.004**	0	.088***	.003	0	0
	(.019)	(.014)	(.01)	(.01)	(.006)	(.006)	(.001)	(0)	(.014)	(.006)	(.001)	(0)
N_L	1224	1949	1551	2957	3293	5822	10118	31593	1377	1424	12127	51591
N_R	1262	1041	1501	856	1424	1259	6926	33211	1949	2357	15025	62761

Notes: Difference between mean values below and above birth weight threshold T within a 50 gram bandwidth (X_{below}-X_{above}). Standard errors below mean difference in parentheses. N_L: Number of observations left/below threshold within 50 gram bandwidth. N_R: Number of observations right/above threshold within 50 gram bandwidth. *** p<0.001; ** p<0.01 and * p<0.05. Procedures are the coded number of OPS (German modification of ICPM) codes. Mortality is the share of newborns died in hospital. Extreme prematurity is defined as birth with gestational age less than 28 weeks (ICD-10-GM: P07.2). Prematurity is defined with 28 completed weeks or more but less than 37 completed weeks of gestation (ICD-10-GM: P07.3). **Source:** Own calculations based on the DRG-Statistic.

Table 3: Regression Discontinuity Estimates - Length of stay in days, 2005-2011

	Coeff	Robust SE	Controls	Fit	N_L	N_R
T=600	7.659	4.144		Q	3,059	4,588
	3.373	3.744	√			
	18.166**	5.562		C		
	8.619	5.078	√			
T=750	10.490**	3.381		Q	4,588	3,461
	4.738	3.132	√			
	18.965***	4.473		C		
	12.623**	4.118	√			
T=875	-0.110	2.667		Q	3,461	5,135
	-0.532	2.453	√			
	4.766	3.605		C		
	4.076	3.295	√			
T=1000	2.852	1.741		Q	5,135	9,698
	1.041	1.603	√			
	6.746**	2.511		C		
	4.572*	2.292	√			
T=1250	0.277	1.066		Q	9,698	15,044
	-0.035	1.005	√			
	2.523	1.527		C		
	1.839	1.445	√			
T=1500	0.844	0.537		Q	15,044	49,707
	0.456	0.505	√			
	3.104***	0.809		C		
	1.953*	0.766	√			
T=2000	1.551***	0.194		Q	49,707	167,621
	1.159***	0.184	√			
	3.460***	0.257		C		
	2.407***	0.244	√			
T=2500	0.647***	0.054		Q	167,621	727,562
	0.633***	0.050	√			
	1.305***	0.073		C		
	1.046***	0.069	√			
T=700	5.086	4.862		Q	1,377	3,859
	5.102	4.658	√			
	8.413	6.188		C		
	5.709	6.002	√			
T=1300	-0.063	2.317		Q	1,424	14,879
	0.614	2.208	√			
	8.098*	3.675		C		
	7.717*	3.647	√			
T=2200	-0.189	0.191		Q	38,955	128,666
	-0.095	0.179	√			
	0.507	0.269		C		
	0.358	0.252	√			
T=2700	0.027	0.041		Q	170,695	556,867
	0.023	0.038	√			
	0.004	0.058		C		
	-0.026	0.054	√			

Notes: Regression Discontinuity Estimations using two birth weight intervals as bandwidth. Coeff: Coefficient for observations below threshold. Robust SE: robust standard errors. Controls: Using control variables for gender, prematurity status, multiple births, asphyxia, IRDS, ward type, and maternal smoking, alcohol consumption and drug use. Fit: Birth weight functions, fitted separately on each side of threshold; Q: quadratic function, C: cubic function. N_L: Number of observations left/below threshold. N_R: Number of observations right/above threshold. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. **Source:** Own calculations based on the DRG-Statistic.

Table 4: Regression Discontinuity Estimates - Number of Procedures, 2005-2011

	Coeff	Robust SE	Controls	Fit	N_L	N_R
T=600	1.556*	0.758		Q	3,059	4,588
	0.598	0.662	√			
	3.705***	1.012		C		
	1.582	0.895	√			
T=750	1.166	0.686		Q	4,588	3,461
	0.060	0.625	√			
	2.214*	0.886		C		
	0.949	0.799	√			
T=875	0.525	0.600		Q	3,461	5,135
	0.439	0.555	√			
	1.026	0.847		C		
	0.897	0.783	√			
T=1000	0.181	0.387		Q	5,135	9,698
	-0.248	0.356	√			
	0.813	0.528		C		
	0.256	0.479	√			
T=1250	0.367	0.241		Q	9,698	15,044
	0.283	0.220	√			
	0.885**	0.322		C		
	0.704*	0.294	√			
T=1500	0.190	0.132		Q	15,044	49,707
	0.048	0.120	√			
	0.718***	0.200		C		
	0.361	0.185	√			
T=2000	0.480***	0.054		Q	49,707	167,621
	0.323***	0.049	√			
	1.046***	0.073		C		
	0.655***	0.066	√			
T=2500	0.312***	0.019		Q	167,621	727,562
	0.295***	0.017	√			
	0.570***	0.026		C		
	0.448***	0.023	√			
T=700	0.700	0.998		Q	1,377	3,859
	0.712	0.934	√			
	1.209	1.364		C		
	0.754	1.291	√			
T=1300	0.061	0.546		Q	1,424	14,879
	0.315	0.516	√			
	1.123	0.896		C		
	1.093	0.880	√			
T=2200	-0.041	0.059		Q	38,955	128,666
	-0.001	0.053	√			
	0.135	0.083		C		
	0.108	0.075	√			
T=2700	0.001	0.015		Q	170,695	556,867
	0.000	0.014	√			
	0.011	0.022		C		
	0.004	0.019	√			

Notes: Regression Discontinuity Estimations using two birth weight intervals as bandwidth. Coeff: Coefficient for observations below threshold. Robust SE: robust standard errors. Controls: Using control variables for gender, prematurity status, multiple births, asphyxia, IRDS, ward type, and maternal smoking, alcohol consumption and drug use. Fit: Birth weight functions, fitted separately on each side of threshold; Q: quadratic function, C: cubic function. N_L: Number of observations left/below threshold. N_R: Number of observations right/above threshold. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. **Source:** Own calculations based on the DRG-Statistic.

Table 5: Regression Discontinuity Estimates - Hours of Ventilation, 2005-2011

	Coeff	Robust SE	Controls	Fit	N_L	N_R
T=600	92.055	58.990		Q	3,059	4,588
	32.262	53.507	√			
	207.192**	78.420		C		
	68.949	71.945	√			
T=750	57.307	49.549		Q	4,588	3,461
	-7.859	46.633	√			
	125.418	66.301		C		
	45.766	62.561	√			
T=875	-21.093	37.496		Q	3,461	5,135
	-26.286	34.069	√			
	19.311	50.927		C		
	6.296	45.846	√			
T=1000	45.380*	22.666		Q	5,135	9,698
	23.316	20.580	√			
	63.153	33.057		C		
	31.083	30.131	√			
T=1250	6.704	10.944		Q	9,698	15,044
	5.030	10.229	√			
	24.525	14.292		C		
	20.794	13.438	√			
T=1500	2.626	5.277		Q	15,044	49,707
	-1.298	5.089	√			
	13.002	9.348		C		
	5.340	9.132	√			
T=2000	2.817	1.585		Q	49,707	167,621
	0.669	1.540	√			
	5.794*	2.271		C		
	1.247	2.205	√			
T=2500	1.279**	0.464		Q	167,621	727,562
	0.753	0.454	√			
	2.131**	0.665		C		
	0.844	0.652	√			
T=700	42.339	67.038		Q	1,377	3,859
	47.564	63.567	√			
	95.700	87.146		C		
	70.308	83.431	√			
T=1300	-37.087	24.801		Q	1,424	14,879
	-30.047	23.377	√			
	-12.548	37.701		C		
	-14.675	35.997	√			
T=2200	1.190	1.347		Q	38,955	128,666
	1.492	1.296	√			
	-2.176	1.858		C		
	-2.245	1.789	√			
T=2700	0.232	0.412		Q	170,695	556,867
	0.212	0.407	√			
	-0.919	0.681		C		
	-0.822	0.674	√			

Notes: Regression Discontinuity Estimations using two birth weight intervals as bandwidth. Coeff: Coefficient for observations below threshold. Robust SE: robust standard errors. Controls: Using control variables for gender, prematurity status, multiple births, asphyxia, IRDS, ward type, and maternal smoking, alcohol consumption and drug use. Fit: Birth weight functions, fitted separately on each side of threshold; Q: quadratic function, C: cubic function. N_L: Number of observations left/below threshold. N_R: Number of observations right/above threshold. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. **Source:** Own calculations based on the DRG-Statistic.

Table 6: Regression Discontinuity Estimates - Mortality, 2005-2011

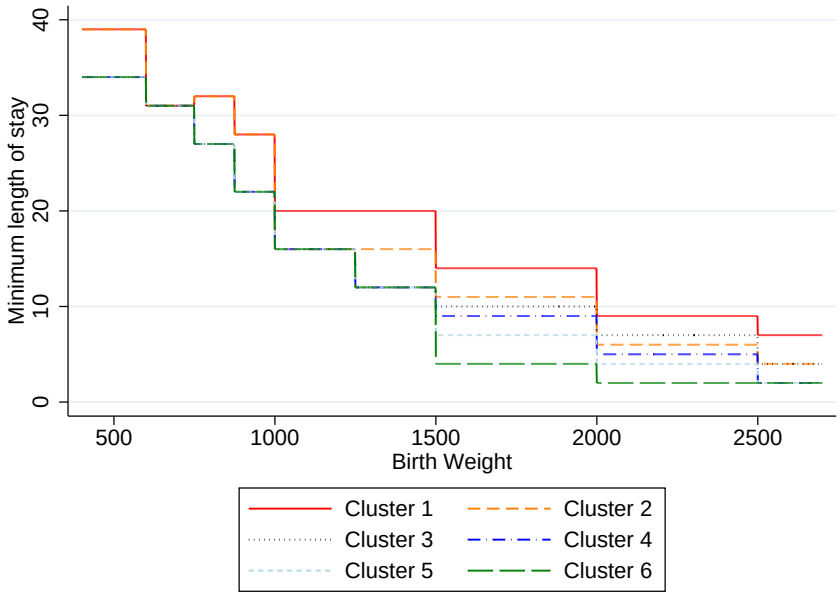
	Coeff	Robust SE	Controls	Fit	N_L	N_R
T=600	-0.074*	0.033		Q	3,059	4,588
	-0.056	0.031	√			
	-0.145**	0.044		C		
	-0.101*	0.041	√			
T=750	-0.118***	0.027		Q	4,588	3,461
	-0.089**	0.026	√			
	-0.162***	0.036		C		
	-0.134***	0.034	√			
T=875	0.025	0.020		Q	3,461	5,135
	0.026	0.019	√			
	0.005	0.026		C		
	0.002	0.026	√			
T=1000	-0.016	0.013		Q	5,135	9,698
	-0.014	0.013	√			
	-0.034	0.020		C		
	-0.031	0.019	√			
T=1250	-0.015	0.008		Q	9,698	15,044
	-0.014	0.008	√			
	-0.028*	0.011		C		
	-0.027*	0.011	√			
T=1500	-0.015***	0.004		Q	15,044	49,707
	-0.016***	0.004	√			
	-0.028***	0.006		C		
	-0.028***	0.006	√			
T=2000	-0.001	0.001		Q	49,707	167,621
	-0.002	0.001	√			
	-0.007***	0.002		C		
	-0.007***	0.002	√			
T=2500	0.001*	0.000		Q	167,621	727,562
	0.001	0.000	√			
	0.000	0.000		C		
	0.000	0.000	√			
T=700	-0.055	0.036		Q	1,377	3,859
	-0.056	0.035	√			
	-0.079	0.047		C		
	-0.061	0.046	√			
T=1300	-0.006	0.016		Q	1,424	14,879
	-0.007	0.015	√			
	-0.064**	0.025		C		
	-0.064**	0.024	√			
T=2200	0.002	0.001		Q	38,955	128,666
	0.002	0.001	√			
	-0.001	0.002		C		
	-0.001	0.002	√			
T=2700	0.000	0.000		Q	170,695	556,867
	0.000	0.000	√			
	-0.001	0.000		C		
	-0.001	0.000	√			

Notes: Regression Discontinuity Estimations using two birth weight intervals as bandwidth. Coeff: Coefficient for observations below threshold. Robust SE: robust standard errors. Controls: Using control variables for gender, prematurity status, multiple births, asphyxia, IRDS, ward type, and maternal smoking, alcohol consumption and drug use. Fit: Birth weight functions, fitted separately on each side of threshold; Q: quadratic function, C: cubic function. N_L: Number of observations left/below threshold. N_R: Number of observations right/above threshold. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. **Source:** Own calculations based on the DRG-Statistic.

Figure 1: Hospital Reimbursement by Birth Weight

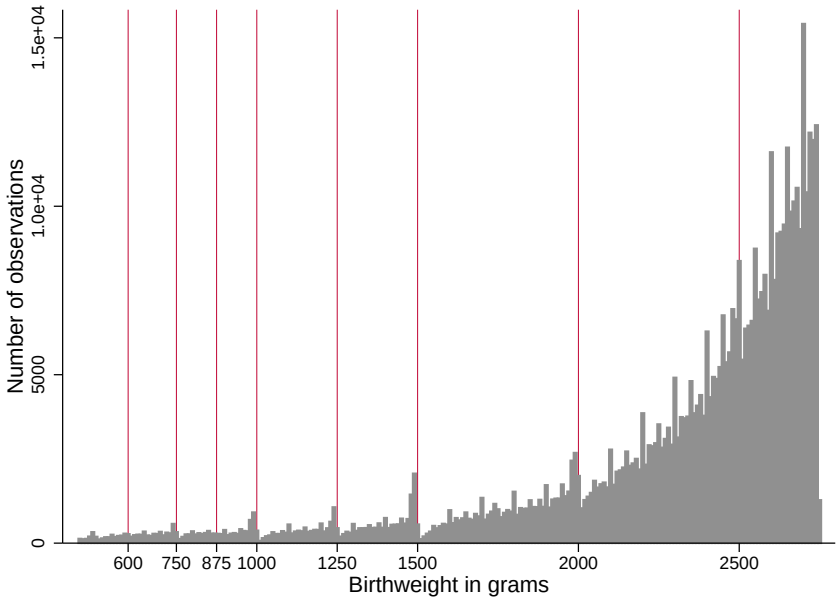
Notes: Development of case-weights for the 6 DRG-clusters defined in tables A.1 and A.2 for the year 2008.

Figure 2: Minimum Length of Stay by Birth Weight



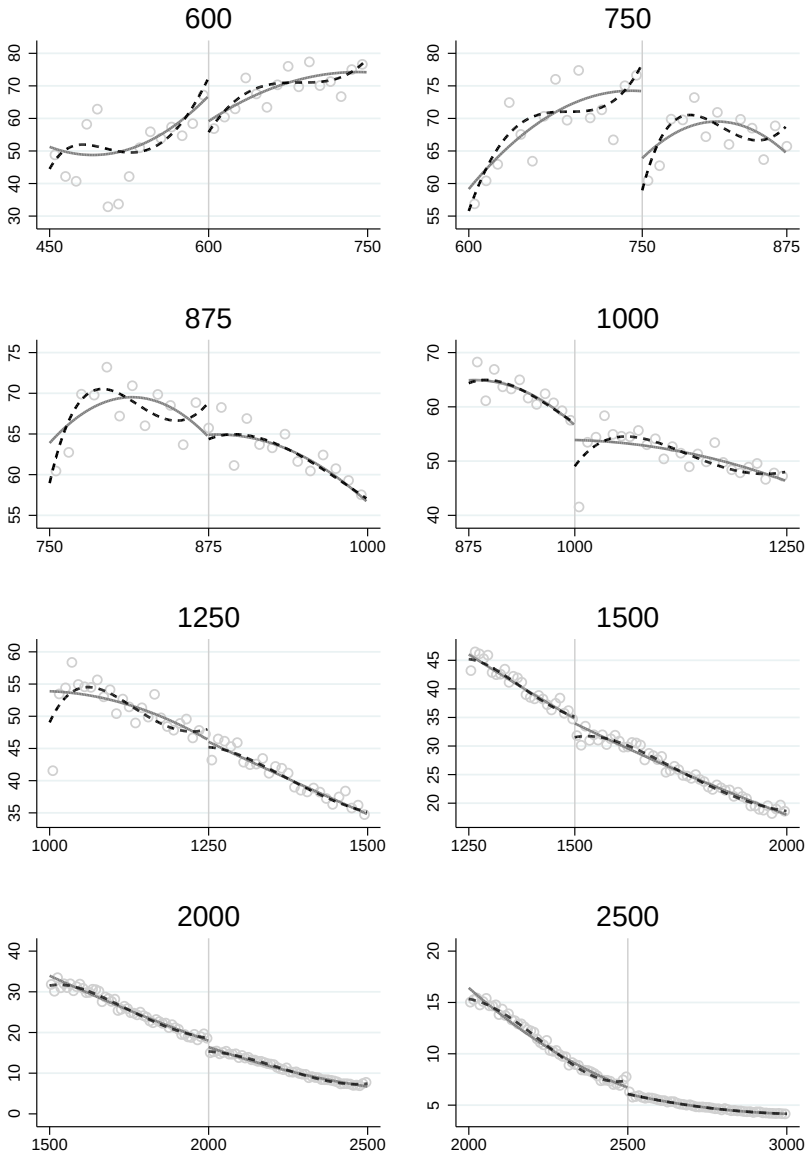
Notes: Development of minimum length of stay for the 6 DRG-clusters defined in tables A.1 and A.2 for the year 2008.

Figure 3: Low Birth Weight Frequencies (Years 2005-2011)



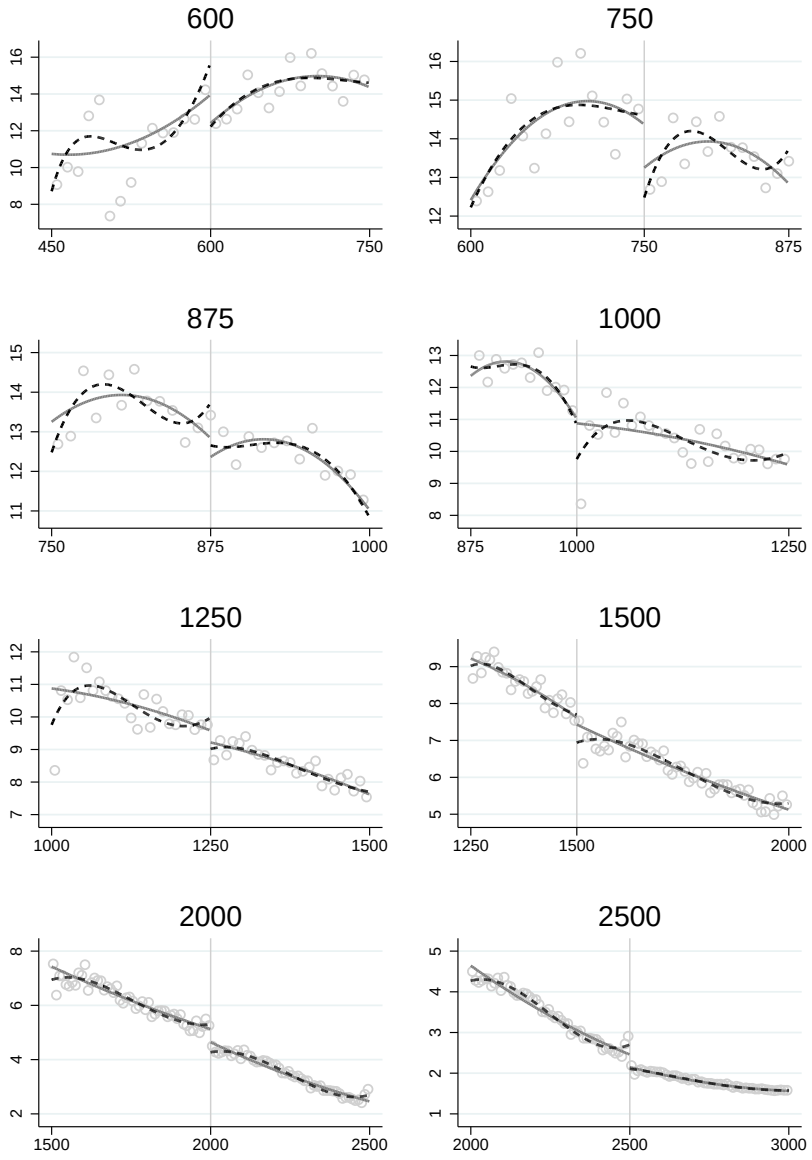
Notes: Distribution of birth weights below 2700 grams in the years 2005-2011. Red lines mark birth weight thresholds in the G-DRG system. All birth weight bins with less than three observations were omitted due to confidentiality. This however does not change the look of the graph. **Source:** Own calculations based on the DRG-Statistic.

Figure 4: Length of Stay



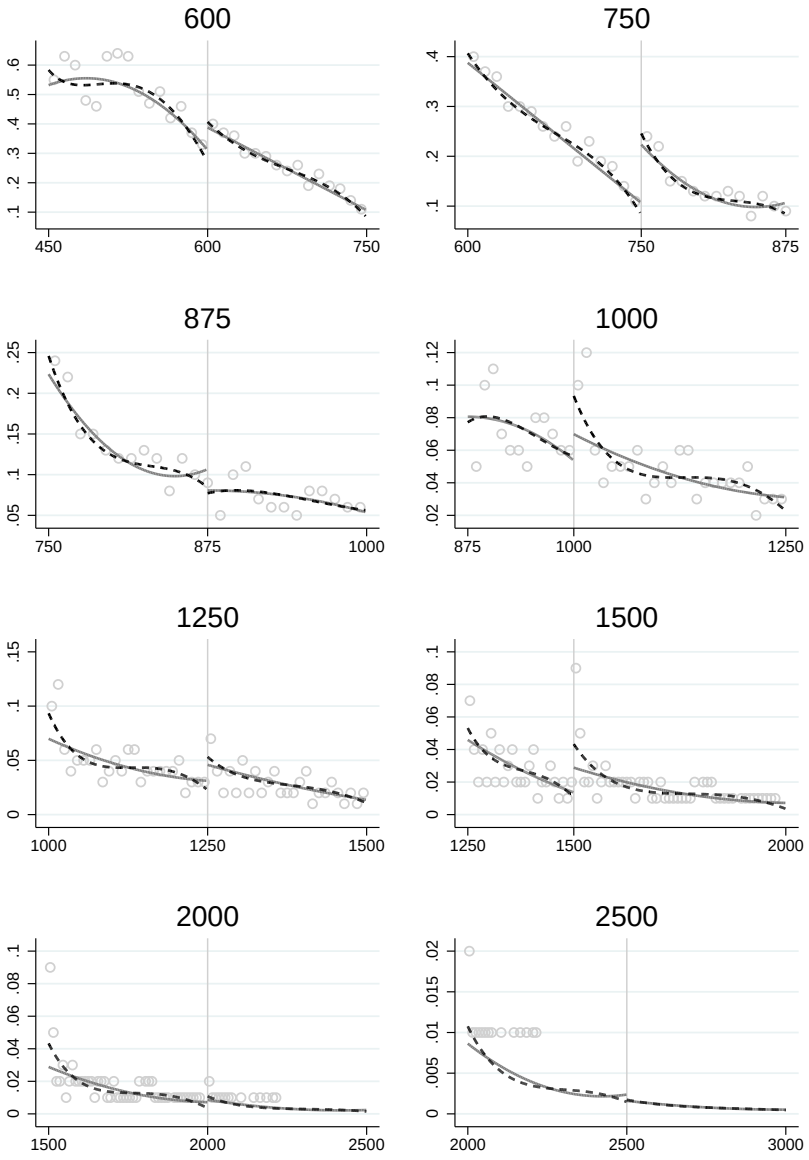
Notes: All birth weights with less than three observations were omitted due to confidentiality. This however does not change the look of the graph. **Source:** Own calculations based on the DRG-Statistic.

Figure 5: Number of Procedures



Notes: All birth weights with less than three observations were omitted due to confidentiality. This however does not change the look of the graph. **Source:** Own calculations based on the DRG-Statistic.

Figure 6: Mortality



Notes: All birth weights with less than three observations were omitted due to confidentiality. This however does not change the look of the graph. **Source:** Own calculations based on the DRG-Statistic.

References

Almond, Douglas, Joseph J Doyle, Amanda E Kowalski and Heidi Williams (2010) "Estimating Marginal Returns To Medical Care: Evidence from at-risk newborns" In: *The Quarterly Journal of Economics* 125 (2), pp. 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" In: *The Quarterly Journal of Economics* 126 (4), pp. 2117–2123.

Bauer, Karl et al. (2006) "Empfehlungen für die strukturellen Voraussetzungen der perinatologischen Versorgung in Deutschland" In: *Zeitschrift für Geburtshilfe und Neonatologie* 210 (01), pp. 19–24.

Busse, Reinhard and Annette Riesberg (2005) "Gesundheitssysteme im Wandel, Deutschland." WHO Regionalbüro für Europa im Auftrag des Europäischen Observatoriums für Gesundheitssysteme und Gesundheitspolitik.

Gemeinsamer Bundesausschuss (2013) "Beschluss über eine Änderung der Vereinbarung über Maßnahmen zur Qualitätssicherung der Versorgung von Früh-und Neugeborenen vom 20 Juni 2013".

Jürges, Hendrik and Juliane Köberlein (2014) "First do no harm. Then do not cheat: DRG upcoding in German neonatology" In: *Schumpeter Discussion Papers* (14001).

Loytved, Christine (2014) "Qualitaetsbericht 2012: Ausserklinische Geburtshilfe in Deutschland".

Schreyögg, Jonas, Reinhard Busse, Matthias Bäuml, Jonas Krämer, Tilman Dette and Alexander Geissler (2014) "Forschungsauftrag zur Mengenentwicklung nach § 17b Abs. 9 KHG - Endbericht". Hamburg Center for Health Economics.

Shigeoka, Hitoshi and Kiyohide Fushimi (2014) "Supplier-induced demand for newborn treatment: Evidence from Japan" In: *Journal of Health Economics* 35, pp. 162–178.

Zimmer, Klaus-Peter (2012) "Neonatology departments under economic pressure" In: *Deutsches Ärzteblatt International* 109 (31-32), p. 517.

Appendix

A DRGs and Reimbursement

Table A.1: DRG-Clusters for Newborns in Germany (2005-2011)

Cluster	Year	<600	600 - 749	750 - 874	875 - 999	1000 - 1249	1250 - 1499	1500 - 1999	2000 - 2499	> 2499
	2005	P61A	P61B	P62A	P62B		P03B	P04B	P05B	P06B
Gruppe 1	**2006-2010**	P61A	P61C	P62A	P62C		P03B	P04B	P05B	P06B
	2011	P61A	P61C	P62A			P03B	P04B	P05B	P06B
	2005	P61A	P61B	P62A	P62B		P03D	P04C	P05C	P06C
Gruppe 2	**2006-2010**	P61A	P61C	P62A	P62C		P03C	P04C	P05C	P06C
	2011	P61A	P61C	P62A			P03C	P04C	P05C	P06C
	2005	P61A	P61B	P62A	P62B	P63Z	P64Z	P65A	P66A	P67A
Gruppe 3	**2006-2010**	P61B	P61D	P62B	P62D	P63Z	P64Z	P65A	P66A	P67A
	2011	P61B	P61D	P62B	P62C	P63Z	P64Z	P65A	P66A	P67A
	2005	P61A	P61B	P62A	P62B	P63Z	P64Z	P65B	P66B	P67B
Gruppe 4	**2006-2010**	P61B	P61D	P62B	P62D	P63Z	P64Z	P65B	P66B	P67B
	2011	P61B	P61D	P62B	P62C	P63Z	P64Z	P65B	P66B	P67B
	2005	P61A	P61B	P62A	P62B	P63Z	P64Z	P65C	P66C	P67C
Gruppe 5	**2006-2010**	P61B	P61D	P62B	P62D	P63Z	P64Z	P65C	P66C	P67C
	2011	P61B	P61D	P62B	P62C	P63Z	P64Z	P65C	P66C	P67C
	2005	P61A	P61B	P62A	P62B	P63Z	P64Z	P65D	P66D	P67D
Gruppe 6	**2006-2010**	P61B	P61D	P62B	P62D	P63Z	P64Z	P65D	P66D	P67D
	2011	P61B	P61D	P62B	P62C	P63Z	P64Z	P65D	P66D	P67D

Notes: DRG-clusters and birth weight thresholds in gram. Within each cluster, birth weight is the only grouping criterion. This table is based on simulations, performed using the G-DRG Webgrouper: `http://drg.uni-muenster.de/index.php?option=com_webgrouper&view=webgrouper&Itemid=26`

Table A.2: Case-weights for selected DRGs 2005-2011

DRG	2005	2006	2007	2008	2009	2010	2011
P01Z	2.668	2.672	3.115	3.662	3.223	3.083	3.651
P02A	16.215	17.452	20.835	21.666	35.132	40.813	45.644
P02B	9.244	10.075	9.186	10.233	16.651	18.013	16.598
P02C					10.024	10.042	11.188
P03A	17.005	16.571	18.494	18.988	19.765	19.534	21.008
P03B	13.176	13.023	12.616	13.782	13.793	12.865	14.084
P03C	11.146	10.943	11.191	10.845	11.368	10.641	11.701
P03D	9.035						
P04A	9.329	9.122	11.871	18.017	18.236	16.186	15.822
P04B	6.338	7.506	8.631	9.253	8.759	8.567	9.183
P04C	5.765	7.338	7.419	7.68	7.488	7.166	8.099
P05A	8.571	8.95	9.482	10.075	9.079	9.287	8.876
P05B	8.003	6.604	6.243	6.564	6.206	6.127	6.125
P05C	4.677	4.722	5.095	4.438	4.543	4.304	5.603
P06A	7.22	7.898	7.858	9.47	9.48	9.083	8.352
P06B	5.367	4.701	5.276	5.059	5.013	4.991	5.014
P06C	2.394	2.575	2.298	2.454	2.055	2.216	2.904
P60A	0.593	0.523	0.499	0.533	0.748	0.869	0.737
P60B	0.312	0.526	0.53	0.576	0.559	0.556	0.536
P60C		0.271	0.23	0.219	0.198	0.169	0.161
P61A	32.531	33.844	31.47	44.193	45.09	39.671	35.119
P61B	24.55	26.111	27.065	32.255	29.685	28.73	28.05
P61C	5.742	28.966	28.644	29.327	33.79	33.521	36.802
P61D		22.582	23.264	25.344	27.876	28.086	25.709
P61E		4.54	6.076	5.847	4.728	5.667	5.367
P62A	20.217	28.428	27.145	28.98	29.305	29.446	32.739
P62B	12.933	17.497	18.992	21.128	19.757	20.433	20.679
P62C	5.532	24.96	22.794	22.831	25.632	25.106	17.427
P62D		13.528	13.177	15.656	16.412	15.51	7.237
P62E		4.993	4.01	5.825	6.096	6.052	
P63Z	9.38	9.163	8.627	9.145	8.808	8.776	9.477
P64Z	6.054	6.682	6.503	6.601	7.086	7.051	7.205
P65A	5.463	5.519	5.182	5.52	5.442	5.164	5.715
P65B	4.185	4.4	4.333	4.448	4.445	4.622	4.749
P65C	3.151	3.173	3.493	3.457	3.321	3.588	3.569
P65D	2.373	1.908	1.68	1.954	1.757	1.673	1.945
P66A	2.87	2.997	3.059	3.425	3.155	3.611	3.552
P66B	2.329	2.441	2.386	2.493	2.575	2.517	2.642
P66C	1.75	1.755	1.849	1.628	1.846	1.886	1.862
P66D	0.506	0.504	0.47	0.411	0.446	0.419	0.349
P67A	2.01	2.045	1.863	1.865	1.814	1.797	1.839
P67B	1.085	1.01	0.989	1.096	1.075	1.079	1.042
P67C	0.641	0.654	0.669	0.64	0.667	0.638	0.647
P67D	0.299	0.287	0.272	0.271	0.27	0.26	0.259

Notes: Tables with all case-weights can be found at http://www.g-drg.de.

B Upcoding in German Neonatologies 2008 - 2012

We follow Jürges and Köberlein (2014), and calculate two measures to quantify upcoding. For this analysis we use a 10% subsample of the DRG-statistic. First, a uniform measure where birth weight in small intervals around a given threshold T is assumed to be uniformly distributed. Under this assumption, the number of observations above and below T should be equal. For distance d, the number of upcoded cases is given by

$$U_{d,T} = 0.5[\sum_{w=T-d}^{w=T+d-1}(N_w)] - [\sum_{w=T-d}^{w=T-1}(N_w)] \tag{5.1}$$

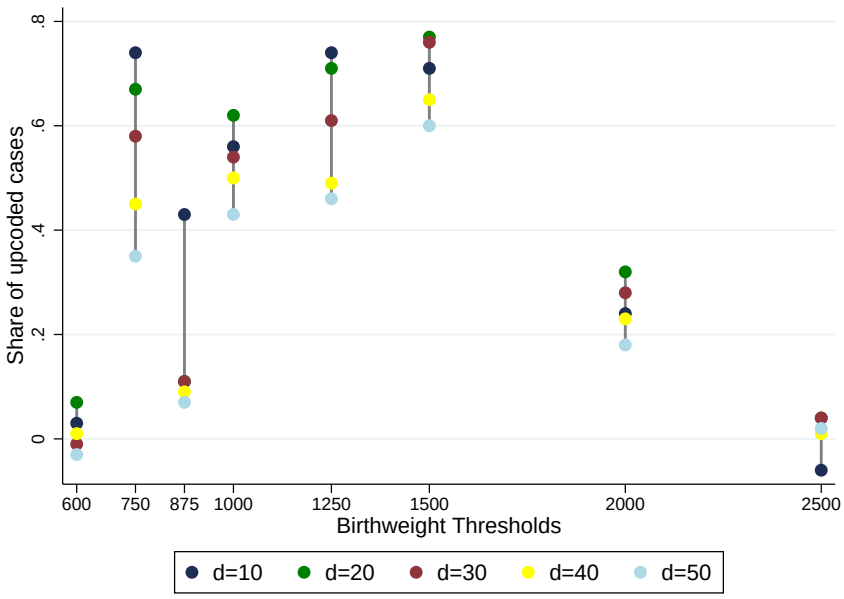
where N_w is the number of observations with birth weight w. The proportion of cases that were upcoded ($\gamma_{d,T}$) from above to below T is then defined as:

$$\gamma_{d,T} = 2[\frac{\sum_{w=T-d}^{w=T-1}(N_w)}{\sum_{w=T-d}^{w=T+d-1}(N_w)}] - 1 \tag{5.2}$$

.

Figure A.1 is a graphical representation of the measure using different bandwidths.

Figure A.1: Uniform Upcoding Measure (Years 2008-2012)



Notes: Uniform upcoding measure at DRG-relevant thresholds using different bandwidths. **Source:** Own calculations based on a 10% sample of the DRG-Statistic, analysed at the Research Data Center of the Bavarian State Office for Statistics.

According to the uniform measure, upcoding is highest at $T = [750, 1000, 1250, 1500]$ where 40-80% of newborns were upcoded. There also appears to be substantial upcoding at 875 and 2000 grams (at least 10%). No upcoding took place at 600 and 2500 grams. Further, $U_{d,T}$ and $\gamma_{d,T}$ are negative at placebo Ts which shows that statistically implausible jumps of birth weight frequencies are indeed only present at DRG relevant Ts. Not surprisingly, the choice of d heavily influences $\gamma_{d,T}$. As a rule of thumb, $\gamma_{d,T}$ decreases with d because obviously stronger manipulation of birth weight is needed to claim excess reimbursement for newborns with birthweights farther away from thresholds.

A clear drawback of the uniform upcoding measure is that actually birth weight is not uniformly distributed around Ts as the number of observations above every T is higher than below. While the uniform measure can be seen as a lower bound for the degree of upcoding, we also employ the second upcoding measure suggested by Jürges and Köberlein (2014). Using a local polynomial regression, the birthweight distribution without jumps is predicted and differences between predicted and actual frequencies are then taken as an indicator for upcoding. The number of upcoded cases $P_{d,T}$ is then obtained by

$$P_{d,T} = \sum_{w=T-d}^{w=T-1} (N_w) - \sum_{w=T-d}^{w=T-1} (\hat{N}_w) \tag{5.3}$$

where $\sum_{w=T-d}^{w=T-1}(\hat{N}_w)$ is the number of predicted observations within distance d from T. From $P_{d,T}$, the share of upcoded cases $\rho_{d,T}$ follows directly as it is

$$\rho_{d,T} = [P_{d,T}]/[\sum_{w=T-d}^{w=T-1} (N_w)] \tag{5.4}$$

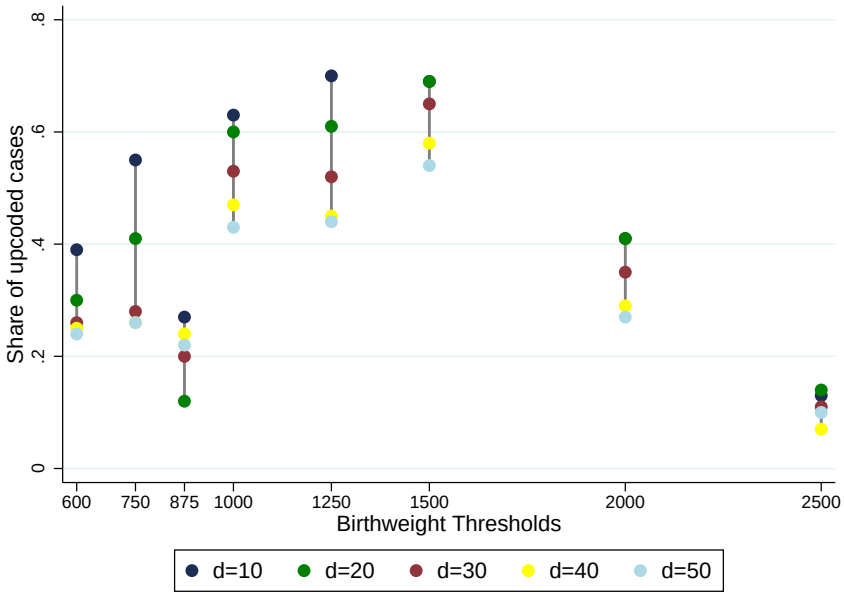
. A graphical illustration of the upcoding measure is given in figure A.2.

The overall upcoding pattern using the polynomial measure is the same as above. However, $P_{d,T}$ and $\rho_{d,T}$ indicate upcoding at two placebo Ts (700 and 1700 grams) of 4% to 18%. As $\rho_{d,T}$ is based on less restrictive distributional assumptions than $\gamma_{d,T}$, it is slightly less sensitive to variations in d. For example the range of upcoding at $T = 875$ narrows to 12-27%. The only striking difference between the two measures is that in contrast to $\gamma_{d,600}$, the polynomial measure $\rho_{d,600}$ indicates that upcoding took place for 20 to 40% of newborns at $T = 600$ in contrast to no upcoding when the uniform measure is used.

We also compare the degree of upcoding by hospital ward. Figure A.3 shows that the bounds for the share of upcoded cases in maternity clinics, and specialised pediatric and neonatal wards overlap for all but the last two thresholds.

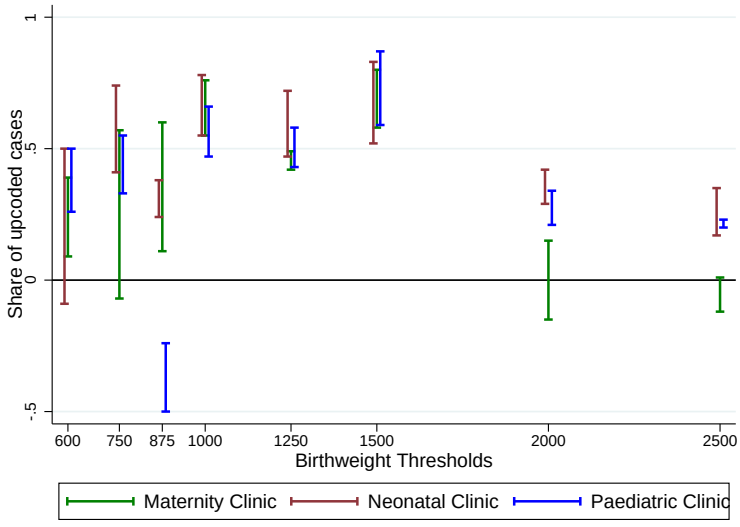
Finally we also analyse whether upcoding takes place rationally in the sense that it

Figure A.2: Polynomial Upcoding Measure (Years 2008-2012)



Notes: Polynomial upcoding measure at DRG-relevant thresholds using different bandwidths. **Source:** Own calculations based on a 10% sample of the DRG-Statistic, analysed at the Research Data Center of the Bavarian State Office for Statistics.

Figure A.3: Share of Upcoded Cases by Clinic Speciality

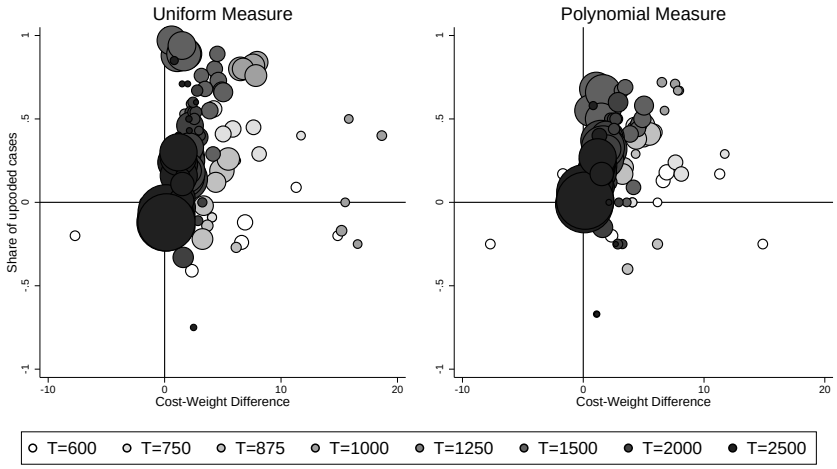


Notes: Upper and lower bounds for upcoding taking maximum and minimum of γ, ρ at d=30 and d=50 per clinic and threshold. **Source:** Own calculations based on a 10% sample of the DRG-Statistic, analysed at the Research Data Center of the Bavarian State Office for Statistics.

takes place more frequently if excess reimbursement is higher. We use the cost-weight differences at each threshold within the severity clusters for the years 2008 - 2012 and

compare this excess reimbursement with the share of upcoded newborns with the respective DRGs. Statistical analysis confirms the visual impression from figure A.4 namely that upcoding is (if at all) only loosely linked to excess reimbursement.

Figure A.4: Share of Upcoded Cases by Cost-Weight Difference



Notes: Upcoded cases using both measures with d=50. Uniform Measure based on 118 observations, Polynomial Measure based on 112 observations (five extreme values for $T = 1000$ from P63Z to P62D were omitted and not enough cases were available to calculate the measure in cluster 2 at 2500 grams in 2009). **Source:** Own calculations based on a 10% sample of the DRG-Statistic, analysed at the Research Data Center of the Bavarian State Office for Statistics. Cost-weights can be found at `http://www.g-drg.de`.