Groß, Mona; Jürges, Hendrik; Wiesen, Daniel

The effects of audits and fines on upcoding in neonatology

Health Economics

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Groß, Mona; Jürges, Hendrik; Wiesen, Daniel (2021) : The effects of audits and fines on upcoding in neonatology, Health Economics, ISSN 1099-1050, Wiley, Hoboken, NJ, Vol. 30, Iss. 8, pp. 1978-1986, https://doi.org/10.1002/hec.4272

Received: 16 April 2020 | Revised: 6 December 2020 | Accepted: 29 March 2021

DOI: 10.1002/hec.4272

HEALTH ECONOMICS LETTER

Health Economics WILEY

The effects of audits and fines on upcoding in neonatology

Mona Groß[1] | **Hendrik Jürges**[2] | **Daniel Wiesen**[1]

[1]Department of Business Administration and Healthcare Management, University of Cologne, Cologne, Germany

[2]Schumpeter School of Business and Economics, University of Wuppertal, Wuppertal, Germany

Correspondence
Hendrik Jürges, Schumpeter School of Business and Economics, University of Wuppertal, Haspeler Str. 27, 42285 Wuppertal, Germany.
Email: juerges@wiwi.uni-wuppertal.de

Daniel Wiesen, Department of Business Administration and Healthcare Management, University of Cologne, Albertus Magnus Platz, 50923 Cologne, Germany.
Email: wiesen@wiso.uni-koeln.de

Funding information
University of Cologne, Germany; University of Wuppertal, Germany
Open access funding enabled and organized by Projekt DEAL.

Abstract

Upcoding is a common type of fraud in healthcare. However, how audit policies need to be designed to cope with upcoding is not well understood. We provide causal evidence on the effect of random audits with different probabilities and financial consequences. Using a controlled laboratory experiment, we mimic the decision situation of obstetrics staff members to report birth weights of neonatal infants. Subjects' payments in the experiment depend on their reported birth weights and follow the German non-linear diagnosis-related group remuneration for neonatal care. Our results show that audits with low detection probabilities only reduce fraudulent birth-weight reporting, when they are coupled with fines for fraudulent reporting. For audit policies with fines, increasing the probability of an audit only effectively enhances honest reporting, when switching from detectable to less gainful undetectable upcoding is not feasible. Implications for audit policies are discussed.

KEYWORDS
audit policies, audits and fines, behavioral experiment, reporting of birth weights, upcoding

JEL CLASSIFICATION
D03, 118

1 | INTRODUCTION

Upcoding of patients to attract higher diagnosis-related group (DRG) payments is a common problem in several healthcare systems, leading to inefficiencies and financial losses (e.g., Barros & Braun, 2017; Bastani et al., 2019; Carter et al., 1990; Dafny, 2005; Januleviciute et al., 2016; Silverman & Skinner, 2004). Incentives to upcode are particularly prevailing in neonatal intensive care (e.g., Hochuli, 2020; Jürges & Köberlein, 2015; Reif et al., 2018; Shigeoka & Fushimi, 2014). The reimbursement for neonatal care is typically determined through birth weights reported by obstetrics staff. DRG payments non-linearly increase with decreasing birth weights at birth-weight thresholds. In Germany, for example, reporting weights just below a threshold may yield additional payments of more than EUR 17,000 (a relative increase of 40%; Jürges & Köberlein, 2015).[1]

To cope with upcoding, policy-makers often intend to increase the frequencies of audits.[2] However, empirical evidence on the effectiveness of such a means is mostly lacking. Only Hennig-Schmidt et al.'s (2019) experiment with neonatal framing shows that a random audit coupled with a fine effectively reduces dishonesty. Similarly, Angerer

et al. (2020) find that non-optimal treatment decreases with more frequent audits in a credence goods experiment.[3] Nevertheless, it is not well understood whether a random audit alone, which upon detection confronts individuals with their dishonesty alluding to self-image concerns (e.g., Bénabou & Tirole, 2006), is sufficient or whether financial consequences are needed to induce honesty.

Using a behavioral experiment, we investigate how (i) coupling random audits with fines and (ii) increasing the audit probability affects upcoding. A controlled laboratory experiment renders control over the decision situation and allows us to observe upcoding which is typically hidden in the field.[4] We thus complement field evidence on upcoding in neonatal care by analyzing *individual* behavior in situations in which upcoding is either detectable or undetectable through audits.

2 | EXPERIMENTAL DESIGN AND PROCEDURE

In a neonatal framing (Hennig-Schmidt et al., 2019), subjects role-play obstetrics staff members charged with the task of entering birth weights $\widehat{w}_j$ of six preterm infants j in birth reports. The birth weights w_j are drawn in random order and shown on subjects' screens: 1200, 1250, 1300, 1350, 1400, and 1500 g. After having seen an infant's weight, subjects are asked to report the birth weight $\widehat{w}_j = [1150; ...; 1550]$ in 50-gram increments.

Subjects receive a fixed lump-sum payment F and variable DRG-based payments $r(\widehat{w}_j^i)$ depending on subject i's *reported* weight per infant. Based on empirical evidence (e.g., Reif et al., 2018), subjects are informed that infants receive optimal care according to their *true* (not the reported) birth weights, which excludes non-financial motivations to upcode. Consequently, treatment costs $c(w_j)$ also depend on *true* weights. Subject i's overall profit is: $\pi^i(w, \widehat{w}) = F + \sum_{j=1}^{6} r(\widehat{w}_j) - c(w_j)$. For an illustration of a decision situation, see Supplementary Appendix A1.

The range of weights in the experiment comprises thresholds at 1250 and 1500 g, following the German DRG-scheme. Payments within a DRG are set such that average treatment costs of an infant are mostly covered. For profits of all combinations of reported and true weights, see Supplementary Appendix A2. We refer to upcoding whenever a birth weight is fraudulently reported to be below a threshold implying a higher DRG-based payment.

Between-subjects, we test the effects of different audit probabilities and of a fine on individuals' reporting behavior; see Table 1. In our baseline (treatment NANF), subjects report birth weights without audits and fines. To investigate audit-effects without fines, a 10%-random audit is introduced in 10ANF. A fine for fraudulent reporting is added in 10AF. If upcoding is detected, all DRG-based payments are withheld and subjects only receive the lump-sum F. In 75AF, the probability of an audit (hence detection and fine) is increased to 75%.[5] We compare reporting behavior between NANF and 10ANF and between 10AF and 75AF to analyze the *effects of audits*. To analyze the *effect of fines* (at low detection probability), we compare treatments 10ANF and 10AF.[6]

The audit mechanism in 10ANF, 10AF, and 75AF relies on comparisons of reported birth weights with infants' weights recorded on their second day of life. Assuming that hospital records show the correct weight on the second day

TABLE 1 Overview on experimental treatments

Detection probability	Financial consequences	
	No fine	**Fine**
0% (no audit)	**NANF** (No-audit-no-fine/Baseline): No random audit, upcoding cannot be detected, no fine ($n = 56$).	—
10%	**10ANF** (10%-audit-no-fine): Random audit of subjects' reported birth weights with 10% probability. If upcoding is detected, subjects are informed about the detection of their fraudulent behavior, but they are not fined ($n = 38$).	**10AF** (10%-audit-and-fine): Random audit is equivalent to 10ANF. If upcoding is detected, subjects are informed about the detection of their fraudulent behavior, and they are fined. That means they only receive the fixed amount F ($n = 65$).
75%	—	**75AF** (75%-audit-and-fine): Random audit of subjects' reported birth weights with 75% probability. If upcoding is detected, subjects are informed about the detection of their fraudulent behavior, and they are fined. That means they only receive the fixed amount F ($n = 38$).

Note: The number of participants per treatment is reported in parentheses.

after birth, and taking into account that newborns lose about 5% of their initial weight within the first 24 h (e.g., Flaherman et al., 2015), the second-day weight cannot be higher than the first-day weight. The opposite would indicate fraudulent reporting. The second-day weight plus 5% thus represents a lower bound for the true birth weight which enables subjects to upcode without the risk of being detected. The analogy for our experiment is that upcoding by 50 g is undetectable and upcoding by more than 100 g is detectable. For true birth weights just above the thresholds (1250 and 1500 g), upcoding to the next DRG-threshold goes undetected. Upcoding 1500 g-infants by two DRG-thresholds is then detectable.[7] In contrast, subjects always face a risk of detection for true birth weights of 1300, 1350, and 1400 g, for which upcoding is to report 1200 g. We are thus able to investigate effects of audits and fines on honest reporting when either (i) only detectable upcoding is possible, (ii) undetectable upcoding is possible, or (iii) detectable and undetectable upcoding are possible. In case a subject is audited and upcoding is detected (for at least one infant), subjects are informed about the detection of fraudulent behavior.

The computerized experiment programmed in z-Tree (Fischbacher, 2007) was conducted via the Cologne Laboratory for Experimental Research in May and June 2019. For the experimental protocol, see Supplementary Appendix A5. For treatments NANF and 10AF, we also include data from the experiments by Hennig-Schmidt et al. (2019) which were conducted at the BonnEconLab in 2014 and 2015. Characteristics and behavior of subjects from Bonn and Cologne did not differ significantly; see Supplementary Appendix B1. Overall, 197 students participated in our experiments, 99 in Cologne and 98 in Bonn; 61% were medical students, the average age was 23 years, and 45% were female.

3 | RESULTS

In all treatments, we observed substantial upcoding of neonatal cases between 37% and 82%. We further differentiate between three types of reporting behavior: detectable upcoding, undetectable upcoding, and honest reporting.[8] Figure 1 shows the average proportions for the reporting types differentiated by treatments. The observed pattern is that, the introduction of a 10%-random audit without fine (10ANF), the introduction of a fine to the 10%-audit (10AF), and the increase in the audit probability from 10% to 75% (75AF), reduced detectable upcoding while undetectable upcoding

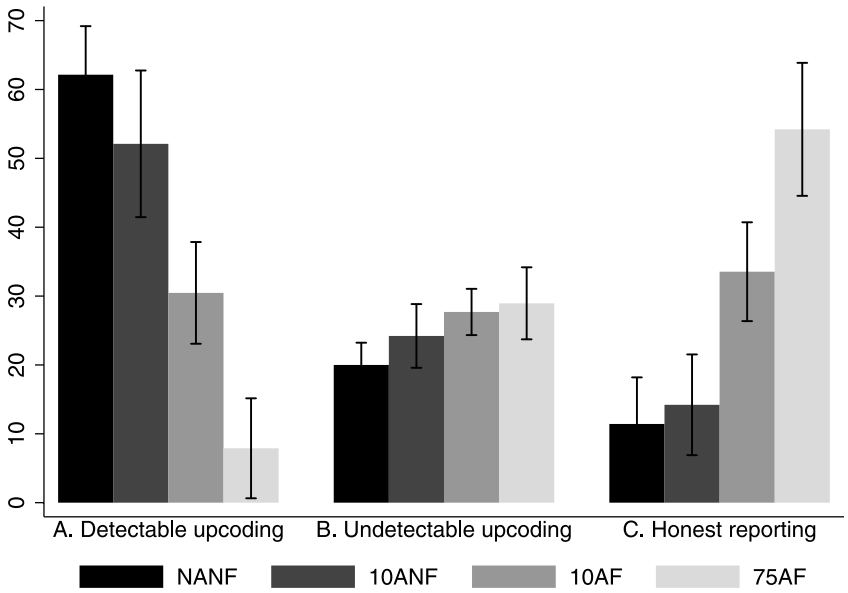

FIGURE 1 **Proportions of reporting behavior by treatments (in %).** The bar charts represent the average shares of upcoding and honest reporting differentiated by treatments, with 95% confidence intervals. Upcoding is defined as payment-increasing misreports of birth weights. Undetectable upcoding is achieved by misreporting birth weights by only 50 g, which cannot be detected by an audit (Panel A). Detectable upcoding implies misreporting birth weights by 100 g or more, which can always be detected by audit (Panel B). Honest reporting refers to reporting the true birth weight (Panel C). We included all birth weights except 1200 g where (payment-increasing) upcoding is not possible ($k = 985$ decisions). NANF: No-audit-no-fine, 10ANF: 10%-audit-no-fine, 10AF: 10%-audit-and-fine, 75AF: 75%-audit-and-fine

increased. Honest reporting increased at a small rate for 10ANF and more substantially for audit policies with fines (10AF, 75AF).

We next separately analyze individuals' honest reporting as opposed to situations in which either only detectable or undetectable upcoding is feasible. We apply logit regression models to analyze individuals' behavior controlling for subjects' characteristics: age, gender, medical major, personality traits (Big-Five Inventory; Rammstedt & John, 2007), and integrity (Schlenker, 2008).[9] Table 2 reports predictive margins for honest birth-weight reporting differentiated by treatments. Treatment differences expressed as marginal effects at the means reflect the effects of audits and fines on the likelihood of honest reporting in percentage points (pp).

First, we analyze *honest reporting* only for true birth weights, at which upcoding is always detectable (1300, 1350, and 1400 g); see Panel A of Table 2. Introducing a 10%-audit without a fine slightly increased honest reporting by about 7 pp ($p = 0.248$). Adding a fine to the 10%-random audit significantly increased honest reporting by about 33 pp ($p < 0.001$). An increase in audit probability from 10% to 75% (with fines) increased honest reporting by roughly 21 pp ($p = 0.015$).

TABLE 2 Predictive margins from logit regressions on differences in honest reporting between experimental treatments

A. Honest reporting when only detectable upcoding is possible (at 1300, 1350, and 1400 g; $k = 591$ decisions)			
Detection prob.	**No fine**	**Fine**	**△, in pp (*p*-value)**
0% (no audit)	12.4% (NANF)	—	
10%	19.8% (10ANF)	52.4% (10AF)	32.6 (<0.001)
75%	—	73.8% (75AF)	
△, in pp (*p*-value)	7.4 (0.248)	21.4 (0.015)	
B. Honest reporting when undetectable upcoding is possible (at 1250 g; $k = 197$ decisions)			
Detection prob.	**No fine**	**Fine**	**△, in pp (*p*-value)**
0% (no audit)	7.4% (NANF)	—	
10%	1.8% (10ANF)	5.7% (10AF)	3.9 (0.307)
75%	—	21.7% (75AF)	
△, in pp (*p*-value)	−5.6 (0.194)	16.0 (0.070)	
C. Honest reporting when undetectable and detectable upcoding are feasible (at 1500 g; $k = 196$ decisions)			
Detection prob.	**No fine**	**Fine**	**△, in pp (*p*-value)**
0% (no audit)	4.3% (NANF)	—	
10%	3.5% (10ANF)	16.3% (10AF)	12.8 (0.023)
75%	—	17.4% (75AF)	
△, in pp (*p*-value)	−0.8 (0.837)	1.2 (0.887)	

Note: This table reports predictive margins at the means of the covariates based on logit models on honest reporting for Panels A and B and based on a multinomial logit model on reporting behavior at 1500 g for Panel C. Upcoding is defined as payment-increasing misreports of birth weights. Panel A estimates honest reporting for infants for whom upcoding is always detectable in case of an audit (1300, 1350, and 1400 g; $k = 591$ decisions). Panel B estimates honest reporting for infants of 1250 g for whom undetectable upcoding is possible ($k = 197$ decisions). Undetectable upcoding means that upcoding is achieved by misreports of birth weights by only 50 g, which cannot be detected by an audit. Panel C estimates honest reporting for infants of 1500 g for whom undetectable upcoding is possible but less gainful than detectable upcoding ($k = 196$ decisions). All predictive margins are adjusted for individual characteristics, that is, gender, age, medical major, personality traits (Big-Five Inventory; Rammstedt & John, 2007), and integrity (Schlenker, 2008). Treatment effects (△) are differences in marginal effects at the means (percentage points, pp). Full regression results are reported in the Appendix, Table B6 and Table B7. NANF, No-audit-no-fine; 10ANF: 10%-audit-no-fine, 10AF: 10%-audit-and-fine, 75AF: 75%-audit-and-fine.

Second, we consider the effects of audits and fines on honest reporting at 1250 g, the true birth weight at which upcoding is undetectable; see Panel B of Table 2.[10] Introducing a 10%-audit led to a decrease in honest reporting by about 6 pp ($p = 0.194$). Adding a fine to an audit with a low detection probability, increased honest reporting by about 4 pp ($p = 0.307$). Raising the audit probability further increased honest reporting by about 16 pp ($p = 0.070$).

Finally, we analyze individuals' honest reporting for the true birth weight of 1500 g. At this weight, subjects have the opportunity to choose between detectable and undetectable upcoding where the former yields a higher gain; see Panel C of Table 2. Introducing a 10%-audit did not significantly affect honest reporting ($p = 0.954$). Adding a fine to it, however, significantly increased honest reporting by about 13 pp ($p = 0.023$). Raising the detection probability of a random audit to 75% did also not significantly affect honesty (1 pp, $p = 0.887$).

While rational choice options at the previously considered true birth weights are binary, subjects have three choice options at 1500 g: honest reporting, *detectable* upcoding (by two DRG-thresholds) and *undetectable* upcoding (by one threshold). Table 3 reports predictive margins based on multinominal logit regressions for detectable and undetectable upcoding differentiated by treatments. Treatment effects are expressed as marginal effects at the means in percentage points (pp). We find that introducing a 10%-random audit decreased detectable upcoding by about 18 pp ($p = 0.089$). At the same time, however, undetectable upcoding increased by about 19 pp ($p = 0.063$). Obviously, honest reporting was hardly affected. Adding a fine to the 10%-audit significantly decreased detectable upcoding by about 33 pp ($p = 0.002$), and in parallel increased undetectable upcoding by about 20 pp ($p = 0.065$). With a fine, an increase in audit probabilities from 10% to 75% reduced detectable upcoding by about 17 pp ($p = 0.022$) and increased undetectable upcoding by about 16 pp ($p = 0.113$). The results reveal an unintended consequence of audits in that they shifted detectable to undetectable upcoding rather than triggering more honest reporting.

In sum, our analyses indicate that, first, random audits with low detection probabilities only reduced upcoding and fostered honest reporting if audits comprise fines. This emphasizes the importance of a financial consequence to cope with dishonest behavior and complements findings from Hennig-Schmidt et al. (2019). Second, raising the probability of an audit increased honest reporting in the decision situations when either only detectable or only undetectable upcoding was possible. As a separate analysis of the true birth weights of 1500 g indicated reductions in detectable upcoding were accompanied by increases in undetectable upcoding.

TABLE 3 Predictive margins from a multinomial logit regression on differences in detectable and undetectable upcoding at 1500 g

A. Detectable upcoding (at 1500 g; $k = 196$ decisions)			
Detection prob.	**No fine**	**Fine**	**△, in pp (*p*-value)**
0% (no audit)	76.2% (NANF)	—	
10%	57.8% (10ANF)	24.9% (10AF)	−32.9 (0.002)
75%	—	7.8% (75AF)	
△, in pp (*p*-value)	−18.4 (0.089)	−17.1 (0.022)	
B. Undetectable upcoding (at 1500 g; $k = 196$ decisions)			
Detection prob.	**No fine**	**Fine**	**△, in pp (*p*-value)**
0% (no audit)	19.5% (NANF)	—	
10%	38.7% (10ANF)	58.8% (10AF)	20.2 (0.065)
75%	—	74.8% (75AF)	
△, in pp (*p*-value)	19.2 (0.063)	16.0 (0.113)	

Note: This table reports predictive margins at the means of the covariates based on a multinomial logit model on upcoding at 1500 g ($k = 196$ decisions). Upcoding is defined as payment-increasing misreports of birth weights. All predictive margins are adjusted for individual characteristics, that is, gender, age, medical major, personality traits (Big-Five Inventory; Rammstedt & John, 2007), and integrity (Schlenker, 2008). "Detectable upcoding" takes place when an individual reports a weight of 1250 g or lower, "Undetectable upcoding" takes place when an individual reports a (fraudulent birth) weight of 1450 g. Estimates for "Honest reporting" are reported in Table 3. We excluded one subject with a birthweight entry of 1550 g. Treatment effects (△) are differences in marginal effects at the means (percentage points, pp). Full regression results are reported in the Appendix, Table B7. NANF, No-audit-no-fine; 10ANF, 10%-audit-no-fine; 10AF, 10%-audit-and-fine; 75AF, 75%-audit-and-fine.

4 | DISCUSSION AND CONCLUSION

Our behavioral experiment provides important insights for healthcare policy-makers on the effects of different audit policies on upcoding in neonatology. First, random audits at low detection probability without financial consequences are not sufficient to foster honest reporting. Only when random audits include fining of fraudulent reporting, honesty increased significantly. A fine seems thus to be an essential instrument carrying a signal that fraudulent reporting of birth weights represents misbehavior and is sanctioned. Second, increasing the frequency of random audits only induces more honesty when individuals are not able to shift from detectable to undetectable upcoding. Hence, differentiating between detectable and undetectable upcoding reveals the unintended consequence of audit policies to foster more undetectable upcoding rather than honest reporting.

When interpreting behavioral consequences for the remuneration within the confines of the experiment, upcoding led to high financial losses for payers. Without an audit, the average payment per infant almost doubled compared to a theoretical payment under fully honest reporting. Introducing a 10%-random audit (without a fine) reduced the financial loss for the insurer by 10%.[11] Only when subjects bore the risk of being fined for fraudulent reporting,

TABLE 4 Overview of mean DRG remunerations per infant by experimental treatment

Treatment	Mean remuneration per infant (in Taler)		Loss due to upcoding (in Taler)	Reduction in loss compared to NANF (in %)
	If fully honest	Observed behavior		
A. At all birth weights; $k = 985$ decisions				
NANF	184	343	159	—
10ANF	184	327	143	−10.1
10AF	184	286	102	−35.8
75AF	184	238	54	−66.0
B. At 1300, 1350, and 1400 g; when only detectable upcoding is possible; $k = 591$ decisions				
NANF	200	344	144	—
10ANF	200	323	123	−14.6
10AF	200	273	73	−49.3
75AF	200	219	19	−86.8
C. At 1250 g; when undetectable upcoding is possible; $k = 197$ decisions				
NANF	200	364	164	—
10ANF	200	371	171	+4.3
10AF	200	369	169	+3.0
75AF	200	337	137	−16.5
D. At 1500 g; when undetecable is possible but less gainful than detectable upcoding; $k = 197$ decisions				
NANF	120	320	200	—
10ANF	120	295	175	−12.5
10AF	120	243	123	−38.5
75AF	120	197	77	−61.5

Note: This table reports the average diagnosis-related group (DRG) remuneration the insurer has to pay per infant. Upcoding is defined as payment-increasing misreports of birth weights. In Panel A, we only consider the infants for whom upcoding is possible (birth weight of 1200 g is excluded for our calculations). In Panel B, we only consider the infants for whom upcoding only detectable upcoding is possible (1300, 1350, and 1400 g). In Panel C, we only consider the infant with birth weight of 1200 g for whom undetectable upcoding is possible. In Panel D, we only consider the infant with birth weight of 1500 g for whom undetectable upcoding is possible but less gainful detectable upcoding. Under full honest reporting, we report the average hypothetical remuneration for the true birth weight of the respective infants. Observed behavior refers to our behavioral data differentiated by treatments. We have calculated the mean remuneration of every subject per infant based on the reported birth weights. We report the financial loss for the insurer due to DRG upcoding as the difference between fully honest reporting and our behavioral data. In the last column, we calculate the relative differences of financial loss between the respective audit and our baseline treatment. All monetary amounts are given in Taler, our experimental currency, the exchange rate being 1 Taler = 0.01 EUR. NANF, No-audit-no-fine; 10ANF, 10%-audit-no-fine; 10AF, 10%-audit-and-fine; 75AF, 75%-audit-and-fine.

however, there was a noticeable drop. When a 10% or 75%-audit came with a fine, the financial loss declined by 36% or 66% compared to no audit; see Panel A of Table 4. The effects are more pronounced focusing on infants for whom only detectable upcoding is possible. Introducing a 10%-audit reduced the financial loss by 15%, a 10%-audit and fine by 49%, and a 75%-audit and fine by 87%; see Panel B of Table 4.

We now separately consider remuneration effects per infant at 1250 and 1500 g for whom undetectable upcoding was possible. While treatment comparisons reveal that the insurer's financial loss due to DRG upcoding increased by 4% (10ANF) and by 3% (10AF) for infants at 1250 g, it can be reduced by 17% only when audits occurred with high detection probability and a fine (75AF); see Panel C of Table 4. However, the insurer's loss can be reduced by 13% (10ANF), by 39% (10ANF), and by 62% (75AF) for infants with true birth weights of 1500 g; see Panel D of Table 4. This decline can be explained by upcoding infants with true birth weights of 1500 g by one instead of two DRG thresholds. Thus, switching from detectable to less gainful undetectable upcoding reduced the payments to some extent.

The results should be interpreted within the confines of our experimental setup. The implementation of audit policies comes at costs, such as set-up costs for monitoring programs, personnel costs, and costs for potential false positive findings which need to be weighed against the savings in expenditures due to less upcoding. Further, beyond the financial savings which can be realized through changing individuals' reporting behavior when introducing audits, fines for detected fraudulent behavior under audits help to reduce the losses insurers face due to upcoding. When considering a real-world health setting, it remains unclear whether the detection of dishonest behavior would imply psychological costs, for example derived by concerns for social reputation or self-respect (e.g., Bénabou & Tirole, 2006) and observed lying aversion (e.g., Dufwenberg & Dufwenberg, 2018; Gneezy et al., 2018), which could vary between the lab and the field and might thus lead to different kinds of upcoding behavior. While evaluating costs and benefits is at the discretion of health policy-makers, our findings at least provide some directional guidance for the ongoing debates on the design and implementation of audit policies. Our experimental study suggests interesting paths for future research; for example, the analysis of reporting behavior under a high detection probability without financial consequences in case of detection, and situations in which upcoding does not yield financial gains. In these ways, the preferences of individuals for dishonesty could be investigated further.

In sum, our results suggest that audits with fines can, on the aggregate, reduce upcoding while not necessarily inducing more honesty. Audits might still decrease dishonesty by pushing dishonest individuals into reporting fraudulently to an extent that is not detectable. This calls for a design of audit policies that makes the detection of dishonest behavior more likely, for example, through audit mechanisms that reduce measurement errors.

ACKNOWLEDGMENTS

We are grateful for valuable comments and suggestions by Matteo M. Galizzi, Heike Hennig-Schmidt, Tor Iversen, Ludwig Kuntz, and Christian Waibel as well as conference and seminar participants at the BEHnet Workshop Innsbruck and the University of Cologne. We also thank Leonie Offergeld, Anna Hanel, Julie Damm, and Helena Müller for their excellent research assistance and help in conducting the experiments. Financial support from the University of Cologne and the University of Wuppertal is gratefully acknowledged.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data available in article supplementary material.

ORCID

Mona Groß https://orcid.org/0000-0002-3259-7839
Daniel Wiesen https://orcid.org/0000-0003-4627-1730

ENDNOTES

[1] According to the German DRG schedule, reported weights that fall into the category of 750 to 874 g instead of 875 to 999 g, increases the average reimbursement by EUR 17,555, from EUR 45,985 to EUR 63,540; see Jürges and Köberlein (2015) for further details.

[2] Increasing audit frequencies for a given fine level would make dishonest behavior less attractive (e.g., Becker, 1968). In a Beckerian sense, a utility-maximizing decision-maker weighs the expected utilities from dishonest and honest behavior. This logic motivated health-policy

reforms that were recently implemented after much debate. For example, in Germany a recent reform of the Medical Service of the Sickness Funds (MDK) came into effect in early 2020. It intends to sanction fraudulent reporting of hospital bills. In detail, depending on the overall share of unobjected (or correctly billed) hospital invoices, hospitals must pay a fine to the health insurance funds in addition to the repayment of the difference between the true and the wrongly billed amount. In 2020, the fine amounts to 10% of the difference, but is never less than EUR 300. From 2021, the size of the fine will differ depending on the overall share of contested bills in 2020 (*Gesetz für bessere und abhängige Prüfungen (MDK-Reformgesetz) 2019*, Art. 1, §275c). In the past, German hospitals did not face any consequences beyond the repayment of the falsely billed amounts. Moreover, hospitals have received, and will continue to receive, EUR 300 for every audited invoice which has been correctly billed as a lump-sum expense allowance from the health insurance fund.

[3] For a definition of credence goods, see, for example, Dulleck and Kerschbamer (2006) and for excellent surveys of the literature, see, for example, Kerschbamer and Sutter (2017) and Balafoutas and Kerschbamer (2020).

[4] For a definition of behavioral experiment in health and more on the discussion of the use of experiments in health economics, see Galizzi and Wiesen (2017, 2018).

[5] While the fine seems substantial under 10AF, a risk-neutral profit-maximizer would still upcode weights in all decisions. An audit probability of 75% represents the cut-off value which (assuming common knowledge about randomly drawn birth weights) implies that a profit-maximizer would only engage in undetectable upcoding which cannot be detected by audits.

[6] For a detailed description of the decomposition of the effects of audits and fines, see Supplementary Appendix A.3.

[7] In detail, when $w_j = 1250$ g, reporting a birth weight of $\widehat{w}_j = 1200$ g cannot be detected by an audit, as the reported birth weight is higher than the lower bound of the true weight of 1, 187.5g on day two. The other possibility of undetectable upcoding is represented by $w_j = 1500$ g (lower bound of the true weight at the second weighing: 1425 g). Here, reporting $\widehat{w}_j = 1450$ g cannot be detected by an audit, whereas reporting $w_j = 1200$ g can. For a graphical explanation of (detectable and undetectable) upcoding, see Supplementary Figure A.1 in the Appendix A.4.

[8] The vast majority of decisions falls in one of the three categories. Only 8% of reported birth weights reports per treatment deviate from these classifications of behavior and are categorized as unclassified. The proportions of unclassified birth weight reports did not vary systematically with the treatment; see Supplementary Table B.3 in Appendix B.3. For the frequencies of individuals' choices, see Supplementary Table B.8 to B.11 in Appendix B.4.

[9] For descriptive statistics on individual characteristics by treatments, see Supplementary Table B.2 in Appendix B.2. Descriptive statistics on proportions in upcoding behavior and reported marginal effects based on regressions without individual controls yield very similar results on the effects of an audit and fine; see Supplementary Table B.4 to B.7 in the Appendix B.3.

[10] Note that at 1250 g, reporting 1150 g would be detectable but it yields no financial gain compared to reporting 1200 g (and remaining undetectable). Participants who choose 1150 g make an inferior decision and we thus classify it "other" reporting behavior.

[11] Note that the reduction in financial loss only refers to the reduced payments the insurer has to pay per infant based on observed reporting behaviors. Potential fines which hospitals have to pay if they are found out for misreporting birth weights are not considered in the calculation and would even lead to higher reductions in financial losses.

REFERENCES

Angerer, S., Rützler, D., & Waibel, C. (2020). Monitoring institutions in health care markets: Experimental evidence. *SSRN Journal*. https://doi.org/10.2139/ssrn.3372994

Balafoutas, L., & Kerschbamer, R. (2020). Credence goods in the literature: What the past fifteen years have taught us about fraud, incentives, and the role of institutions. *Journal of Behavioral and Experimental Finance*, *26*, 100285. https://doi.org/10.1016/j.jbef.2020.100285

Barros, P., & Braun, G. (2017). Upcoding in a national health service: The evidence from Portugal. *Health Economics*, *26*, 600–618. https://doi.org/10.1002/hec.3335

Bastani, H., Goh, J., & Bayati, M. (2019). Evidence of upcoding in pay-for-performance programs. *Management Science*, *65*, 1042–1060. https://doi.org/10.1287/mnsc.2017.2996

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, *76*, 169–217.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*, 1652–1678. https://doi.org/10.1257/aer.96.5.1652

Carter, G. M., Newhouse, J. P., & Relles, D. A. (1990). How much change in the case mix index is DRG creep?. *Journal of Health Economics*, *9*, 411–428. https://doi.org/10.1016/01676296(90)90003l

Dafny, L. S. (2005). How do hospitals respond to price changes? *American Economic Review*, *95*, 1525–1547. https://doi.org/10.1257/000282805775014236

Dufwenberg, M., & Dufwenberg, M. A. (2018). Lies in disguise–a theoretical analysis of cheating. *Journal of Economic Theory*, *175*, 248–264. https://doi.org/10.1016/j.jet.2018.01.013

Dulleck, U., & Kerschbamer, R. (2006). On doctors, mechanics, and computer specialists: The economics of credence goods. *Journal of Economic Literature*, *44*, 5–42. https://doi.org/10.1257/002205106776162717

Fischbacher, U. (2007). Z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*, 171–178. https://doi.org/10.1007/s1068300691594

Flaherman, V. J., Schaefer, E. W., Kuzniewicz, M. W., Li, S. X., Walsh, E. M., & Paul, I. M. (2015). Early weight loss nomograms for exclusively breastfed newborns. *Pediatrics*, *135*, 16–23. https://doi.org/10.1542/peds.20141532

Galizzi, M. M., & Wiesen, D. (2017). Behavioural experiments in health: An introduction. *Health Economics*, *26*, 3–5. https://doi.org/10.1002/hec.3629

Galizzi, M. M., & Wiesen, D. (2018). Behavioral experiments in health economics. In J. Hamilton (Ed.), *Oxford research encyclopedia of economics and finance*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190625979.013.244

Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, *108*, 419–453. https://doi.org/10.1257/aer.20161553

Hennig-Schmidt, H., Jürges, H., & Wiesen, D. (2019). Dishonesty in health care practice: A behavioral experiment on upcoding in neonatology. *Health Economics*, *28*, 319–338. https://doi.org/10.1002/hec.3842

Hochuli, P. (2020). Losing body weight for money: How provider-side financial incentives cause weight loss in Swiss low-birth-weight newborns. *Health Economics*, *29*, 406–418. https://doi.org/10.1002/hec.3991

Januleviciute, J., Askildsen, J. E., Kaarboe, O., Siciliani, L., & Sutton, M. (2016). How do hospitals respond to price changes? evidence from Norway. *Health Economics*, *25*, 620–636. https://doi.org/10.1002/hec.3179

Jürges, H., & Köberlein, J. (2015). What explains DRG upcoding in neonatology? The roles of financial incentives and infant health. *Journal of Health Economics*, *43*, 13–26. https://doi.org/10.1016/j.jhealeco.2015.06.001

Kerschbamer, R., & Sutter, M. (2017). The economics of credence goods–a survey of recent lab and field experiments. *CESifo Economic Studies*, *63*, 1–23. https://doi.org/10.1093/cesifo/ifx001

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. *Journal of Research in Personality*, *41*, 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Reif, S., Wichert, S., & Wuppermann, A. (2018). Is it good to be too light? Birth weight thresholds in hospital reimbursement systems. *Journal of Health Economics*, *59*, 1–25. https://doi.org/10.1016/j.jhealeco.2018.01.007

Schlenker, B. R. (2008). Integrity and character: Implications of principled and expedient ethical ideologies. *Journal of Social and Clinical Psychology*, *27*, 1078–1125. https://doi.org/10.1521/jscp.2008.27.10.1078

Shigeoka, H., & Fushimi, K. (2014). Supplier-induced demand for newborn treatment: Evidence from Japan. *Journal of Health Economics*, *35*, 162–178. https://doi.org/10.1016/j.jhealeco.2014.03.003

Silverman, E., & Skinner, J. (2004). Medicare upcoding and hospital ownership. *Journal of Health Economics*, *23*, 369–389. https://doi.org/10.1016/j.jhealeco.2003.09.007

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Groß, M., Jürges, H., & Wiesen, D. (2021). The effects of audits and fines on upcoding in neonatology. *Health Economics*, *30*(8), 1978–1986. https://doi.org/10.1002/hec.4272