ESSAYS ON INCENTIVES, MORAL HAZARD AND OPTIMAL POLICY

JAKOB SCHLOCKERMANN

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF ECONOMICS
ADVISOR: ALEXANDRE MAS

JUNE 2018
Abstract:

This collection of essays studies incentives and public policy in health care and labor markets. Chapter 1 investigates the causal effect of financial incentives on hospital discharge decisions. Making use of a kink in the German reimbursement system for hospitals, I employ a bunching design to tightly bound the effect of marginal pay on patients’ length of stay around zero. I discuss how the contrast between my result and the previous evidence is explained by differences in the institutional setting. In Chapter 2, I provide new evidence which improves our understanding of provider as well as patient behavior in response to incentives in health care. In particular, I show that a well-known piece of evidence in favor of moral hazard on the patient side with respect to health insurance coverage is better interpreted as limited attention behavior by doctors. Chapter 3 turns towards labor markets. It studies the impacts of heterogeneity in earnings potential on the optimal design of unemployment insurance and estimates the newly relevant parameters empirically.
Acknowledgments:

I am indebted to Alex Mas for his guidance and support throughout my whole PhD experience and, especially, for pushing me a little in the right moments. I am also very grateful to Ilyana Kuziemko and David Silver for their help and comments which have improved my projects substantially. My work also benefitted from comments by David Arnold, Janet Currie, Marc Fleurbaey, Mike Golosov, Felipe Goncalves, Daniel Herbst, Bo Honoré, Amanda Kowalski, Andrew Langan, Steve Mello, Steve Woodbury and David Zhang. Moreover, I thank the many participants of the Princeton Labor Lunch, the Princeton Public Finance Lunch, the Princeton CHW lunch, the Princeton Labor Seminar, the 2017/2018 RES PhD Meetings and the 2017/2018 SAEe for their helpful comments. Special thanks to Melanie Scheller from the German Federal Statistical Agency for her patient help with accessing the data. The Princeton University Industrial Relations Section provided generous financial support. Any errors are my own.
Contents

Abstract ................................................................. iii
Acknowledgments ....................................................... iv
List of Tables .......................................................... viii
List of Figures .......................................................... x

1 Do Provider Incentives Always Affect Health Care Costs? New Evidence from Germany 1
  1.1 Introduction ....................................................... 1
  1.2 Institutional Background ........................................ 6
  1.3 Theory ............................................................. 13
  1.4 Data and Sample Selection ...................................... 19
  1.5 Results ............................................................ 21
  1.6 Discussion ......................................................... 31
References ............................................................... 39
  1.A Measurement of Length of Stay ............................... 62
  1.B Generalized Model ............................................... 63
  1.C Further Results .................................................. 70
  1.D Bunching Estimate for Einav et al. Setting .................... 76
  1.E Back-of-the-Envelope Calculation for the Comparison to the Medicare Reform ...................... 80
List of Tables

1.1 Summary Statistics Analysis Sample vs Remaining Sample . . . . . . . 57
1.2 How Often Do Kink Locations Change? . . . . . . . . . . . . . . . . . 57
1.3 Causal Effect of Increasing Marginal Reimbursement per Day in Hos-
pital by 1,000€ on Length of Stay . . . . . . . . . . . . . . . . . . . 58
1.4 Causal Effect of Increasing Marginal Reimbursement by 1,000€ per
Day in Hospital on Length of Stay - for Transfers . . . . . . . . . . 59
1.6 Fixed Effect Regression for the Effect of the Kink Location on Covari-
ates - Age . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
1.5 Fixed Effect Regression for the Effect of the Kink Location on Length
of Stay in Days . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 60
1.7 Fixed Effect Regression for the Effect of the Kink Location on Covariates-
Gender . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 61
1.8 Fixed Effect Regression for the Effect of the Kink Location on Covariates-
Number Diagnoses . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 61
1.9 Fixed Effect Regression for the Effect of the Kink Location on Covariates-
Number Procedures . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62
1.10 Distribution of Kink Locations in the Cross section . . . . . . . . . . 76
2.1 Summary statistics - analysis sample vs excluded births . . . . . . . 89
2.2 Summary statistics — 10p-2a vs 2a-10p . . . . . . . . . . . . . . . . . 89
2.3 Frequency of births (for the analysis sample) . . . . . . . . . . . . . . . 92
2.4 Length of stay ................................................. 96
2.5 Frequency of births (for the excluded births) ...................... 121
2.6 Birth weight .................................................. 122
2.7 Share female ................................................... 123
2.8 Number of Diagnoses ......................................... 124
2.9 Number of Procedures ........................................ 125
2.10 Number of additional midnights ................................. 126
2.11 Time of the day at discharge .................................. 127
2.12 Hazard rates .................................................. 128

3.1 Summary Statistics — Above and Below Median Income .......... 137
3.2 Response of Unemployment Hazards to Benefits by Previous Income . 140
3.3 Response of Unemployment Hazards to Benefits by Assets .......... 141
List of Figures

1.1 Within-DRG Payment Schedule as a Function of Length of Stay - Generic 42
1.2 Within-DRG Payment Schedule as a Function of Length of Stay -
   Example DRG I51Z ......................................................... 43
1.3 Histogram of the Revenue Loss when Discharging the Day Before the
   Kink Instead of the Kink Day (in €) .................................. 44
1.4 Histogram of the Revenue Loss when Discharging the Day Before the
   Kink Instead of the Kink Day (in Percent) .............................. 45
1.5 Average Length of Stay in Germany Over Time ......................... 46
1.6 Histogram Kink Locations ................................................ 47
1.7 Histogram Length of Stay ............................................. 48
1.8 Histogram of Length of Stay Relative to the Patient’s Kink Location 49
1.9 Hazard Rates Around Kink - Example DRG I51Z ....................... 50
1.10 Hazard Rates Around Kink - Pooled Sample ........................... 51
1.11 Payment Schedule Change - Example DRG H62A for Which the Kink
    Location Goes Down by One Day ....................................... 52
1.12 Hazard Rates Around Kink from One Year to the Next - Example DRG
    H62A for Which the Kink Location Goes Down by One Day .......... 53
1.13 Hazard Rates Around Kink from One Year to the Next if the Kink
    Location Goes Down by One Day - Pooled Sample ................... 54
1.14 Hazard Rates Around Kink from One Year to the Next if the Kink
Location Goes Down by One Day - Pooled Sample - October until March 55
1.15 Hazard Rates Around Kink from One Year to the Next if the Kink
Location Does Not Change - Pooled Sample ............................. 56
1.16 Deviating billed number of days and difference between discharge and
admission date - by admission month ................................. 63
1.17 Time Series Length of Stay - Selection of OECD Countries ......... 70
1.18 Time Series Length of Stay - Procedural vs Medical DRGs ......... 71
1.19 Time Series Length of Stay - by Quartile of Hospital w.r.t. 2005
Average Length of Stay ......................................................... 72
1.20 Heterogeneity in Hazard Rates Around Kink - by Quartile of Hospital
Size .......................................................... 73
1.21 Heterogeneity in Hazard Rates Around Kink - by Discharge Reason . 74
1.22 Birth Weight Manipulation ............................................. 75
2.1 Frequency of births ..................................................... 91
2.2 Covariates around midnight ........................................... 93
2.3 Length of stay ......................................................... 95
2.4 Number of additional midnights and average time of the day at discharge 98
2.5 Cumulative share of discharged newborns .......................... 99
2.6 Payment schedule for a healthy newborn ............................ 101
2.7 Payment schedule for John and Maynard ............................ 102
2.8 Fraction of significant hospital-specific RDD coefficients for the cumu-
lative fraction of discharged newborns after three additional midnights
plotted against the (binned) hospital-specific RDD coefficient for the
cumulative fraction after two add. midnights .......................... 105
2.9 C-Section probability of mothers admitted around midnight ...... 106
2.10 Frequency of births and average length of stay around the clock ... 108
2.11 Length of stay around the clock .......................... 109
2.12 Count and C-Section probability of mothers admitted in the morning 110
Chapter 1

Do Provider Incentives Always Affect Health Care Costs? New Evidence from Germany

1.1 Introduction

In the face of rising health care costs, countries around the world have reformed the way they reimburse health care providers, since traditional payment systems were thought to give a financial incentive to overprovide. An early and very famous example of such a reform took place in 1983 when Medicare changed hospital reimbursement for inpatient care from a fee-for-service system, which pays the hospital for each individual service provided, to a so-called prospective payment system. Prospective payment systems make the hospital’s reimbursement less dependent on the actually provided services and length of stay, but more strongly tied to the expected costs based on case-characteristics, hence giving less financial incentive to increase the number of services and the patient’s length of stay.
The 1983 Medicare reform is widely perceived as having reduced the number of services and, in particular, average length of stay (Coulam and Gaumer 1992) and more recent quasi-experimental evidence from the U.S. confirms that financial incentives for U.S. hospitals affect treatment and discharge decisions (Einav et al. 2017 and Eliason et al. 2016).\(^1\) Since the financial incentives apply to the hospitals and not the treating doctors, this evidence implies that U.S. hospitals are able to influence the treatment decisions in a way that benefits hospital profits.\(^2\) Therefore, the U.S. evidence might be very specific to its institutional and cultural context. While Germany in 2004 followed Medicare in moving to a prospective payment system for hospitals in the hope of reducing health care costs, it is an open question whether the U.S. experience of shorter hospital stays carries over to the German institutional context.

This paper provides quasi-experimental evidence on the causal impact of hospital financial incentives on length of stay in Germany. I make use of a unique feature in Germany’s reimbursement schedule for hospitals. Like their Medicare counterparts, German patients are grouped into Diagnostic Related Groups (DRGs) based on diagnoses, major procedures and patient demographics. Within a DRG, the reimbursement increases linearly in length of stay up to a certain number of days at which it kinks and becomes flat. Hence, the hospital’s marginal reimbursement for keeping a patient in the hospital for another day drops discontinuously at the kink. Thus, discharging a patient becomes discontinuously more attractive the moment the kink is reached. If the hospital’s financial incentives affect the decision about its patients’ day of discharge, more patients will be discharged on the kink day than what would be expected under a smooth payment schedule without kinks, i.e. there will be

\(^1\)Bajari et al. (2017) is another paper using non-linear schedules to study the impact of marginal reimbursement rates for hospitals on treatment decisions in the U.S. — they focus on organ transplants and on heterogeneity in the response across patients.

\(^2\)There is also a large literature documenting that financial incentives that directly apply to the doctors affect treatment choices, e.g. Rice 1983, Nguyen and Derrick 1997 Yip 1998, Gruber, Kim, and Mayzlin 1999, Jacobsen et al 2010, Clemens and Gottlieb 2014, Coey 2015, Alexander 2015
bunching of discharges on the kink day. I demonstrate theoretically from estimating the amount of bunching one can infer the causal impact of changing the marginal reimbursement to the hospital for keeping a patient another day on average length of stay.

The empirical analysis is conducted using administrative data covering the universe of in-patient hospitalizations in Germany from 2005-2013, amounting to more than 130 million cases. I start by presenting suggestive evidence on the effect of the marginal reimbursement for another day in the hospital on length of stay in Germany. First, Germany’s major 2004 reform that introduced a prospective payment system based on DRGs and reduced the marginal reimbursement for most patients to zero produced no notable break in average length of stay in the time series, suggesting little causal impact. Second, examining the hazard rates around the kink day reveals no notable excess mass of patients being released, again pointing towards very small effects.

This simple static bunching analysis cannot be used to tightly bound the causal effect though, since due to the discrete nature of the assignment variable—days in the hospital—one would need to make strong functional form assumptions regarding the shape of the hazard function. For my main analysis, I therefore make use of the fact that the exact day at which the payment schedule kinks is not only DRG-specific, but can also change from one year to the next. Hence, one can directly evaluate how a DRG’s patients’ hazard rates compare from one year to the next as the DRG’s kink location changes, allowing for a compelling visual assessment to which degree the discharge decisions respond to the financial incentives. Econometrically, the changing kink locations make it possible to estimate the amount of bunching purely from changes in hazard rates from one year to the next without the need for any functional form or smoothness assumptions regarding the shape of the hazard.
Consistent with the suggestive evidence, the visual assessment shows no indications that the hazard rates respond to the changing kink location. Moreover, I can tightly bound the effect of changing the marginal reimbursement hospitals receive for keeping patients another day. Specifically, I can reject that length of stay would fall by more than 0.05 days if the marginal reimbursement for another day in the hospital was reduced by 1,000€.³

What do my results imply for the effectiveness of the 2004 German reform? Germany introduced its prospective payment system based on DRGs coming from a per diem system, that is a system that pays a fixed amount per hospitalization day.⁴ German politicians reformed the hospitals’ incentives aiming for a reduction in length of stay. Ulla Schmidt, who was the federal minister of health when the German prospective payment system was introduced, expresses her belief that the system was successful in achieving this goal in a 2009 interview as follows:⁵

The parliament has passed many reforms in the last years that have proven to make a positive difference in the health care sector. Some things that caused protests initially are now universally accepted as successful. For instance, the DRG system was portrayed like the end of the hospital as we know it. Today we know: Length of stay has decreased and the system has become more efficient.

Since the reimbursement per day under the pre-2004 per-diem system was, on average, less than 1,000€, the 0.05 provides an upper bound for the reform-induced reduction in length of stay as well. The politicians claiming success apparently confused a secular declining trend in length of stay with the causal effect of their reform.

³Euro in this paper always refers to 2013-Euro.
⁴The system was introduced in 2003 on a voluntary basis. In 2004, it became mandatory for hospitals to participate. Before 2004, the per diem system applied to the majority of cases, but some cases were already reimbursed prospectively.
⁵Dtsch Arztebl 2009; 106(26) - translation by this paper’s author
My results stand in sharp contrast to recent research by Einav et al. (2017), Eliason et al. (2016) and Kim et al. (2015). Their papers study the effects of hospital incentives in the U.S. by exploiting the fact that the Medicare reimbursement for post-acute care hospitals jumps at a certain length of stay. This notch causes a pronounced excess mass of discharges just after the payment amount jumps. In order to gain directly comparable estimates, I use data from Einav et al. (2017) to implement a bunching analysis similar to my main analysis for the Medicare notch and find an implied reduction in length of stay by 0.34 days when cutting marginal reimbursement for another day in the hospital by 1,000€, an estimate seven times as big as the upper bound of my confidence interval.

My results also contrast with the experiences from the 1983 Medicare reform. While the magnitudes cannot be compared directly, I argue later in the paper that my estimates would suggest much smaller effects on length of stay than what is normally attributed to the 1983 Medicare reform.

In the discussion section of this paper, I consider why the effects of marginal incentives for hospitals are an order of magnitude smaller in the German setting than in the United States. Building on facts about the institutional structure, anecdotal evidence and interviews with doctors, I argue that U.S. hospital administrators are very active in shaping the patients’ treatments while German doctors decide very independently of their hospitals’ wishes. I provide evidence that German hospitals are profit-maximizing in principle (specifically, they do make use of the possibility to change the coding to increase revenue), but they are unable to manipulate actual treatment decisions which are entirely decided about by the doctors. Because of such institutional and cultural differences between Germany and the U.S., politicians and researchers should be very cautious when extrapolating reduced-form effects from the U.S. to the very different settings in Europe.
For future policy, my results suggest that politicians—in institutional settings in which doctors act very independently—could be more effective in their efforts to cut health care costs if they shifted their focus on incentivizing these doctors directly instead of incentivizing the hospitals. Moreover, future research should investigate whether—in countries such as Germany with very independently acting doctors—tying hospital reimbursement more closely to length of stay again might improve welfare, because it would make the reimbursement more closely connected to the hospitals’ actually incurred costs. This closer connection might, firstly, reduce the hospitals’ financial risk of drawing particularly sick patients and, secondly, might help eliminate the incentive for hospitals to discriminate against particularly costly patients along the admissions margin.

The rest of the paper is organized as follows. Section 1.2 gives institutional background. In Section 1.3, I demonstrate theoretically that the bunching design identifies the effect of marginal reimbursement on length of stay. Section 1.4 discusses the data and the sample selection. Section 1.5 presents the results. Section 1.6 provides a discussion and concludes.

1.2 Institutional Background

Germany’s health care system is one of the most expensive among OECD countries. In 2013, Germany spent 11.0% of its GDP on health care (OECD average 8.9%, U.S. 16.4%) putting it on fifth position in the OECD. With 8.3 hospital beds per 1,000 people in 2013 (OECD average 4.8 beds, U.S. 2.9 beds) the hospital sector has a high level of utilization in international comparison, reflecting a high average length of stay (9.1 days in 2013, OECD average 8.1 days, U.S. 6.1 days) as well as a large total number of hospitalizations (252 per 1,000 people in 2013, OECD average 155, U.S. 125).
Hospital reimbursement in Germany is determined at the federal level and is the same nationwide (except for hospital-specific proportional shift factors as discussed below), irrespective of the patient’s health insurer.

Until 2004\textsuperscript{6}, Germany reimbursed hospitals using a cost-based per diem system in about 80\% of cases (the remaining 20\% were already reimbursed using a fixed prospective payment, see Theilen 2004). That is, the fee payable to the hospital increased linearly (with a hospital- and department-specific slope depending on the hospital’s historical costs) in the number of days a patient stayed hospitalized. In the face of rising health care costs, the German government decided to transition to a prospective payment scheme based on DRGs. The vast majority of cases (more than 94\% in 2013) are now reimbursed according to the DRG system (the most prominent exception are the psychiatric cases which only in recent years started to transition to a separate prospective payment system). Based on diagnoses, major procedures and the patient’s age, each case is grouped into one out of more than 1,000 DRGs. Due to the complexity of the grouping, more than 75\% of hospitals in 2011 employed clinical coders whose main duty is to correctly code diagnoses, procedures and, ultimately, DRGs (Franz et al 2011).

In some cases, the DRG classification can also depend on further variables like birth weight, the discharge reason (e.g., whether the person died) or length of stay. In particular, there are many one-day-DRGs which determine reimbursement in the special case of a patient having a certain diagnosis and staying just one day. These one-day-DRGs do not pose a problem for my design, however, since for my main research design I only use year-to-year changes in hazard rates for DRGs for which the patient composition is the same from one year to the next according to the official DRG migration tables.\textsuperscript{7} That is, DRGs for which the patient composition changes

\textsuperscript{6}Technically, the system already switched in 2003, but it only became compulsory in 2004.

\textsuperscript{7}The DRG migration table from $t-1$ to $t$ considers all patients from $t-2$ and groups them into the appropriate DRG according to the system in $t-1$ and according to the system in $t$. The table then shows how DRGs from $t-1$ map into DRGs from $t$. For the analysis, I restrict the attention
mechanically because, e.g., a new one-day-DRG is introduced are not part of the sample.

The DRG definitions are updated every year and designed to maximize cost homogeneity within DRGs while keeping the number of different DRGs within reasonable limits. The definitions for year \( t \) are based on cost data that are collected from a sample of hospital in \( t-2 \).

**Payment Scheme**

Within a DRG, the fee payable to the hospital (if the patient is not transferred to or from the hospital — transferred patients are subject to a different payment schedule and analyzed later in a separate section) is a function of the hospital stay length as depicted in Figure 1.1 (Figure 1.2 shows a specific example of a DRG payment schedule. As it is apparent from the graph, this DRG has its kink point at five days.). The parameters of the payment scheme are — as the DRG definitions — based on the hospital cost data from two years before. The payment increases linearly until a third of the average length of stay (rounded and measured two years prior) of all patients in this DRG is reached (but at least until day 2 is reached). The slope is determined by dividing average variable costs (that is, total costs excluding costs of major procedures, e.g., bypass surgery) of all patients with this DRG by the number of days at which the kink occurs (again, costs measured two years prior). After the kink, the payment schedule remains flat until the average plus two times the standard deviation of the length of stay of all patients with this DRG two years prior is reached.\(^8\) From then on it increases again linearly. Any out-patient treatments to DRGs that have a one-to-one mapping from \( t-1 \) to \( t \), that is DRGs with an unchanged patient composition.

\(^8\)To be precise, the upper kink point is the average length of stay plus the maximum of two times the standard deviation or a maximum difference that is determined every year (e.g., in 2005 the upper kink point could at most be the average length of stay plus 17 days)
- prior to admission or post discharge - by the hospital are included in the DRG payment (but do not count towards the number of days in the hospital).\(^9\)

Interestingly, while the overall goal of the DRG reform was to reduce length of stay, the reduced reimbursement to the left of the lower kink point was introduced in order to discourage hospitals from discharging patients extremely early.

Doctors can easily get information on the DRG’s kink location - either because they code the diagnoses and procedures themselves (in which case the software tells them all the information about the patient’s DRG) or because they can ask their clinical coder with whom they work closely.

Figure 1.3 demonstrates how strongly marginal reimbursement changes at the kink. For each DRG, I calculate the slope in € to the left of the lower kink, that is how much revenue the hospital loses if the patient is discharged the day before the kink instead of the kink day. The graph is a histogram of this measure across all years and DRGs. The distribution is centered around about 1,000€, but with a heavy right tail.

In Figure 1.4, I also show a histogram of the percentage loss in hospital revenue if the patient is discharged the day before the kink instead of the kink day. I.e. for each DRG I calculate the ratio of the slope to the left of the lower kink and the total amount that the hospital gets paid in the flat part of the schedule and Figure 1.4 is a histogram of this measure across all DRGs and years. In general, the financial loss of discharging a patient a day before her kink day is quite substantial, although there is a lot of heterogeneity across DRGs.

One concern for identification is that a patient’s DRG is not necessarily fixed throughout her hospital stay, but can change if, for instance, the patient gets an infection and therefore a new major diagnosis. In the interviews I conducted, doctors

\(^9\)This is unless the total number of days (in-patient days plus treatment days pre-admission and post-discharge) exceeds the upper kink point (average length of stay plus two times the standard deviation)
working in German hospitals confirmed that in typical cases the patient’s DRG is very predictable from day one.

**Demand Side Institutions**

Nearly 90% of Germans are in the public health insurance system. There are more than 100 different public health insurers (all public corporations) which compete with each other for patients. The rules for the reimbursement of providers as well as copayments are, however, highly regulated and very similar across insurers. Public health insurance covers all costs of hospital stays except for a copay of 10 Euro per day. The copay is only payable for up to 28 days a year, so there is a minor kink from the patient’s perspective at 28 days, but she is not affected by the kink in the hospital payment schedule. Self-employed workers, civil servants as well as employees who earn above a certain threshold can opt to be privately insured. About 10% of people are covered by private health insurers. Private health insurance contracts do typically include a deductible. Hence, these patients have an incentive to contain costs. Unfortunately, I cannot distinguish between publicly and privately insured patients in my data. So only about 10% of patients have to pay a deductible and hence have financial incentives inverse to those of the hospital. These patient incentives contrast with Medicare which features a deductible for most patients, i.e. in Medicare patient incentives are a stronger force countering the hospital incentives than in Germany.\(^\text{10}\)

The health insurers in Germany can —as in Medicare— audit bills and appeal. In 2013, 4.4% of cases were successfully audited concerning the length of stay. If an audit is unsuccessful, the health insurer has to pay 300 Euros to the hospital in compensation for the wrong accusation and the resulting work load for the hospital. If the audit is successful, on the other hand, the bill is simply adjusted, but there

---

\(^{10}\)Medicare patients do not pay a deductible if they are also on Medicaid or pay for supplemental Medigap.
is no fine for the hospital. According to the health insurers, there therefore is little incentive for the hospitals to adjust their bills in anticipation of the audits.\footnote{see, e.g., Faktenblatt Thema: Abrechnungsprüfung in Krankenhäusern from 06/06/2014} As demonstrated in the generalized model in appendix 1.B, the presence of audits would not confound the research design even with anticipatory behavior, since the causal effect of interest is the effect given the presence of demand side constraints. The audits can, however, introduce measurement error into the length of stay variable. I discuss the issue of measurement error in detail in Section 1.4.

Supply Side Institutions

In 2013, there were 1,995 hospitals in Germany, with 596 being public, 706 non-profit and 693 for-profit. This contrasts with a market for long-term care hospitals studied by Einav et al. (2017), Eliason et al. (2016) and Kim et al. (2015). The long term care hospital market in the U.S. is dominated by for-profit hospitals (Eliason et al. 2016).

The payment schedule that hospitals face (as shown in Figure 1.1) is identical across hospitals except for a proportional shift factor. The shift factor was different for each hospital when the new system was introduced in 2004 (the hospital-specific shift factors were introduced in order to have a smooth transition and no sudden jump in hospital revenue relative to the old system which paid hospitals a hospital-specific amount for each day a patient stayed in the hospital), but has since converged to a single factor within each state. Since 2010 the statewide shift factors have been converging towards a factor that is common nationwide. Every year, each hospital and the health insurers agree on a hospital budget, the expected total revenue of the hospital. Deviations from this expected revenue are only partially compensated in order to insure the hospital against random fluctuations in revenue - i.e. changes in
treatment decisions that affect hospital reimbursement only partially translate into actual revenue initially until the expected revenue is adjusted.

Hospitals have further revenue streams besides the DRG reimbursement. Hospitals can bill separately for a specified list of rare and highly expensive procedures that are not tied to one specific DRG (e.g., implementation of a vagus nerve stimulator). Moreover, hospitals receive additional funds depending on, for example, the amount of investments, whether the hospital provides an emergency room or the degree to which the hospital participates in the training of new doctors. However, these additional funds do not affect the discontinuous break in marginal reimbursement at the kink and are therefore no threat to identification.

German doctors working in hospitals are salaried and unionized, except for the head physicians whose pay is individually contracted and does often depend on economic outcomes in her department (e.g., the head physician’s pay can depend on the number of times a specific procedure like hip replacement takes place in the head physician’s department). In the case of privately insured patients (or publicly insured patients who are willing to pay extra money in order to be treated by the head physician) the head physician can charge additionally per service. Typically, these additional charges are then shared with the other doctors in her department. The discharge decisions is usually made by the patient’s responsible doctor. While patients can choose to leave the hospital against their doctors advice, such cases are coded in the data and very rare events. Discharges typically take place midday after the doctor’s ward round.

The employer-employee relationship between U.S. hospitals and their doctors differs from the one in Germany in that U.S. doctors are typically not hospital employees.\textsuperscript{12} Instead, most doctors working in hospitals are contracting independently with

\textsuperscript{12}In 2013, “roughly 25 percent of all specialty physicians who see patients at hospitals are employed” (\textquotedblleft 7 Trends in Hospital-Employed Physician Compensation\textquotedblright in Becker’s Hospital Review 01-25-2013)
the hospital (often times as a group of physicians and often times with several hospitals for the same doctor) regarding the use of the facilities and are reimbursed for treating Medicare patients by Medicare directly using a fee for service system. What is similar between the German and the U.S. system, however, is that the financial incentives that the hospitals face are at odds with those faced by the doctors (Alexander 2017). The fact that financial incentives for hospitals work so well in the U.S. despite doctors not being directly affected is therefore particularly interesting and discussed more in section 1.6.

Medical liability risk in Germany is generally perceived to be small relative to that in the U.S. due to comparatively small awards against physicians. Richard A. Epstein, director of the law and economics program at the University of Chicago Law School, can be quoted with "Nobody is as hospitable to potential liability as we are in this country. The unmistakable drift is we do much more liability than anybody else, and the evidence on improved care is vanishingly thin".

1.3 Theory

Summary

This section demonstrates that a bunching design identifies the causal impact of marginal reimbursement for another day in the hospital on average length of stay. The model features heterogeneous patients, allows for generic health production and costs functions as well as agency frictions between the hospital and the medical personnel. Under one additional regularity assumption relative to the standard bunching design as in Kleven (2016) — the assumption is needed due to the discreteness of the assignment variable ‘days in the hospital’ —, the causal impact of marginal reimbursement can be calculated from the estimated amount of bunching and the estimated mass

\footnote{see, e.g., Law Library Of Congress - Medical Malpractice Liability Systems In Selected Countries} \footnote{American Medical News, May 3, 2010}
of patients that would have been discharged at the days above the kink day under a counterfactual smooth payment schedule.

Setup

I present a simplified model here - in appendix 1.B I show that the result holds in a more general model with risk-averse hospitals (the identified effect then applies to a change in marginal reimbursement that also adjusts the fixed payment component such that hospital profits remain constant in equilibrium) as well as the possibility of audits by the health insurers.

The hospital admits a continuum of patients of type $\theta_i$ who stay $d_i$ days and enjoy health benefit $h(d_i, \theta_i)$ which is concave in the number of days in the hospital, i.e. $\frac{\partial^2 h(d_i, \theta_i)}{\partial d_i^2} < 0$. Since there are no functional form assumptions on how $\theta_i$ affects $h(d_i, \theta_i)$, one can assume a uniform distribution $\theta_i \sim U[0, 1]$ without loss of generality - i.e. $\theta_i$ is patient $i$’s quantile in the distribution of patient sickness. Sicker patients benefit more from staying in the hospital for longer. That is, $\frac{\partial^2 h(d_i, \theta_i)}{\partial d_i \partial \theta_i} > 0$.

The hospital receives payment $P(d_i)$ and incurs costs $C(d_i)$. The hospital gains utility from profits and from its patients’ health (either because of an intrinsic concern for their patients’ health or because they fear lawsuits or reputational costs if patients are mistreated). The hospital values profits relative to patient health benefits according to preference parameter $\lambda_h \geq 0$. Since in practice, the medical personnel and not the hospital shareholders make the discharge decision, the hospital faces an agency problem in implementing its objective function. I model this agency problem in the form of parameter $0 \leq \lambda_d \leq 1$ which dampens the degree to which profits are taken into account. That is, the patients’ length of stay is determined by the solution to

$$\max_{\{d_i(\theta_i)\in \mathbb{N}\}} \lambda_d \lambda_h \int_i [P(d_i) - C(d_i)] + \int_i h(d_i, \theta_i)$$
If the agency problem were modeled in a different way, the bunching design would not necessarily identify the causal effect of interest anymore, because the kinks might have effects on hazard rates away from the kink point. I discuss this point in more detail later in the results section and provide evidence against such effects of the kink on hazard rates away from the kink.

Note also that changes in admission and coding behavior - while interesting subjects to study in their own right - do not threaten the validity of this paper’s findings. If anything, adjustments in coding and admission behavior would lead to an overestimate of the bunching mass in my setting. This is because the incentive to deny admission or to upcode to a different diagnosis with a different kink location is smallest for those patients who would otherwise be discharged on the profit maximizing kink day.

**Optimal Hospital Behavior**

I assume that $P(d_i), C(d_i)$ and $h(d_i, \theta_i)$ are shaped such that the objective function is globally concave. At baseline, consider a linear payment schedule $P^{baseline}(d_i) = \bar{p} + p \cdot d_i$.

Optimization than amounts to choosing cutoff values for $\theta_i$ determining which patient types are kept for how many days. $\bar{\theta}_d^{baseline}$ denotes the highest $\theta_i$ for which the patient stays $d$ days under the baseline schedule. A patient with a $\theta_i$ just above $\bar{\theta}_d^{baseline}$ would stay $d + 1$ while a patient with a $\theta_i$ just beneath $\bar{\theta}_d^{baseline}$ would stay $d$ days. The cutoff values defining the range of patients who are discharged on day $d^*$ are implicitly defined by equations

$$
\lambda_d \lambda_h \left[ C(d^* + 1) - C(d^*) - p \right] = h\left(d^* + 1, \bar{\theta}_d^{baseline}\right) - h\left(d^*, \bar{\theta}_d^{baseline}\right)
$$
\[ \lambda_d \lambda_h [C(d^*) - C(d^* - 1) - p] = h\left(d^*, \bar{\theta}_{d^* - 1}^{\text{baseline}}\right) - h\left(d^* - 1, \bar{\theta}_{d^* - 1}^{\text{baseline}}\right) \]

That is, for patient type \( \bar{\theta}_{d^*}^{\text{baseline}} \) the hospital is just indifferent between the net profit valued with \( \lambda_d \lambda_h \) of keeping her \( d^* + 1 \) instead of \( d^* \) days and the net health benefit it would bring to the patient. A patient with \( \theta_i \) a little bigger than \( \bar{\theta}_{d^*}^{\text{baseline}} \) would be kept \( d^* + 1 \) days, since her health benefit of staying another day is higher than for the \( \bar{\theta}_{d^*}^{\text{baseline}} \) patient. Similarly, for patient type \( \bar{\theta}_{d^* - 1}^{\text{baseline}} \) the hospital is indifferent between the marginal health benefit of keeping her \( d^* \) instead of \( d^* - 1 \) days and the profit impact.

Now consider the policy experiment of interest, reducing marginal reimbursement by \( \Delta p > 0 \) throughout the schedule, i.e. \( P_{\text{reform}}(d_i) = \bar{p} + (p - \Delta p) \cdot d_i \), with \( \Delta p > 0 \). Note that with risk-neutral hospitals, the fixed payment amount \( \bar{p} \) does not affect hospital behavior - in appendix 1.B, I consider risk-averse hospitals. The cutoff values are now defined by

\[ \lambda_d \lambda_h [C(d^* + 1) - C(d^*) - (p - \Delta p)] = h\left(d^* + 1, \bar{\theta}_{d^*}^{\text{reform}}\right) - h\left(d^*, \bar{\theta}_{d^*}^{\text{reform}}\right) \]

\[ \lambda_d \lambda_h [C(d^*) - C(d^* - 1) - (p - \Delta p)] = h\left(d^*, \bar{\theta}_{d^* - 1}^{\text{reform}}\right) - h\left(d^* - 1, \bar{\theta}_{d^* - 1}^{\text{reform}}\right) \]

which implies that the cutoff values increase, i.e. \( \bar{\theta}_{d}^{\text{reform}} > \bar{\theta}_{d}^{\text{baseline}} \forall d \). That is, the patients stay on average for a shorter time. Due to the discreteness of the assignment variable — length of stay — I need to make an additional regularity assumption relative to the standard bunching setting. Specifically, I assume that patients who share the same length of stay \( d \) under the old schedule \( P_{\text{baseline}}(d_i) \) move towards at most two different length of stay values under the new schedule \( P_{\text{reform}}(d_i) \). That is, those
patients who stay, e.g., 5 days under the old schedule, stay for 3−4 days or for 4−5 days under the new schedule, but never for 2−4 or 3−5 days.

Now consider the introduction of a convex kink at $d^*$. That is, the payment schedule becomes

$$
P_{kink}^{}(d_i) = \begin{cases} 
\bar{p} + p \cdot d_i & d_i \leq d^*

\bar{p} + (p - \triangle p) \cdot d_i & d_i > d^*
\end{cases}
$$

Under the new kinked payment schedule, the new cutoff values defining who is discharged at $d^* - \bar{\theta}_{d^*}^{kink}$ and $\bar{\theta}_{d^* - 1}^{kink}$— are defined by

$$
\lambda_d \lambda_h [C(d^* + 1) - C(d^*) - (p - \triangle p)] = h\left(d^* + 1, \bar{\theta}_{d^*}^{kink}\right) - h\left(d^*, \bar{\theta}_{d^*}^{kink}\right)
$$

$$
\lambda_d \lambda_h [C(d^*) - C(d^* - 1) - p] = h\left(d^*, \bar{\theta}_{d^* - 1}^{kink}\right) - h\left(d^* - 1, \bar{\theta}_{d^* - 1}^{kink}\right)
$$

Hence, $\bar{\theta}_{d^* - 1}^{kink} = \bar{\theta}_{d^* - 1}^{baseline}$ and $\bar{\theta}_{d^*}^{kink} = \bar{\theta}_{d^*}^{reform} > \bar{\theta}_{d^*}^{baseline}$ and $\bar{\theta}_{d^*}^{kink} - \bar{\theta}_{d^*}^{baseline}$ is the excess mass or bunching at $d^*$ under the kinked schedule. Hence, $\bar{\theta}_{d^*}^{kink}$ is the marginal buncher who responds to the introduction of $P_{kink}^{}(d_i)$ the same way as to the introduction of $P_{reform}^{}(d_i)$.

**What Does a Bunching Design Identify?**

We established that the marginal buncher responds to the introduction of the kink the same way as to the policy experiment of interest (that is, changing marginal pay by $\triangle p$ throughout the schedule). Let $\tilde{d}$ denote the length of stay that the marginal buncher $\bar{\theta}_{d^*}^{kink}$ would have enjoyed under the baseline linear schedule. For this marginal buncher, the causal effect of interest —the effect of changing marginal reimbursement per day by $\triangle p$ on length of stay— is $\frac{d(d_i)}{d(p)} \triangle p = \tilde{d} - d^*$. Using the assumption discussed
above, the causal effect $\frac{d(d_i)}{d(p)} \Delta p$ is equal to $\tilde{d} - d^*$ for all patients who would have stayed $\tilde{d}$ under the baseline linear schedule and for whom $\theta_i < \tilde{\theta}_d^{kink}$, but the causal effect $\frac{d(d_i)}{d(p)} \Delta p$ is $\tilde{d} - (d^* + 1)$ for all patients who would have stayed $\tilde{d}$ under the baseline linear schedule and for whom $\theta_i > \tilde{\theta}_d^{kink}$. Therefore, the total causal effect on patients staying $\tilde{d}$ under the old baseline schedule is

$$E \left[ \frac{d(d_i)}{d(p)} \Delta p \mid \theta_d^{baseline} > \theta_i > \theta_d^{baseline} - 1 \right] = \left( \tilde{d} - d^* \right) \frac{\tilde{\theta}_{d^*}^{kink} - \tilde{\theta}_d^{baseline}}{\tilde{\theta}_d^{baseline} - \theta_i}$$

$$+ \left( \tilde{d} - (d^* + 1) \right) \frac{\tilde{\theta}_d^{baseline} - \tilde{\theta}_d^{kink}}{\tilde{\theta}_d^{baseline} - \theta_i}$$

$$= \tilde{d} - d^* - 1 + \frac{\tilde{\theta}_{d^*}^{kink} - \tilde{\theta}_d^{baseline}}{\tilde{\theta}_d^{baseline} - \theta_i}$$

A simple example makes the formula intuitive: If the observed bunching mass is only a small fraction of the observed mass at $d^* + 1$—say, 10%—$\tilde{d} = d^* + 1$, because the marginal buncher is coming from $d^* + 1$, and $\frac{\tilde{\theta}_{d^*}^{kink} - \tilde{\theta}_d^{baseline}}{\tilde{\theta}_d^{baseline} - \theta_i} \approx 0.1$, since only patients who are at $d^* + 1$ under the counterfactual linear schedule and whose $\theta_i < \tilde{\theta}_d^{kink}$ bunch at $d^*$ together with the marginal buncher. In the example, the formula tells us that the average causal effect on the patients staying $d^* + 1$ days under the counterfactual linear schedule is 0.1 days, since that is the fraction of patients who move from $d^* + 1$ to $d^*$ due to the kink.

Since $d^*$ is known, we need to estimate $\tilde{d}$ and $\frac{\tilde{\theta}_{d^*}^{kink} - \tilde{\theta}_d^{baseline}}{\tilde{\theta}_d^{baseline} - \theta_i}$ in order to get the causal effect of interest. Let $B$ denote the bunching mass estimated from the data and $f(d)$ the estimated expected mass of patients at $d$ under the counterfactual linear schedule. Then $\tilde{d}$ can be inferred from the data by finding the value for $\tilde{d}$ for which

$$f(d^* + 1) + ... + f(\tilde{d}) \geq B$$
\[ f(d^* + 1) + ... + f(\tilde{d} - 1) \leq B, \]

since the bunching mass is equal to the mass at the days from \( d^* + 1 \) up to \( \tilde{d} - 1 \) plus the fraction of the mass at \( \tilde{d} \) that bunches. This fraction is the bunching mass that is not explained by the mass coming from \( d^* + 1 \) up to \( \tilde{d} - 1 \), i.e.

\[
\frac{\bar{\theta}_k - \bar{\theta}_{baseline}}{\bar{\theta}_{baseline} - \bar{\theta}_{baseline}^{d-1}} = B - f(d^* + 1) + ... + f(\tilde{d} - 1)
\]

Hence, the remaining challenge is to estimate \( B \) and \( f(d^* + 1) \), etc. in order to estimate the parameter of interest \( E\left[ \frac{\Delta p}{\bar{\theta}_{baseline}} \bigg| \bar{\theta}_d > \bar{\theta}_{baseline} \right] \).

### 1.4 Data and Sample Selection

**Data**

I use administrative data from the Federal Statistical Agency in Germany.\(^{15}\) It covers the universe of in-patient hospitalizations covered by the DRG system\(^{16}\) from 2005 - 2013, more than 10,000,000 cases each year with variables including baseline patient-characteristics like sex, age and region as well as case-characteristics like diagnoses, procedures, length of stay, hospital identifier and admission and discharge date. All hospitals are required by law to report all of the previous year’s hospitalizations until March 31st to the Federal Statistical Agency. The data the hospitals send to the agency is based on the data generated for billing purposes and hence of very high quality.

**Measurement Error in Length of Stay**

\(^{15}\)The official source for all reported results in this paper is: 'FDZ der Statistischen Ämter des Bundes und der Länder, DRG-Statistik 2005-2013, eigene Berechnungen' (Translation: ‘Research Data Center of the Federal and State Statistical Agencies, DRG-Statistics 2005-2013, own calculations’)

\(^{16}\)As mentioned before, the vast majority - more than 94% in 2013 - are reimbursed according to the DRG system. The only major exception are psychiatric patients.
Length of stay may be mismeasured for two reasons: First, if a patient is readmitted within 30 days or before her DRG’s upper kink point is reached —counting from the first admission date— and if the patient obeys certain criteria, the two hospitalizations are merged into one case and length of stay is summed up. That is, only one case would show up in my data and the hospital is reimbursed as if it were one case. Second, health insurer’s audits might introduce measurement error into my length of stay variable. If a bill is successfully audited the actual length of stay and the billed one (which is the one that shows in the data) can deviate. If patients discharged on the kink day are especially likely to be audited or to be readmitted, this measurement error could lead to an underestimate of bunching.

There are two reasons why my results appear to be robust with respect to the measurement error problem: First, I have two different measures for length of stay which are differentially affected by the two types of measurement error. The billed number of days —which yields the correct length of stay in the case of a readmission, but misreports length of stay in case of an audit, since the billed number of days is adjusted after an audit— and the difference between discharge and admission date —which produces some measurement error in the case of readmissions (since days after the first discharge and before the readmission would wrongly be counted towards total length of stay), but is not affected by audits, since admission and discharge date are unadjusted after a successful audit. I conducted my analysis using both measures and the results are very similar, suggesting that the measurement error is of little importance. The reported results in this paper are for the difference between discharge and admission date.

Second, while the billed number of days mismeasures length of stay in the case of a successful audit, this problem is much larger for cases admitted early in the calendar year than for cases admitted closer to the end of the calendar year due to the point in time at which the data is collected - for a detailed discussion see appendix 1.A. Using
the billed number of days, I find no evidence that there is more bunching in the data for cases from later in the year, again suggesting that the measurement error is not the driving force behind the results.

**Sample**

Throughout the analysis, I exclude those with missing data on DRG, length of stay, discharge reason or admittance reason. Also I focus on discharges the timing of which actually are under the doctor’s control - i.e. I drop deaths or discharges against the doctor’s advice from the sample. Furthermore, I analyze transfers to or from other hospitals in a separate section, since those are subject to special reimbursement rules. After applying these restrictions I am left with more than 85% of the overall number of cases. The remaining 15% are mostly due to deaths and transfers.

For data privacy reasons I cannot make use of observations for which there are less than 3 patients that are discharged with a certain DRG, in a certain year and after a certain number of days in the hospital. This measurement error will only affect very uncommon DRGs and is unlikely to significantly affect any of my results, especially for the weighted regressions.

### 1.5 Results

**Time Series Evidence**

Figure 1.5 shows the development of average length of stay in Germany over time around the 2004 introduction of the DRG system.\footnote{Note that the source is the official German hospital statistic here which is why the numbers are different than the OECD numbers for length of stay for Germany. I use the OECD numbers only when comparing Germany to other countries.} Average length of stay is on a secular declining trend (the trend is also present before and after the time window shown in Figure 1.5). This trend is not specific to Germany, but can be observed
in many developed countries. Figure 1.17 in appendix 1.C shows how length of stay evolved over the 2000s in a sample of OECD countries.

The secular decline in length of stay appears to not only be common across countries, but also across hospitalization reasons and hospitals within Germany. First, technical progress making surgeries less taxing is not the sole driver of the decline. Figure 1.18 in appendix 1.C shows how length of stay evolves for DRGs that involve a major procedure compared to DRGs that do not. While the absolute decline in length of stay is larger for DRGs involving procedures, the trend for the other DRGs is similar. Second, the trend is not specific to the hospitals that started out with particularly long durations. I split up the hospitals into quartiles based on their average residualized length of stay in 2005. Figure 1.19 in appendix 1.C shows how average length of stay evolves for those four groups of hospitals. The hospitals in the fourth quarter that had unusually long durations in 2005 do indeed see the largest fall in length of stay afterwards, but a declining trend is also visible for the remaining hospitals.

As discussed in more detail in Section 1.6, the German 2004 reform changed hospital reimbursement for most patients from being linearly increasing towards being flat in length of stay, i.e. the marginal reimbursement was reduced to zero for most patients. Had the marginal reimbursement for another day in the hospital any meaningful effect on length of stay in Germany, we would expect to see a downward jump in length of stay around the time of introduction. Yet there is no apparent break in the time series around 2004 suggesting little causal impact. However, the pre-existing trend towards shorter stays and the missing control group as well as the possibility of changes in coding and admission behavior make it difficult to draw confident conclusions just from the time series.

Static Bunching Analysis
Therefore, I next turn to the static bunching analysis making use of the kink in the payment schedule. To provide some sense for the distribution of the kink locations as well as length of stay in the cross section, Figure 1.6 shows the distribution of kink locations across DRGs, Figure 1.7 provides a histogram for length of stay and Figure 1.8 shows how often patients stay shorter respectively longer than their DRGs’ kink locations.

If marginal reimbursement had an influence on the discharge decision, the hazard rates on the kink day should be unusually large. Figure 1.9 presents the hazard rates around the kink for the example DRG discussed earlier (the payoff schedule is shown in Figure 1.2). There is no bunching apparent at the kink at five days.

To graphically analyze the degree of bunching for the pooled data, I restrict the sample to DRGs with a kink location at six days or higher and center all observations around their respective kink location and pool them. Figure 1.10 shows the resulting hazard rates plotted against the number of days in the hospital relative to the kink location of the patient’s DRG. If the marginal reimbursement for the hospital had a meaningful impact on discharge decisions, we would see an unusually large hazard rate at 0. Yet, there is no apparent excess hazard at the kink point, again pointing towards no major effects of marginal reimbursement on length of stay decisions. The graph looks similar for other restrictions on the data like selecting only DRGs with a kink location of at least 5 or 7 days.

**Dynamic Bunching Analysis**

The coarse nature of the running variable ‘days in the hospital’ makes it difficult to implement the standard static bunching design econometrically, since the smooth counterfactual hazard rate in the absence of the kink cannot be pinned down precisely. Therefore, I make use of DRGs with changing kink locations over time to get precise counterfactual hazards without functional form assumptions.
Since the DRG definitions are updated every year, there are mechanical changes in the patient composition for some DRGs from one year to the next. As discussed previously, the official DRG migration table from $t - 1$ to $t$ considers all actually observed patients in $t - 2$ and groups them into the appropriate DRG according to the system in $t - 1$ and according to the system in $t$. The table then shows how DRGs from $t - 1$ map into DRGs from $t$. For the analysis, I restrict the attention to DRGs with an unchanged patient composition (that is, DRGs for which the migration table shows a one-to-one mapping from $t - 1$ to $t$) as well as a change in kink location from $t - 1$ to $t$.

Table 1.1 shows descriptive statistics for the analysis sample in comparison with the remaining patient population. The analysis sample is comparable to the remaining cases in terms of covariates, but patients in the analysis sample have a longer average length of stay by construction, since in order for a DRG to have a changing kink location it must necessarily have had an average duration of at least 7.5 days at some point (the analysis of transfers later in the paper will provide some sense for whether the results carry over to less serious diseases). The stability of covariates from one year to the next within a DRG for which the kink location changes is discussed and shown in the form of regression analyses at the end of this section.

To provide some sense of how often kink locations change, Table 1.2 shows for how many DRGs the kink location changes from one year to the next in a certain way as well as how many patients these DRGs cover. It is apparent that the kink locations decrease more often than rise due to the secular trend towards shorter stays.

**Graphical Analysis**

I start with an example. Figure 1.11 shows the payoff schedule for DRG H62A in years 2005 and 2006. The kink moves from four to three days. In Figure 1.12, I
present this DRG’s hazard rates for the two years. There is no noticeable increase in
the hazard at three or decrease in the hazard at four days in 2006.

In order to conduct the graphical analysis for a pooled sample of DRGs, I focus
on the most common change in kink location, that is on DRGs for which the kink
location decreases by one day from year \( t \) to year \( t+1 \). I center all observations around
their respective kink location in \( t \) and pool them. Figure 1.13 shows the hazard rates
in \( t \) and in \( t + 1 \) for this pooled sample plotted against days relative to the kink
location in \( t \). If there were meaningful bunching behavior, we would see relatively
higher hazard rates at 0 for year \( t \) (since 0 is that year’s kink) and relatively higher
hazard rates at \(-1\) (which is the kink location in \( t + 1 \)) for year \( t + 1 \). The shape
of the hazard rates, however, looks very similar in \( t \) and \( t + 1 \) except for a tendency
towards higher hazards in \( t + 1 \) in general. This tendency is reflective of the trend
towards shorter stays. The fact that there is notable move of discharges away from 0
and towards \(-1\) provides strong evidence that the hospitals do not adjust treatment
length in response to marginal reimbursement.\(^{18}\)

Figure 1.14 shows the same graph for a tighter time window, specifically with the
data restricted to October until December for year \( t \) and January to March for year
\( t + 1 \). Figure 1.14 supports the conclusions from Figure 1.13, albeit being a little bit
less precise due to the smaller underlying mass of data.\(^{19}\)

One possible identification concern is that doctors might need more than a year
to adjust to a new kink location. In order to test for this possibility, Figure 1.15
shows the evolution of hazard rates from one year to the next for DRGs for which the
kink location did not change. If doctors learned over time to optimize their discharge

\(^{18}\)Note that the hazard rates at \(-3\) drop relative to the hazard rates at \(-2\) by construction, because
if, e.g., a DRG has its kink location at 3 days in \( t \), it will necessarily have a hazard of zero at \(-3\).
But this effect is identical for year \( t \) and year \( t + 1 \) and the hazard rates can be compared directly.

\(^{19}\)In contrast to Figure 1.13, Figure 1.14 features hazard rates in \( t \) that are generally higher than
in \( t + 1 \). This most likely due to month effects. Patients admitted in December typically feature
higher hazard rates - possibly, because major surgeries with a long expected duration in the hospital
are postponed until after Christmas and new year.
behavior with regards to the kink, we would expect the hazard rate at the kink day to rise (relative to the hazard rates for the other days) in year $t+1$ compared to year $t$. The graph shows no indication of such learning behavior.

Regression Analysis

For the econometric analysis, I calculate the hazard rate $hazard_{drg,t,d}$ for each DRG $drg$ for each year $t$ for each possible length of stay $d$. I then estimate the following type of specifications

$$\ln hazard_{drg,t,d} = \delta \cdot \Delta p_{d,drg,t} + \alpha_{drg,d} + \gamma_{drg,t} + \epsilon_{drg,d,t}$$

$\Delta p_{d,drg,t}$ is zero if $d$ is not the kink day for DRG $drg$ in year $t$ and denotes the decrease in marginal reimbursement at the kink (measured in 1,000€) otherwise - for instance, $\Delta p_{d,drg,t} = 3$ for the kink day if the payoff schedule features slope 3,000€ per day to the left of the kink and becomes flat afterwards. Hence, $\delta$ is the percentage change in the hazard rate if at $d$ there is a kink at which marginal reimbursement is decreased by 1,000€. The $\alpha_{drg,d}$ are indicators for each DRG-day-combination allowing for an arbitrary hazard rate pattern across days for each DRG. Note that this implies that there are no functional form assumptions regarding the shape of the hazards across days within a DRG. Instead, $\delta$ is identified purely via how strongly the hazard rate for the same $drg - d$-pair changes from one year to the next when $\Delta p_{d,drg,t}$ changes. That is, the identifying variation comes from changing kink locations for the same DRG from year to year as well as from changes in the size of the jump of marginal reimbursement at the kink for the same DRG from year to year.

The indicators for each DRG-year-pair $\gamma_{drg,t}$ allow for a DRG-specific proportional shift in hazard rates each year. These indicators capture general time effects for each DRG, e.g., a trend towards shorter stays/higher hazards for some DRGs. Note that
for a bunching design with a kink (in contrast to one with a notch) there is no hole in the distribution to be expected to the right of the kink which is why no indicator for the day above the kink day is needed.

All standard errors are bootstrapped with $N = 400$ and clustered at the DRG-level. I convert the estimated parameter $\delta$ to the parameter of interest — the causal effect of cutting marginal reimbursement by 1,000 2013-Euro on length of stay $E \left[ \frac{d(d_i)}{dp} \mid \theta_d > \theta_i > \bar{\theta}_d \right]$ as discussed in the theory section. The conversion takes place inside the bootstrap so that the standard errors are accurately adjusted. The conversion procedure requires the expected mass of patients at the days above the kink day $f(d^* + 1)$, etc. under the hypothetical linear schedule without a kink. I use the observed mass at $d^* + 1$, etc., since if the bunching mass is small the difference is negligible (Kleven 2016). In order to limit the influence of days $d$ that are far away from the kink on the estimation of $\gamma_{drg,t}$, I restrict the range of $d$ over which I estimate the equation to the lowest observed kink location of DRG $drg$ minus ten days and the largest one plus ten days. The results are barely changed by varying this range or not restricting the range at all.

Table 1.3 presents the regression results. Columns one and two present the causal effect of interest for the unweighted and weighted regression equation without the $\gamma_{drg,t}$ fixed effects. Columns three and four repeat these specifications including $\gamma_{drg,t}$. The estimates are similar and close to zero for all four specifications, yet a little more precise if weighted and if one includes the $\gamma_{drg,t}$ fixed effects. The specification reported in column four — weighted and including the $\gamma_{drg,t}$ — is the most precise and implies an increase of 0.019 days in average length of stay when increasing marginal reimbursement by 1,000€. Across the four specifications one can reject an implied causal effect of more than a 0.05 days increase.

Next, I investigate the possibility of downward rigidity in treatment decisions. Specifically, it is imaginable that it is easier for doctors to keep patients a day longer in
response to an increase in kink location than it is to keep them shorter when the kink location goes down. Columns five and six restrict the sample to DRGs that feature a decreasing respectively increasing kink location over time. Note that the samples are not entirely distinct, because some DRGs have increases as well as decreases in kink location over time. Only using the DRGs with kink location decreases, one finds results very consistent with the prior results. Restricting the sample to DRGs with increases shows a somewhat larger effect albeit with reduced precision. But even this specification still allows me to bound the causal effect below 0.12 days.

Lastly, I test whether hospitals learn how to optimize profits in the presence of the DRG system over time. I repeat the specification from column four, but restrict the sample to data from 2010 or later. There is no clear upward jump in the coefficient, suggesting that even after some years of learning the new system the hospitals still do not discharge patients in a profit maximizing manner.

**Bunching Estimates for Less Serious Diseases and Transferred Patients**

A possible concern with this paper’s analysis is that the causal effect is identified only for fairly serious diseases, since the average duration for a DRG must necessarily be above 7.5 days at some point in order for the DRG to possibly have a changing kink location. Transferred patients - who have been excluded from the analysis so far - can shed some light on whether the results carry over to less seriously sick patients, because transfers from or to the hospital obey special reimbursement rules. For transfers, the reimbursement for the hospital is linearly increasing until the rounded average duration from two years prior is reached. Afterwards, the payment schedule becomes flat and identical to the one for non-transferred patients. Since the kink for transfers is located at the rounded average duration measured two years prior, transfers allow me to implement the dynamic bunching design for all DRGs that are comparable from one year to the next and for which the average duration —rounded
and measured two years prior—changes. Importantly, this includes DRGs with fairly small average durations such as two or three days. Table 1.4 presents the results when estimating the main analysis equation \( \ln \text{hazard}_{\text{drg},t,d} = \delta \cdot \Delta p_{d,\text{drg},t} + \alpha_{\text{drg},d} + \gamma_{\text{drg},t} + \epsilon_{\text{drg},d,t} \) using transferred patients and the changing kink locations of their respective payoff schedules. The results are again very close to zero and even more precise than those of the main analysis, suggesting that this paper’s conclusions are not limited to severely sick patients.

**Robustness to Alternative Forms of Agency Frictions**

The bunching design identifies the causal effect of marginal reimbursement on length of stay if the agency friction takes the form modeled in Section 1.3, i.e. if the hospitals can make the medical decision makers take hospital profits into account, albeit to a dampened degree. It is imaginable, however, that the hospitals evaluate their doctors based on some rule of thumb - e.g., it could be that the hospital responds to a decrease in the kink location by asking the doctors to reduce the patients’ average length of stay. In this case, the payment schedule would have an effect on length of stay, yet there would not be any bunching.

In order to investigate whether the kink has effects on hazard rates of patients away from the kink, I again make use of the changing kink locations over time. Specifically, I estimate the following fixed-effect regressions to test whether the kink location has any impact on *average* length of stay:

\[
\text{days}_{i,\text{drg},t} = \beta \cdot \text{kinklocation}_{\text{drg},t} + \alpha_{\text{month}} + \alpha_{\text{drg}} + \sum_{j=1}^{J} \delta_j \cdot (\text{avgduration}_{\text{drg},t-2})^j + \gamma X_{it} + \epsilon_{i,\text{drg},t}
\]

\( \text{days}_{i,\text{drg},t} \) denotes the patient’s (who is in DRG \( \text{drg} \) in year \( t \)) length of stay and \( \text{kinklocation}_{\text{drg},t} \) is the location of the kink of DRG \( \text{drg} \) in year \( t \) (for instance,
kinklocation\textsubscript{drg,t} = 5 in the case of the example DRG in Figure 1.2). \( \alpha_{\text{month}} \) and \( \alpha_{\text{drg}} \) are fixed effects for each month and each DRG.

Recall that the kink location of a DRG is determined by the average location measured two years prior, specifically \( kinklocation_{\text{drg},t} = \max \left( 2, \text{round} \left[ \frac{\text{avgduration}_{\text{drg},t-2}}{3} \right] \right) \). Hence, DRGs for which the kink locations decrease are on a downward trend in length of stay. In order to account for that, I control for polynomials in \( \text{avgduration}_{\text{drg},t-2} \) in most specifications. After controlling for \( \text{avgduration}_{\text{drg},t-2} \) there is still variation in \( kinklocation_{\text{drg},t} \) left due to the rounding. Essentially, controlling for \( \text{avgduration}_{\text{drg},t-2} \) allows to compare the development of length of stay for a DRG for which the kink location changes with another DRG for which the average duration evolved similarly in the past, but for which the kink location does not change because of the rounding. \( X_{it} \) are further controls that I employ in some specifications, specifically hospital-month fixed effects and DRG-specific age effects and indicators for gender.

The results are presented in Table 1.5. The first column shows the result without controlling for the average duration from two years prior. As expected, this specification shows a positive effect of the kink location on length of stay, since any time trend in length of stay for a DRG will also affect the kink location. Controlling for the average duration two years prior, however, moves the coefficient very close to zero and adding further controls does not change the estimates meaningfully. Hence, the kink does neither have an effect on the hazards at the kink nor away from the kink, implying that the results are robust to alternative models of the agency friction.

Recall also that the time series evidence from Figure 1.5 shows no effect of going from a schedule that is increasing in length of stay for most patients to a mostly flat schedule, which —irrespective of how one models the agency friction— further supports the case that the marginal reimbursement for the hospital has no impact on length of stay in Germany.
Kink Location and Covariates

The dynamic bunching design analyzing the hazard rates at the kink locations when the kink locations change did not require the observables and unobservables of a DRG’s patient population to be orthogonal to the DRG’s kink location. Instead, the dynamic bunching design only required that any other determinants of length of stay that are correlated with the kink location do not differentially affect the hazard rates depending on whether the day is a kink day or not.

The fixed effect regressions in the last subsection, however, do require other factors to be orthogonal to the kink location changes. While the robustness of the results to including individual controls already demonstrated this stability, Pischke and Schwandt (2015) show that cofounders should be analyzed as the dependent variable also.

I investigate whether changes in observables are correlated with changes in the kink location by running specifications of the form

\[ \text{covariate}_{i,\text{drg},t} = \beta \cdot \text{kinklocation}_{\text{drg},t} + \alpha_{\text{drg}} + \alpha_{\text{month}} + \sum_{j=1}^{J} \delta_j \cdot (\text{avgduration}_{\text{drg},t-2})^j + \epsilon_{i,\text{drg},t} \]

with \( \text{covariate}_{i,\text{drg},t} \) denoting the the observable of interest for patient \( i \) with DRG \( \text{drg} \) in year \( t \). The other variables are defined as in the regressions in the last subsection.

Tables 1.6 to 1.9 show the results for age, gender, the number of diagnoses and the number of procedures. None of these regressions show a significant association between the covariate and the kink location after controlling for the fixed effects and the polynomials in the average duration measured two years prior.

1.6 Discussion

Contrast to the recent U.S. evidence
The results presented in the last section stand in a sharp contrast to recent evidence from the United States. Einav et al. (2017), Eliason et al. (2016) and Kim et al. (2015) study the effects of a jump in the Medicare reimbursement for post-acute hospitals after patients pass a certain length of stay threshold. These papers find a pronounced excess mass of discharges right after the payment jumps.

I used data from Einav et al. (2017) to construct an estimate that can be compared quantitatively to my paper’s numbers, i.e. an estimate of how length of stay responds to a cut in the marginal reimbursement per day by 1,000€. The point estimate is a reduction in length of stay by 0.34 days, seven times more than the upper bound of my paper’s confidence interval. The contrast is especially remarkable, since hospital care —as measured by the number of hospital beds per capita, the number of hospitalizations and average length of stay— is generally much more extensive in Germany than in the U.S. (the U.S. has had a comparatively —by OECD standards— low utilization of hospital services throughout the last decades, for the recent OECD numbers see Section 1.2). The larger baseline amount of hospital services suggests lower medical returns to care at the margin in Germany than in the U.S. and, hence, if anything, more ability for the German health care providers to adjust care in response to financial incentives.

I interviewed two doctors who work or recently worked in German hospitals. Both of them were aware of the kink in the schedule, both of them realized that a discharge on the kink day is particularly financially attractive for the hospitals and both of them confirmed that they were aware of the kink locations for often occurring DRGs. Moreover, getting to know a DRG’s kink location is easy by using the coding software or —if they have one working with them— by asking the medical coder.

---

20I use the probability mass distribution of discharges across days reported in Einav et al. (2017) for before and for after the jump in the Medicare reimbursement was introduced. I compare the probability mass distribution before and after the introduction in order to estimate the effect of the jump in reimbursement on length of stay. For details, see appendix 1.D.

21The interviews were conducted August 25th and October 19th 2017 via phone.
While one doctor recalled a general information session that informed about DRGs and the payoff structure for the hospitals, none of them reported any pressure from the hospital in their day-to-day activities. One of the interviewed doctors denied taking the kink location into account in her decision making at all. The second one stated “It [the kink] is present in my mind and is being communicated by the hospital. But it’s not like they [the hospital] are doing enough so that I identify so much with them [the hospital]. If it were like a family business then maybe it would be different. [...] I do what makes sense medically. I would take the UGV [the kink] into account only if otherwise I were completely undecided whether to discharge today or tomorrow.”

This anecdotal evidence stands in a sharp contrast to stories from the U.S. (originally reported in an article by Weaver et al. 2015 in the Wall Street Journal (WSJ) and cited in Eliason et al. 2016) that suggest a strong pressure to discharge patients after the threshold at which the hospital’s Medicare payment jumps is reached. The WSJ article describes “meetings in which hospital staffers would discuss plans for each patient at the facility—armed with printouts from a computer tracking system that included, for each patient, the date at which reimbursement would shift to a higher, lump-sum payout.” Reports also allege that “doctors, pressured by hospital administrators, sometimes ordered extra care or services intended in part to retain patients until they reached their thresholds, or discharged those who were costing the hospitals money regardless of whether their medical conditions had improved.”

This more active role of administrators in shaping treatment decisions in U.S. hospitals explains the much larger effect of financial incentives for hospitals in the U.S. than in Germany. The influence of administrators also explains why financial incentives for hospitals have an effect in the U.S. at all, despite the fact that in the U.S. —as in Germany— the doctors face financial incentives that are at odds with those of the hospital and despite it being legally difficult for U.S. hospitals to directly

---

22Translation by this paper’s author.
pass on their incentives to the doctors, since it is prohibited for any hospital or critical access hospital from knowingly making a payment directly or indirectly to a doctor as an inducement to reduce or limit services to Medicare or Medicaid beneficiaries under the doctor’s care (Alexander 2017).

Since other institutional features—discussed in Section 1.2—such as patient incentives and liability risk cannot explain the observed differences between the U.S. and Germany, I interpret my findings as the agency friction $\lambda_d$ from the model being smaller in Germany than in the U.S., i.e. as German hospitals being less able to make the medical decision makers take hospital profits into account. The importance of agency frictions in hospitals is also a key theme in recent work by Sacarny (2016) who finds that U.S. hospitals differ drastically in their ability to make doctors code diagnoses in a way that benefits the hospital financially.

**Are German hospital managers trying to maximize profits?**

The preceding section demonstrates that the hospital management has less influence on treatment decisions in Germany than in the U.S., rendering financial incentives that apply at the hospital level an ineffective policy tool. An interesting question that follows is whether German hospital managers just have no interest in profit-maximizing or whether it is the cultural and institutional context that prevents them from doing so.

In order to distinguish between these two possibilities, I test whether hospitals maximize profits in settings in which the culture and the institutions might be less of a hurdle. Specifically, I make use of the fact that the hospital’s reimbursement for treating a newborn drops discontinuously if the birth weight exceeds 1,500g. Hence, downward manipulation of the birth weight of newborns just above this cutoff can increase hospital profits. The hospitals might be better able to respond to these financial incentives than to the incentives studied in the main part of this paper for
two reasons. Firstly, the birth weight is typically recorded by the midwives who might be easier to control for the hospital than the doctors and, secondly, manipulating the birth weight in a way that it drops below 1,500g is not a manipulation of the newborn’s treatment which the doctors may feel obligated to decide about entirely based on medical grounds. Birth weight manipulation in Germany was also studied in Jürges and Köberlein (2015) and Reif et al. (2017). The latter paper demonstrates that the manipulation of the birth weight does not lead to different treatment choices later.

Figure 1.22 in appendix 1.C shows the frequency of births plotted against the birth’s birth weight. The graph shows strong evidence of manipulation behavior with a clear excess mass of births with birth weight just below beneath 1,500g as well as missing mass just above 1,500g.

To measure the degree to which hospitals manipulate the birth weight, I calculate for each hospital the number of newborns with a birthweight between 1,480g and 1,499g and divide it by the number of newborns in the hospitals with a birthweight between 1,460g and 1,519g. If there were no manipulation, the resulting fraction —denote it $\sigma_h$— should average about 33%. In the data, $\sigma_h$ averages 69.3%, implying drastic manipulation behavior by the German hospitals.

Hence, just like their U.S. counterparts, the German hospital managers are striving for profit-maximization, but they are constrained in achieving their goals in areas with large agency frictions such as when deciding about the actual treatment of the patients. These agency frictions render financial incentives for hospitals a poor policy tool, since to be truly cost-saving the incentives have to affect real treatment choices.

**Implications for the German reform and contrast to the Medicare reform**

This paper bounds the causal effect of reducing marginal reimbursement for another day by 1,000€ below 0.05 days. As discussed in Section 1.2, prior to 2004
Germany used a cost-based per diem system in about 80% of cases, that is hospitals were reimbursed using a linear schedule in length of stay with hospital- and department-specific slopes that depended on the hospital’s historical costs. While I do not have data on these hospital- and department-specific slope parameters, the average slope must have been smaller than 1,000€ per day, because —given the average length of stay before 2004— the inflation-adjusted pre-2004 total yearly hospital revenue would otherwise have been bigger than the current level while in reality it was smaller. Hence, in 2004 Germany switched from a system that paid on average less than 1,000€ per day to the current system which is mostly flat in length of stay. Therefore, the estimated 0.05 days provides an upper bound for the effect of the 2004 reform on length of stay. Thus, German politicians —like the minister of health quoted in the introduction— took false credit for the fall in length of stay after 2004, since it apparently was just a continuation of the previous trend and not a causal effect of the reform.

The 1983 Medicare reform is widely perceived as having reduced length of stay. While the exact size of the reform’s impact is contested, the discussion evolves around magnitudes that are an order of magnitude larger than what my paper attributes to the German reform. For instance, Russell (1989) states:

Historically, length of stay for the elderly had declined steadily, drifting slowly downward from 13.8 days in 1968 to 10.1 days in 1982. The declines in the two years before prospective payment were unusually steep by historical standards, but the decline between 1983 and 1984, when the average dropped by nearly a day, was unprecedented, ample reason to suspect that prospective payment was the cause.

The Medicare reform moved to prospective pay coming from a fee for service system, i.e. a system that pays for each individual provided service. A fee for service system

\[23\] The remaining 20% were reimbursed using a fixed prospective payment, see Theilen 2004
does provide the hospitals with a financial incentive to increase length of stay, since additional services can be provided (including the service of providing a bed, etc. for another night). Since pre-1983 Medicare did pay per service and not per day, my estimates do not directly apply. But a rough back-of-envelope calculation suggests that my estimates from Germany would imply a smaller effect for the 1983 reform than what has been observed:

I approximate the 'reimbursement per day' for the pre-1983 Medicare system by dividing Medicare’s total expenditures for hospitals by the total number of Medicare hospital days. I describe the procedure in appendix 1.E. The calculation provides an estimate of the average reimbursement per day, but an overestimate of the marginal reimbursement for keeping a patient another day at the end of her spell, since the costs-weighted total amount of services provided normally increases less than proportionally with the patient’s length of stay (Ishak et al. 2012).

I find that Medicare paid approximately 750€ per hospitalization day in 1984, implying that the 1983 reform—which reduced the marginal reimbursement to zero for most patients—cut the marginal reimbursement per day by less than 1,000€. Hence, my 0.05 days upper bound for the effects of a cut of marginal reimbursement by 1,000€ stands in sharp contrast to what conventional wisdom attributes to the 1983 reform, again supporting the case that incentives for hospitals are more effective in the U.S. than in Germany.

**Implications for policy and future research**

My results suggest that future research should investigate whether countries with institutions and cultural norms as in Germany could improve welfare by making hospital reimbursement depend more strongly on length of stay again. Paying hospitals in a way that is more closely tied to the actual incurred costs would better financially compensate hospitals that draw a lot of patients who are comparatively
sick conditional on their DRG. The closer tie between costs and reimbursement could have two advantages.

First, compensating hospitals better for patients who are relatively expensive conditional on their DRG would reduce the incentive for the hospitals to discriminate against such patients along the admission margin. Such discrimination is a concern despite this paper’s results, since the hospital management might have more ability to manipulate admissions than actual treatment once admitted.\textsuperscript{24}

Second, paying hospitals in a way that corresponds more closely to the actual incurred costs might raise welfare, because —given this paper’s results— such a policy change would barely affect total treatment volume, but it would reduce the hospitals’ financial risk of drawing patients that are comparatively expensive conditional on their DRG. The financial risk would be reduced across hospitals (i.e. the risk of a hospital drawing comparatively sick patients permanently due to, e.g., its location) and within hospitals (i.e. the risk of a hospital drawing many comparatively sick patients within a particular year).

The within-hospital financial risk is non-negligible for smaller hospitals, since fluctuations in length of stay —conditional on the DRG composition— cancel imperfectly over the course of a year.\textsuperscript{25} Reducing the financial risk for hospitals would —if hospital owners are risk-averse—allow to reduce equilibrium hospital profits without inducing hospital exit.

For future policy in Germany and countries with similar institutions, my results also suggest that politicians who want to reduce health care costs should consider changing the financial incentives for the doctors directly. In the U.S., various pilot programs test the effect of paying the doctors working in hospitals in a way that

\textsuperscript{24}Alexander (2016) provides evidence of such discrimination at the admission margin in the U.S.

\textsuperscript{25}For instance, the average length of stay at the hospital-year level, residualized for hospital and DRG-year fixed effects, has a standard standard deviation of 0.78 days for the smallest quartile of hospitals (for the largest quartile it is only 0.24), nearly a tenth of the average length of stay.
encourages costs saving behavior — see, for instance, the evaluation of such a program in Alexander (2016).

The experiences from the 1983 Medicare reform have shaped the way researchers and politicians around the globe think about designing incentives for health care providers. For future research, this paper advises caution when extrapolating reduced form effects from one cultural and institutional setting to another.

References


Coey, D.: 2015, Physicians’ financial incentives and treatment choices in heart attack management, Quantitative Economics 6(3), 703–748.


The parameters of the payment scheme are — as the DRG definitions — based on the hospital cost data from two years before. The payment increases linearly until a third of the average length of stay (rounded and measured two years prior) of all patients in this DRG is reached (but at least until day 2 is reached). The slope is determined by dividing average variable costs (that is, total costs excluding costs of major procedures, e.g., bypass surgery) of all patients with this DRG by the number of days at which the kink occurs (again, costs measured two years prior). After the kink, the payment schedule remains flat until the average plus two times the standard deviation of the length of stay of all patients with this DRG two years prior is reached. From then on it increases again linearly.
Example of a payment schedule. The graph shows the pay for a patient with DRG I51Z in 2005 as a function of number of days in the hospital. DRG I51Z is for ‘other procedures at the hip joint or femur, without major complications’. Pay does shift proportionally vertically across locations - the graph corresponds to the average proportional shift factor for the state of Hamburg.
For each DRG, I calculate the change in slope at the lower kink in 2013-Euro, i.e. how much money the hospital loses by discharging the day before the kink day instead of on the kink day. The plot is a histogram of this measure across all DRGs and all years 2005-2013.
For each DRG I calculate the ratio of the slope to the left of the lower kink and the amount that is paid to the hospital in the flat part of the schedule - i.e. the share of revenue that is lost by discharging the patient a day earlier than the kink day instead of on the kink day. The graph shows a histogram of this measure across all DRGs and all years 2005-2013.
The averages include cases not covered in the DRG system such as psychiatric cases. The source is not the OECD and uses different definitions, which is why the numbers differ from the OECD numbers I use in the international comparisons in the paper.
Figure 1.6: Histogram Kink Locations

Shows the share of DRGs with the respective kink location (for the lower kink) for years 2005-2013. E.g. the example DRG in Figure 1.2 has the kink location 5 days.
Figure 1.7: Histogram Length of Stay

Shows the share of patients staying 1, 2, 3 etc days in the hospitals across all years 2005-2013.
Figure 1.8: Histogram of Length of Stay Relative to the Patient’s Kink Location

Shows the share of patients getting discharged earlier than the lower kink of their respective DRG, between the two kinks or after the right kink of their respective DRG. Graph is a histogram for all patients in the data, i.e. 2005-2013.
Hazard rates for the example DRG I51Z with payment schedule shown in Figure 1.2. Graphs show the hazard rates for hospital discharge in year 2005 for DRG I51Z after 0, 1, 2, etc days in the hospital.
DRG I51Z is for 'other procedures at the hip joint or femur, without major complications'
I normalize the length of stay for each patient with respect to her DRG's lower kink location. E.g., if a patient stays 3 days and her kink is at 4 days, the normalized length of stay is -1. I then pool all cases and calculate the hazard rates w.r.t. the normalized length of stay. I then plot these hazard rates against the normalized length of stay. I restrict the sample to DRGs with kink location of at least 6 days. Graph looks similar for alternative restrictions.
Figure 1.11: Payment Schedule Change - Example DRG H62A for Which the Kink Location Goes Down by One Day

Example of how payment schedules change from one year to the next for the same DRG. The graph shows the pay to the hospital for a patient with DRG H62A in 2005 and 2006 as a function of the number of days in the hospital. The kink location changes from 4 to 3 days. DRG H62A is for 'Diseases of the pancreas, except for malignant neoformation with acute pancreatitis'. Pay does shift proportionally vertically across locations - the graph corresponds to the average proportional shift factor for the state of Hamburg.
Figure 1.12: Hazard Rates Around Kink from One Year to the Next - Example DRG H62A for Which the Kink Location Goes Down by One Day

Shows the hazard rates for the example DRG H62A in 2005 and 2006 - see Figure 1.11 for this DRG’s payment schedule in 2005 and 2006 (the kink location changes from 4 days in 2005 to 3 days in 2006). The graphs plots the hazard rates for discharge from hospital in year 2005 and year 2006 after 0, 1, 2, etc days in the hospital.

DRG H62A is for 'Diseases of the pancreas, except for malignant neoformation with acute pancreatitis'
Figure 1.13: Hazard Rates Around Kink from One Year to the Next if the Kink Location Goes Down by One Day - Pooled Sample

I restrict the sample to DRGs that are comparable from t to t+1 and for which the kink point went down one day from t to t+1. For each patient, I normalize length of stay by the patient’s DRG’s kink location in t (e.g., normalized length of stay is -1 if she stayed 4 days, but her DRG’s kink in t is at 5 days). I then pool the sample and calculate hazard rates w.r.t. the normalized length of stay. That is, the hazard rate at 0 is the hazard rate at the year t’s kink and at -1 is the hazard rate at year t+1’s kink.

Note that the hazard rates to the left of the kink do not necessarily behave smoothly, because e.g. a DRG with kink in t at 3 days has a hazard of 0 at -3 by construction. But the composition effects are identical for t and t+1, so the comparison is still valid.
Figure 1.14: Hazard Rates Around Kink from One Year to the Next if the Kink Location Goes Down by One Day- Pooled Sample - October until March

(Identical to Figure 1.13, except restricted to the narrower time window from October of year t to March of year t+1)

I restrict the sample to DRGs that are comparable from t to t+1 and for which the kink point went down one day from t to t+1. For each patient, I normalize length of stay by the patient’s DRG’s kink location in t (e.g., normalized length of stay is -1 if she stayed 4 days, but her DRG’s kink in t is at 5 days). I then pool the sample and calculate hazard rates w.r.t. the normalized length of stay. That is, the hazard rate at 0 is the hazard rate at the year t’s kink and at -1 is the hazard rate at year t+1’s kink.

Note that the hazard rates to the left of the kink do not necessarily behave smoothly, because e.g. a DRG with kink in t at 3 days has a hazard of 0 at -3 by construction. But the composition effects are identical for t and t+1, so the comparison is still valid.
(Identical to Figure 1.13, except restricted to DRGs for which the kink location stays the same from year \( t \) to year \( t+1 \))

I restrict the sample to DRGs that are comparable from \( t \) to \( t+1 \) and for which the kink point does not change \( t \) to \( t+1 \). For each patient, I normalize length of stay by the patient’s DRG’s kink location (which is the same in \( t \) and \( t+1 \)). E.g., normalized length of stay is -1 if she stayed 4 days, but her DRG’s kink is at 5 days). I then pool the sample and calculate hazard rates w.r.t. the normalized length of stay.

Note that the hazard rates to the left of the kink do not necessarily behave smoothly, because e.g. a DRG with kink at 3 days has a hazard of 0 at -3 by construction. But the composition effects are identical for \( t \) and \( t+1 \), so the comparison is still valid.
Table 1.1: Summary Statistics Analysis Sample vs Remaining Sample

<table>
<thead>
<tr>
<th></th>
<th>Analysis Sample</th>
<th>Remaining Cases</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of Stay</td>
<td>14.05 (16)</td>
<td>6.9 (7.88)</td>
<td>7.15 (0.0091)</td>
</tr>
<tr>
<td>Age</td>
<td>52.2 (26.76)</td>
<td>53.58 (25.47)</td>
<td>-1.38 (0.0293)</td>
</tr>
<tr>
<td>Share Female</td>
<td>0.5 (0.5)</td>
<td>0.54 (0.5)</td>
<td>-0.04 (0.0006)</td>
</tr>
<tr>
<td>Year of Admission</td>
<td>2009 (2.45)</td>
<td>2009.11 (2.58)</td>
<td>-0.11 (0.003)</td>
</tr>
<tr>
<td>Month of Admission</td>
<td>6.47 (3.41)</td>
<td>6.43 (3.46)</td>
<td>0.04 (0.004)</td>
</tr>
<tr>
<td>N</td>
<td>761505</td>
<td>130 million</td>
<td></td>
</tr>
</tbody>
</table>

The table presents summary statistics for the analysis sample of patients (i.e. restricted to DRGs that are comparable from one year to the next and for which the kink location changes) as well as for the remaining observations. For month of admission 1 corresponds to January and 12 to December.

Table 1.2: How Often Do Kink Locations Change?

<table>
<thead>
<tr>
<th>Kink in t</th>
<th>Kink in t+1</th>
<th>DRGs</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>15</td>
<td>23 014</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>33</td>
<td>153 470</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>8</td>
<td>13 539</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>34</td>
<td>169 243</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>7</td>
<td>3 957</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>23</td>
<td>68 280</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>5</td>
<td>11 204</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>2</td>
<td>785</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>11</td>
<td>25 407</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>11</td>
<td>4 728</td>
</tr>
</tbody>
</table>

The sample is restricted to DRGs that are comparable from one year to the next and for which the kink location changes. For each combination of kink location in t and kink location in t+1 the table reports how many DRGs feature this change in kink location from one year to the next and how many patients are grouped into such a DRG. I restrict it to DRGs with a kink location of at most 6 in t. Patients as well as DRGs can appear several times - e.g. because a DRG has kink location 2 in t, then 3 in t+1 and then 2 again in t+2. In that case the DRG (and the patients grouped into this DRG) are counted twice in t+1: once they appear in the 2 to 3 row and once in the 3 to 2 row.
Table 1.3: Causal Effect of Increasing Marginal Reimbursement per Day in Hospital by 1,000€ on Length of Stay

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>all</th>
<th>all</th>
<th>all</th>
<th>increases</th>
<th>decreases</th>
<th>≥2010</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dep. Var.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Causal Effect</td>
<td>-0.004 (0.017)</td>
<td>0.026** (0.01)</td>
<td>0.015 (0.01)</td>
<td>0.019* (0.009)</td>
<td>0.066* (0.027)</td>
<td>0.011 (0.008)</td>
<td>0.028 (0.018)</td>
</tr>
<tr>
<td>Year-DRG-FE</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>weighted</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Patients</td>
<td>6149229</td>
<td>6149229</td>
<td>6144604</td>
<td>6144604</td>
<td>791623</td>
<td>5762258</td>
<td>2623847</td>
</tr>
<tr>
<td>Clusters</td>
<td>107</td>
<td>107</td>
<td>100</td>
<td>100</td>
<td>27</td>
<td>87</td>
<td>48</td>
</tr>
</tbody>
</table>

'Causal Effect' refers to the effect of increasing marginal pay for another day in the hospital by 1,000 2013-Euro. Standard errors are clustered at the DRG level using 400 bootstrap draws. The sample is restricted to DRGs that are comparable from one year to the next and for which the kink location changes. Moreover, Column 5 is restricted to DRGs that feature a decrease in kink location over time, while column 6 is restricted to those featuring an increase. Column 7 only uses post 2010 data. Some DRGs experience both, decreases and increases, which is why the number of clusters in columns 5 and 6 does not add up to the number in columns 1 to 4. Weighting refers to weighting by the number of patients still present in the hospital that day. pvalues: * < .05 ** < .01
Table 1.4: Causal Effect of Increasing Marginal Reimbursement by 1,000€ per Day in Hospital on Length of Stay - for Transfers

<table>
<thead>
<tr>
<th>Dep. Var.</th>
<th>Length of Stay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Causal Effect</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Standard Error</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Year-DRG-FE</td>
<td>no</td>
</tr>
<tr>
<td>weighted</td>
<td>no</td>
</tr>
<tr>
<td>Patients</td>
<td>1205114</td>
</tr>
<tr>
<td>Clusters</td>
<td>140</td>
</tr>
</tbody>
</table>

Causal Effect refers to the effect of increasing marginal pay for another day in the hospital by 1,000 2013-Euro. Standard errors are clustered at the DRG level. The sample is restricted to transfers and to DRGs that are comparable from one year to the next and for which the kink location - in the case of transfers, the average duration two years prior - changes. Weighting refers to weighting by the number of patients still present in the hospital that day. pvalues: * < .05 ** < .01

Table 1.6: Fixed Effect Regression for the Effect of the Kink Location on Covariates - Age

<table>
<thead>
<tr>
<th>Dep. Var.</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kink location</td>
<td>-0.036 (0.126)</td>
</tr>
<tr>
<td>Month FE</td>
<td>no</td>
</tr>
<tr>
<td>DRG FE</td>
<td>yes</td>
</tr>
<tr>
<td>avg duration linear</td>
<td>no</td>
</tr>
<tr>
<td>avg duration quadr.</td>
<td>no</td>
</tr>
<tr>
<td>Observations</td>
<td>929376</td>
</tr>
<tr>
<td>Cluster</td>
<td>147</td>
</tr>
<tr>
<td>Mean Dep.</td>
<td>51.83</td>
</tr>
</tbody>
</table>

Standard errors are clustered at the DRG level. Outcome variable is age. 'avg duration linear' and 'avg duration quadr.' refer to a linear and a quadratic term in the average duration from two years prior. 'Mean Dep.' refers to the mean of the dependent variable. pvalues: * < .05 ** < .01
Table 1.5: Fixed Effect Regression for the Effect of the Kink Location on Length of Stay in Days

<table>
<thead>
<tr>
<th>Dep. Var.</th>
<th>Length of Stay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kink location</td>
<td>0.203 (0.111)</td>
</tr>
<tr>
<td></td>
<td>-0.037 (0.136)</td>
</tr>
<tr>
<td></td>
<td>-0.003 (0.126)</td>
</tr>
<tr>
<td></td>
<td>0.01 (0.1)</td>
</tr>
<tr>
<td></td>
<td>-0.053 (0.092)</td>
</tr>
<tr>
<td>Month FE</td>
<td>yes</td>
</tr>
<tr>
<td>Month-Hospital FE</td>
<td>no</td>
</tr>
<tr>
<td>DRG FE</td>
<td>yes</td>
</tr>
<tr>
<td>avg duration linear</td>
<td>no</td>
</tr>
<tr>
<td>avg duration quadr.</td>
<td>no</td>
</tr>
<tr>
<td>Control set 1</td>
<td>no</td>
</tr>
<tr>
<td>Control set 2</td>
<td>no</td>
</tr>
<tr>
<td>Observations</td>
<td>929376</td>
</tr>
<tr>
<td></td>
<td>929376</td>
</tr>
<tr>
<td></td>
<td>929376</td>
</tr>
<tr>
<td></td>
<td>904527</td>
</tr>
<tr>
<td></td>
<td>904527</td>
</tr>
<tr>
<td>Cluster</td>
<td>147</td>
</tr>
<tr>
<td>Mean Dep.</td>
<td>14.76</td>
</tr>
</tbody>
</table>

Standard errors are clustered at the DRG level. Outcome variable is length of stay (winsorized at the 99th percentile). 'avg duration linear' and 'avg duration quadr.' refer to a linear and a quadratic term in the average duration from two years prior. Control set 1 covers DRG-specific indicators for gender and slopes for age. Control set 2 covers number of diagnoses and number of procedures. 'Mean Dep.' refers to the mean of the dependent variable. pvalues: * < .05 ** < .01
Table 1.7: Fixed Effect Regression for the Effect of the Kink Location on Covariates - Gender

<table>
<thead>
<tr>
<th>Dep. Var.</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kink location</td>
<td>0 (0.002)</td>
</tr>
<tr>
<td>Month FE</td>
<td>no</td>
</tr>
<tr>
<td>DRG FE</td>
<td>yes</td>
</tr>
<tr>
<td>avg duration linear</td>
<td>no</td>
</tr>
<tr>
<td>avg duration quadr.</td>
<td>no</td>
</tr>
<tr>
<td>Observations</td>
<td>929376</td>
</tr>
<tr>
<td>Cluster</td>
<td>147</td>
</tr>
<tr>
<td>Mean Dep.</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Standard errors are clustered at the DRG level. Outcome variable is an indicator for female. 'avg duration linear' and 'avg duration quadr.' refer to a linear and a quadratic term in the average duration from two years prior. 'Mean Dep.' refers to the mean of the dependent variable. pvalues: * < .05 ** < .01

Table 1.8: Fixed Effect Regression for the Effect of the Kink Location on Covariates - Number Diagnoses

<table>
<thead>
<tr>
<th>Dep. Var.</th>
<th>Diagnoses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kink location</td>
<td>-0.19** (0.033)</td>
</tr>
<tr>
<td>Month FE</td>
<td>no</td>
</tr>
<tr>
<td>DRG FE</td>
<td>yes</td>
</tr>
<tr>
<td>avg duration linear</td>
<td>no</td>
</tr>
<tr>
<td>avg duration quadr.</td>
<td>no</td>
</tr>
<tr>
<td>Observations</td>
<td>929376</td>
</tr>
<tr>
<td>Cluster</td>
<td>147</td>
</tr>
<tr>
<td>Mean Dep.</td>
<td>6.72</td>
</tr>
</tbody>
</table>

Standard errors are clustered at the DRG level. Outcome variable is the number of diagnoses. 'avg duration linear' and 'avg duration quadr.' refer to a linear and a quadratic term in the average duration from two years prior. 'Mean Dep.' refers to the mean of the dependent variable. pvalues: * < .05 ** < .01
## Table 1.9: Fixed Effect Regression for the Effect of the Kink Location on Covariates-Number Procedures

<table>
<thead>
<tr>
<th>Dep. Var.</th>
<th>Procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kink location</td>
<td>0.054 (0.037)</td>
</tr>
<tr>
<td>Month FE</td>
<td>no</td>
</tr>
<tr>
<td>DRG FE</td>
<td>yes</td>
</tr>
<tr>
<td>avg duration linear</td>
<td>no</td>
</tr>
<tr>
<td>avg duration quadr.</td>
<td>no</td>
</tr>
<tr>
<td>Observations</td>
<td>929376</td>
</tr>
<tr>
<td>Cluster</td>
<td>147</td>
</tr>
<tr>
<td>Mean Dep.</td>
<td>6.49</td>
</tr>
</tbody>
</table>

Standard errors are clustered at the DRG level. Outcome variable is the number of procedures. ‘avg duration linear’ and ‘avg duration quadr.’ refer to a linear and a quadratic term in the average duration from two years prior. ‘Mean Dep.’ refers to the mean of the dependent variable. pvalues: * < .05 ** < .01

## 1.A Measurement of Length of Stay

In the main text, I argue that the measurement error problem for the billed number of days that is introduced by successful audits is less severe for cases from the end of the calendar year. The reason for this seasonality is because the data for each year is collected on March 31 of the following year. Only if the audit has been completed at that point in time, it will change the billed number of days in the data. Since the median time until an audit completion is over 3 months, the share of completed audits is much smaller for December cases than for patients admitted in January. Therefore, the billed number of days for December cases is often times still pre-audit and identical to the difference between discharge and admission date.

Figure 1.16 shows the share of cases for which the billed number of days is smaller than the difference between discharge and admission date. The share of cases with diverging numbers is clearly higher for cases from early in the year than for cases from later in the year, supporting the case that audits are less of a problem later for data from later in the calendar year.
The graph shows (separately by admission month) the share of cases for which the billed number of days in the data deviates from the difference between discharge and admission date.

1.B Generalized Model

Summary

Here I extend the model from the main text with risk-averse hospitals as well as the possibility for audits by health insurers. The key difference in the results is that the causal effect of changing marginal pay under these generalized assumptions is identified for a reform that changes marginal pay per day while adjusting a fixed payment component to the hospital in a way that keeps hospital profits constant in equilibrium (i.e. taking into account that the hospitals will respond to the changed marginal pay as well as to the changed fixed payment component).
Setup

The hospital admits a continuum of patients of type $\theta_i$ who stay $d_i$ days and enjoy health benefit $h(d_i, \theta_i)$ which is concave in the number of days in the hospital, i.e. $\frac{\partial^2 h(d_i, \theta_i)}{\partial d_i^2} < 0$. Since there are no functional form assumptions on how $\theta_i$ affects $h(d_i, \theta_i)$, one can assume a uniform distribution $\theta_i \sim U[0, 1]$ without loss of generality - i.e. $\theta_i$ is patient $i$’s quantile in the distribution of patient sickness. Sicker patients benefit more from staying in the hospital for longer. That is, $\frac{\partial^2 h(d_i, \theta_i)}{\partial \theta_i \partial d_i} > 0$.

The hospital receives payment $P(d_i)$ and incurs costs $C(d_i)$. The hospital gains utility from profits and from its patients’ health (either because of an intrinsic concern for their patients’ health or because they fear lawsuits or reputational costs if patients are mistreated). The hospital is potentially risk-averse and values profits relative to patient health benefits according to $\lambda_h U(profits)$ with $\lambda_h \geq 0$ and $U'' \leq 0$. Since in practice, the medical personnel and not the hospital shareholders make the discharge decision, the hospital faces an agency problem in implementing its objective function. I model this agency problem in the form of parameter $0 \leq \lambda_d \leq 1$ which dampens the degree to which profits are taken into account.

Moreover, the health insurers audit bills with probability $\gamma \in (0, 1)$ and change the billed number of days to a patient-specific $\tilde{d}(\theta_i)$ with $\tilde{d}'(\theta_i) > 0$. Hence, expected audit costs for patient type $\theta_i$ are $\gamma \left[ P(d_i) - P\left(\tilde{d}(\theta_i)\right)\right]$.

That is, the patients’ length of stay is determined by the solution to

$$\max_{\{d_i(\theta_i) \in \mathbb{N}\}} \lambda_d \lambda_h \int_i U \left( P(d_i) - C(d_i) - \gamma \left[ P(d_i) - P\left(\tilde{d}(\theta_i)\right)\right]\right) + \int_i h(d_i, \theta_i)$$

If the agency problem were modeled in a different way, the bunching design would not necessarily identify the causal effect of interest anymore, because the kinks might have effects on hazard rates away from the kink point. I discuss this point in more
detail in the results section and provide evidence against such effects of the kink on hazard rates away from the kink.

Note also that changes in admission and coding behavior - while interesting subjects to study in their own right - do not threaten the validity of this paper’s findings. If anything, adjustments in coding and admission behavior would lead to an overestimate of the bunching mass in my setting. This is because the incentive to deny admission or to upcode to a different diagnosis with a different kink location is smallest for those patients who would otherwise be discharged on the profit maximizing kink day.

**Optimal Hospital Behavior**

I assume that $P(d_i), C(d_i)$ and $h(d_i, \theta_i)$ are shaped such that the objective function is globally concave. At baseline, consider a linear payment schedule $P_{\text{baseline}}(d_i) = \bar{p} + p \cdot d_i$.

Optimization than amounts to choosing cutoff values for $\theta_i$ determining which patient types are kept for how many days. $\bar{\theta}_{d \text{ baseline}}$ denotes the highest $\theta_i$ for which the patient stays $d$ days under the baseline schedule. A patient with a $\theta_i$ just above $\bar{\theta}_{d \text{ baseline}}$ would stay $d + 1$ while a patient with a $\theta_i$ just beneath $\bar{\theta}_{d \text{ baseline}}$ would stay $d$ days. The cutoff values defining the range of patients who are discharged on day $d^*$ are implicitly defined by equations

$$
\lambda_d \lambda_h U'(\pi) [C(d^* + 1) - C(d^*) - p (1 - \gamma)] = h \left(d^* + 1, \bar{\theta}_{d^* \text{ baseline}}\right) - h \left(d^*, \bar{\theta}_{d^* \text{ baseline}}\right)
$$
\[
\lambda_d \lambda_h U'(\pi) [C(d^*) - C(d^* - 1) - p(1 - \gamma)] \\
= h\left(d^*, \bar{\theta}_{d^* - 1}\right) - h\left(d^* - 1, \bar{\theta}_{d^* - 1}\right)
\]

with \(\pi\) denoting hospital profits in equilibrium.

That is, for patient type \(\bar{\theta}_{d^*}\) the hospital is just indifferent between the net profit valued with \(\lambda_d \lambda_h\) of keeping her \(d^* + 1\) instead of \(d^*\) days and the net health benefit it would bring to the patient. A patient with \(\theta_i\) a little bigger than \(\bar{\theta}_{d^*}\) would be kept \(d^* + 1\) days, since her health benefit of staying another day is higher than for the \(\bar{\theta}_{d^*}\) patient. Similarly, for patient type \(\bar{\theta}_{d^* - 1}\) the hospital is indifferent between the marginal health benefit of keeping her \(d^*\) instead of \(d^* - 1\) days and the profit impact. Note that the individual patient has no impact on hospital profits.

Now consider the policy experiment of interest, reducing marginal reimbursement by \(\Delta p > 0\) throughout the schedule, i.e. \(P^{\text{reform}}(d_i) = \tilde{p} + (p - \Delta p) \cdot d_i\) with \(\Delta p > 0\), while adjusting the fixed payment component such that hospital profits remain at \(\pi\) in equilibrium. That is, \(\tilde{p}\) is chosen such that the profits remain at \(\pi\) taking into account the fact that the patients’ length of stay will change with this change in the payment schedule. The cutoff values are now defined by

\[
\lambda_d \lambda_h U'(\pi) [C(d^* + 1) - C(d^*) - (p - \Delta p)(1 - \gamma)] \\
= h\left(d^* + 1, \bar{\theta}_{d^*}\right) - h\left(d^*, \bar{\theta}_{d^*}\right)
\]

\[
\lambda_d \lambda_h U'(\pi) [C(d^*) - C(d^* - 1) - (p - \Delta p)(1 - \gamma)] \\
= h\left(d^*, \bar{\theta}_{d^* - 1}\right) - h\left(d^* - 1, \bar{\theta}_{d^* - 1}\right)
\]
which implies that the cutoff values increase, i.e. \( \theta_d^{\text{reform}} > \theta_d^{\text{baseline}} \) \( \forall d \). That is, the patients stay on average for a shorter time. Due to the discreteness of the assignment variable —length of stay— I need to make an additional regularity assumption relative to the standard bunching setting. Specifically, I assume that patients who share the same length of stay \( d \) under the old schedule \( P^{\text{baseline}} (d_i) \) move towards at most two different length of stay values under the new schedule \( P^{\text{reform}} (d_i) \). That is, those patients who stay, e.g., 5 days under the old schedule, stay for 3–4 days or for 4–5 days under the new schedule, but never for 2–4 or 3–5 days.

Now consider the introduction of a convex kink at \( d^* \). That is, the payment schedule becomes

\[
P^{kink} (d_i) = \begin{cases} 
\bar{p} + p \cdot d_i & d_i \leq d^* \\
\bar{p} + (p - \Delta p) \cdot d_i & d_i > d^*
\end{cases}
\]

Under the new kinked payment schedule, the new cutoff values defining who is discharged at \( d^* - \theta^{kink}_{d^*} \) and \( \theta^{kink}_{d^*-1} \) are defined by

\[
\lambda_d \lambda_h U'(\pi) [C (d^* + 1) - C (d^*) - (p - \Delta p) (1 - \gamma)] \\
= h (d^* + 1, \bar{\theta}^{kink}_{d^*}) - h (d^*, \bar{\theta}^{kink}_{d^*})
\]

\[
\lambda_d \lambda_h U'(\pi) [C (d^*) - C (d^* - 1) - p (1 - \gamma)] \\
= h (d^*, \bar{\theta}^{kink}_{d^*-1}) - h (d^* - 1, \bar{\theta}^{kink}_{d^*-1})
\]

Hence, \( \bar{\theta}^{kink}_{d^*-1} = \bar{\theta}^{\text{baseline}}_{d^*-1} \) and \( \bar{\theta}^{kink}_{d^*} = \bar{\theta}^{\text{reform}}_{d^*} > \bar{\theta}^{\text{baseline}}_{d^*} \) and \( \bar{\theta}^{kink}_{d^*} - \bar{\theta}^{\text{baseline}}_{d^*} \) is the excess mass or bunching at \( d^* \) under the kinked schedule. Hence, \( \bar{\theta}^{kink}_{d^*} \) is the
marginal buncher who responds to the introduction of $P^{kink}(d_i)$ the same way as to the introduction of $P^{reform}(d_i)$.

**What Does a Bunching Design Identify?**

We established that the marginal buncher responds to the introduction of the kink the same way as to the policy experiment of interest (that is, changing marginal pay by $\Delta p$ throughout the schedule). Let $\bar{d}$ denote the length of stay that the marginal buncher $\theta_{d^*}$ would have enjoyed under the baseline linear schedule. For this marginal buncher, the causal effect of interest —the effect of changing marginal reimbursement per day by $\Delta p$ on length of stay— is $\frac{\partial}{\partial (d_i)} \bar{d} = \bar{d} - d^*$. Using the assumption discussed above, the causal effect $\frac{\partial}{\partial (d_i)} \bar{d}$ is equal to $\bar{d} - d^*$ for all patients who would have stayed $\bar{d}$ under the baseline linear schedule and for whom $\theta_i < \theta_{d^*}^{kink}$, but the causal effect $\frac{\partial}{\partial (d_i)} \bar{d}$ is $\bar{d} - (d^* + 1)$ for all patients who would have stayed $\bar{d}$ under the baseline linear schedule and for whom $\theta_i > \theta_{d^*}^{kink}$. Therefore, the total causal effect on patients staying $\bar{d}$ under the old baseline schedule is

$$E \left[ \frac{\partial}{\partial (d_i)} \Delta p \left| \theta_{d^*}^{baseline} > \theta_i > \theta_{d^* - 1}^{baseline} \right. \right] = \left( \bar{d} - d^* \right) \frac{\theta_{d^*}^{kink} - \theta_{d^* - 1}^{baseline}}{\theta_{d^*}^{baseline} - \theta_{d^* - 1}^{baseline}}$$

$$+ \left( \bar{d} - (d^* + 1) \right) \frac{\theta_{d^*}^{baseline} - \theta_{d^*}^{kink}}{\theta_{d^*}^{baseline} - \theta_{d^* - 1}^{baseline}}$$

$$= \bar{d} - d^* - 1 + \frac{\theta_{d^*}^{baseline} - \theta_{d^*}^{baseline}}{\theta_{d^*}^{baseline} - \theta_{d^* - 1}^{baseline}}$$

A simple example makes the formula intuitive: If the observed bunching mass is only a small fraction of the observed mass at $d^* + 1$—say, 10%—$\bar{d} = d^* + 1$, because the marginal buncher is coming from $d^* + 1$, and $\frac{\theta_{d^*}^{baseline} - \theta_{d^* - 1}^{baseline}}{\theta_{d^*}^{baseline} - \theta_{d^* - 1}^{baseline}} \approx 0.1$, since only patients who are at $d^* + 1$ under the counterfactual linear schedule and whose $\theta_i < \theta_{d^*}^{kink}$ bunch at $d^*$ together with the marginal buncher. In the example, the formula tells us that the average causal effect on the patients staying $d^* + 1$ days under the counterfactual
linear schedule is 0.1 days, since that is the fraction of patients who move from $d^* + 1$ to $d^*$ due to the kink.

Since $d^*$ is known, we need to estimate $\tilde{d}$ and $\frac{\bar{\theta}_{d^* \text{ kink}} - \bar{\theta}_{d^* \text{ baseline}}}{\bar{\theta}_{d-1} - \bar{\theta}_{d-1 \text{ baseline}}}$ in order to get the causal effect of interest. Let $B$ denote the bunching mass estimated from the data and $f(d)$ the estimated expected mass of patients at $d$ under the contrafactual linear schedule. Then $\tilde{d}$ can be inferred from the data by finding the value for $\tilde{d}$ for which

$$f(d^* + 1) + ... + f(\tilde{d}) \geq B$$

$$f(d^* + 1) + ... + f(\tilde{d} - 1) \leq B,$$

since the bunching mass is equal to the mass at the days from $d^* + 1$ up to $\tilde{d} - 1$ plus the fraction of the mass at $\tilde{d}$ that bunches. This fraction is the bunching mass that is not explained by the mass coming from $d^* + 1$ up to $\tilde{d} - 1$, i.e.

$$\frac{\bar{\theta}_{d^* \text{ kink}} - \bar{\theta}_{d^* \text{ baseline}}}{\bar{\theta}_{d-1} - \bar{\theta}_{d-1 \text{ baseline}}} = B - f(d^* + 1) + ... + f(\tilde{d} - 1)$$

Hence, the remaining challenge is to estimate $B$ and $f(d^* + 1)$, etc. in order to estimate the parameter of interest $E\left[ \frac{d(d_i)}{d(p)} \Delta p \left| \bar{\theta}_d^{\text{baseline}} > \theta_i > \bar{\theta}_{d-1}^{\text{baseline}} \right. \right]$.  

69
1.C Further Results

Figure 1.17: Time Series Length of Stay - Selection of OECD Countries

Average length of stay over time for selected OECD countries.
Note that the numbers for Germany do not necessarily agree with the numbers in Figure 1.5 due to the different source.
Source: OECD
The graph shows the average length of stay for each month 2005-2013 separately for DRGs involving a medical procedures and so-called medical DRGs that do not.
Figure 1.19: Time Series Length of Stay - by Quartile of Hospital w.r.t. 2005 Average Length of Stay

The graph shows the average length of stay (not residualized) for each month 2005-2013 separately by quartile of the hospital in terms of the hospital’s residualized length of stay in 2005. I.e. the first quartile are those hospitals that - conditional on the patients’ age, gender, birthweight and DRG - have the lowest length of stay in 2005.
Figure 1.20: Heterogeneity in Hazard Rates Around Kink - by Quartile of Hospital Size

Hazard rates for all patients pooled with a DRG with kink locations of at least 6 days. Done separately by quartile of hospital size.
Hazard rates for all patients pooled with DRGs with kink locations of at least 6 days. Separately for different discharge reasons. Transfer refers to transfers to other hospitals (transfers are not subject to the kink in the payoff schedule). Rehabilitation refers to transfer to a rehabilitation or longterm care unit or a hospice.
The graph shows —pooled across all hospitals and years— the number of births for each 10 gram bin of birth weight. The dots are plotted against the middle point of the respective bin - i.e. the bin from 1,500g to 1,509g is plotted at 1,505g.
Table 1.10: Distribution of Kink Locations in the Cross section

<table>
<thead>
<tr>
<th>Kink Location</th>
<th>DRGs</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>n/a</td>
<td>899</td>
<td>7091698</td>
</tr>
<tr>
<td>2</td>
<td>3259</td>
<td>85555788</td>
</tr>
<tr>
<td>3</td>
<td>1442</td>
<td>18374780</td>
</tr>
<tr>
<td>4</td>
<td>1239</td>
<td>11334270</td>
</tr>
<tr>
<td>5</td>
<td>803</td>
<td>4806553</td>
</tr>
<tr>
<td>6</td>
<td>519</td>
<td>1880018</td>
</tr>
<tr>
<td>7</td>
<td>343</td>
<td>780616</td>
</tr>
<tr>
<td>8</td>
<td>276</td>
<td>438387</td>
</tr>
<tr>
<td>9</td>
<td>183</td>
<td>243005</td>
</tr>
<tr>
<td>10+</td>
<td>551</td>
<td>411956</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>9514</td>
<td>130917071</td>
</tr>
</tbody>
</table>

For each kink location the table gives the number of DRGs that feature this kink location (each year counted separately, so the same DRG is counted several times if it exists in multiple years) as well as the total number of patients who are grouped in one of the respective DRGs. Each patient is only counted once.

1.D Bunching Estimate for Einav et al. Setting

This appendix describes how I construct the bunching estimate of a 0.34 day reduction in length of stay when cutting marginal pay for another day in the hospital by 1,000€ for the setting studied in Einav et al. (2017), Eliason et al. (2016) and Kim et al. (2015). It builds notationwise on the derivation in the main text for my main analysis.

I use the distribution of downstream patient discharges across days reported in Figure 8 of Einav et al. (2017) for the time before the jump in payment at the threshold was introduced and for after it was introduced.\footnote{The January 2017 version of their paper.} I assume that patients only shifted within the ±15 day window around the threshold, that is, I restrict the data to this window.

At the threshold there is a notch (an upward jump in the payment) as well as a kink (the slope of the schedule turns from being increasing towards being flat). To
estimate the amount of bunching I run the following type of specification

\[
share_{d,t} = \alpha_d + \beta_0 I \{post_t\} + \beta_1 d \cdot I \{post_t\} + \beta_2 I \{below\, threshold_d\} \cdot I \{post_t\} \\
+ \beta_3 I \{above\, threshold_d\} \cdot I \{post_t\} + u_{d,t}
\]

with \(share_{d,t}\) denoting the share of patients in time period \(t\) that are discharged after \(d\) days, \(\alpha_d\) indicators for each day, \(I \{post_t\}\) an indicator for whether \(t\) is for the time after the threshold was introduced, \(I \{below\, threshold_d\}\) an indicator for whether \(d\) is for the days just below the threshold (how many days is determined within the process of looking for a solution, because it depends on the bunching mass) and \(I \{above\, threshold_d\}\) an indicator for whether \(d\) is the day just above the threshold.

That is, the \(\alpha_d\) imply that the coefficients of interest are only identified from the changes from the pre- to the post-period and not from smoothness assumptions. \(I \{post_t\} + d \cdot I \{post_t\}\) allows the mass to shift in the post-period relative to the pre-period, i.e. allows for time trends. The \(I \{above\, threshold_d\} \cdot I \{post_t\}\) captures the bunching mass and the \(I \{below\, threshold_d\} \cdot I \{post_t\}\) allows for the hole in the mass of discharges that is expected for a notch with an upward jump.

Due to the presence of the notch as well as the kink, there is a marginal buncher caused by the notch and coming from below the threshold day \(d^*\) —denote the contrafactual day of that marginal buncher by \(d'^{\text{notch}}\) and her type \(\theta^{\text{notch}}\)— as well as a marginal buncher caused by the kink and coming from above the threshold day \(d^*\) —denote the contrafactual day of that marginal buncher by \(d'^{\text{kink}}\) and her type \(\theta^{\text{kink}}\).

I approximate the notch induced change in marginal pay by \(\frac{\Delta p}{d^* - d'^{\text{notch}}}\), with \(\Delta p\) (about $13,000 on average) denoting the payment jump at the threshold. The kink induced change in marginal pay is simply the slope \(\tilde{p}\) that the schedule has to the left of the threshold (about $1,400 on average).
By definition, the marginal bunchers react to the notch and kink the same way as they would to a change in marginal pay throughout the schedule, i.e. $\theta_{\text{notch}}$ would also move from $d_{\text{notch}}$ to $d^*$ if marginal pay were increased by $\frac{\Delta p}{d^* - d_{\text{notch}}}$ throughout the schedule and $\theta_{\text{kink}}$ would also move from $d_{\text{kink}}$ to $d^*$ if marginal were decreased by $\bar{p}$ throughout the schedule. Hence, the causal effect of interest for $\theta_{\text{notch}}$ is $d(d_i) \frac{\Delta p}{d} = d^* - d_{\text{notch}}$ and for $\theta_{\text{kink}}$ is $\frac{d(d_i)}{d} \bar{p} = d_{\text{kink}} - d^*$.

I make the same regularity assumption as in my main text: Patients who share the same length of stay under one payment schedule are at at most two different length of stay values under a different payment schedule - e.g. all patients who stay 6 days under one payment schedule stay 3 or 4 days or they stay 4 or 5 days, but never 3 to 5 days.

Under the regularity assumption, the causal effect for all patients with $\theta_i > \theta_{\text{notch}}$ and who in the absence of the notch would also stay $d_{\text{notch}}$ is $d(d_i) \frac{\Delta p}{d} = d^* - d_{\text{notch}}$, while those with $\theta_i < \theta_{\text{notch}}$ and who in the absence of the notch would also stay $d_{\text{notch}}$, have a causal effect $d(d_i) \frac{\Delta p}{d} = (d^* - 1) - d_{\text{notch}}$. Similarly, all patients with $\theta_i < \theta_{\text{kink}}$ and who in the absence of the kink would also stay $d_{\text{kink}}$ have the causal effect $\frac{d(d_i)}{d} \bar{p} = d_{\text{kink}} - d^*$, while those with $\theta_i > \theta_{\text{notch}}$ and who in the absence of the notch would also stay $d_{\text{kink}}$, have the causal effect $\frac{d(d_i)}{d} \bar{p} = d_{\text{kink}} - (d^* + 1)$.

Let $\theta_{d_{\text{contrafactual}}}$ denote the patient marginal between $d + 1$ and $d$ under the contrafactual distribution in the post period if no jump and kink had been introduced.

The average causal effect for those at $d_{\text{notch}}$ is $\frac{d(d_i)}{d} \frac{\Delta p}{d} = (d^* - d_{\text{notch}}) \frac{\theta_{\text{notch}} - \theta_{d_{\text{contrafactual}}}}{\theta_{\text{notch}} - \theta_{d_{\text{notch}} - 1}} + ((d^* - 1) - d_{\text{notch}}) \frac{\theta_{d_{\text{contrafactual}}}}{\theta_{d_{\text{notch}} - 1}}$. The average causal effect for those at $d_{\text{kink}}$

is $\frac{d(d_i)}{d} \bar{p} = (d_{\text{kink}} - d^*) \frac{\theta_{\text{kink}} - \theta_{d_{\text{contrafactual}}}}{\theta_{\text{kink}} - \theta_{d_{\text{contrafactual}} - 1}} + ((d_{\text{kink}} - 1) - d^*) \frac{\theta_{d_{\text{contrafactual}} - \text{gkink}}}{\theta_{d_{\text{contrafactual}} - \text{gkink}} - 1}$. 


I loop through all possible values for \( \frac{d(d_i)}{dp} \) which imply the same \( \frac{d(d_i)}{dp} \). For each of those values, I run the specification mentioned above

\[
\text{share}_{d,t} = \alpha_d + \beta_0 I \{ \text{post}_t \} + \beta_1 d \cdot I \{ \text{post}_t \} + \beta_2 I \{ \text{below threshold}_d \} \cdot I \{ \text{post}_t \} + \beta_3 I \{ \text{above threshold}_d \} \cdot I \{ \text{post}_t \} + u_{d,t}
\]

The \( I \{ \text{below threshold}_d \} \) is defined as \( d \) being in the \( d^* - d_\text{notch} \) first days to the left of the threshold (note that \( d_\text{notch} \) is implied from the guessed value for \( \frac{d(d_i)}{dp} \)).

I then use the regression result to get the amount of bunching from \( \beta_3 \) and calculate the contrafactual mass distribution if the threshold had not been introduced from the predicted values not using the \( \beta_2 \) and \( \beta_3 \).

I then calculate the bunching mass implied from the guessed \( \frac{d(d_i)}{dp} \) and \( \frac{d(d_i)}{dp} \hat{p} \) together with the contrafactual distribution (if, e.g., \( \frac{d(d_i)}{dp} \Delta p = 2.3 \), I add the contrafactual masses at \( d^* - 1, \ d^* - 2 \) and 0.3 times the mass at \( d^* - 3 \). That way I calculate the total amount of bunching mass that is implied and check whether it is equal to the estimated amount of bunching.

My estimates for \( \frac{d(d_i)}{dp} \) and \( \frac{d(d_i)}{dp} \hat{p} \) are those for which the amount of bunching estimated corresponds to the one implied by \( \frac{d(d_i)}{dp} \) and \( \frac{d(d_i)}{dp} \hat{p} \).

\( \frac{d(d_i)}{dp} \) and \( \frac{d(d_i)}{dp} \hat{p} \) imply a value for \( \frac{d(d_i)}{dp} \) which I then convert to an estimate with respect to a change in 1,000 2013-Euro using the American CPI and the 2013 Euro-Dollar exchange rate.

The result is a 0.34 day reduction in length of stay when cutting the marginal pay per day by 1,000 2013-Euro. The confidence interval (bootstrapped with 250 repetitions) is from 0.31 to 0.37 days.
1.E Back-of-the-Envelope Calculation for the Comparison to the Medicare Reform

Here I describe how I calculate the average reimbursement per hospitalization day in Medicare in 1984.

Levit et al. (1985) reports 44.24 billion dollars in 1984 in Medicare hospital spending. Kominski and Witsberger (1993) and Guterman and Dobson (1986) provide average length of stay and the number of hospitalizations for 1984 Medicare patients, resulting in a total of 103.46 million Medicare hospital days in 1984. Dividing total costs by the number of days and using the U.S. CPI and the 2013 average exchange rate from Dollar to Euro, I end up with approximately 750 2013-Euro per day.
Chapter 2

Why Does Midnight Matter? Moral Hazard vs Limited Attention

2.1 Introduction

In an influential paper, Almond and Doyle (2011) implement a regression discontinuity design (RDD) for Californian newborns born just around midnight. American health insurance typically covers a fixed number of nights in the hospital for a newborn with the number of nights in the hospital being counted as the number of midnights that the newborn was present in the hospital. Newborns born just after midnight therefore effectively enjoy one additional day of insurance coverage compared to those born just before. Almond and Doyle find that newborns born just after midnight stay on average 0.25 nights longer in the hospital than those born just before. They interpret the effect as evidence of moral hazard with respect to health insurance coverage.

There is, however, an alternative interpretation of their findings. If doctors only pay limited attention to the exact time that a newborn has been alive —that is, if they only pay attention to the newborn’s age in days, but not in hours or minutes— newborns born just after midnight will appear 24 hours younger to such doctors
than newborns born just before midnight. To the degree that the perceived age of
the newborn influences the doctor’s discharge decision, this will generate the pattern
observed by Almond and Doyle.

In order to distinguish between these two interpretations, this paper implements a
similar RDD as Almond and Doyle for German newborns born just around midnight.
German health insurance does not feature a maximum number of covered nights
after delivery implying that the moral hazard channel cannot be present. Using
administrative data covering more than 800,000 births in the two-hour window around
midnight, I find that newborns born just after midnight stay about 0.3 nights longer
on average than newborns born just before midnight. The result is very robust and
passes all usual RDD diagnostics such as the required smoothness of covariates at
midnight.

My findings therefore cast doubt on the interpretation of the Almond and Doyle
paper. Moreover, my results provide evidence of limited attention in a high-stakes
professional setting and yield a lower bound of 30% for the share of newborns that
are treated by inattentive doctors.

I reject further possible explanations for the RDD result. The hospitals’ financial
incentives cannot rationalize the jump at midnight. In the main text, I show in detail
that the financial incentives that the hospitals face could explain a jump at midnight
in the probability of being discharged after one additional midnight (excluding the
midnight that defines the RDD threshold). In the data, the jump is much more
pronounced after two additional midnights which the financial incentives cannot
rationalize.

Another possible reason behind the results are administrative rules requiring
patients to stay a minimum number of nights. Just as in the US, the number of nights
in the hospital in Germany is defined as the number of midnights passed. So if patients
are required to stay a minimum number of nights due to some administrative rule,
this could generate the observed jump. Such rules—nationwide or hospital-specific—

driving the results is not consistent with the data, however, as I demonstrate in the
detail in the main text. I also show that neither changes in shift nor changes in
C-Section probability at midnight are behind this paper’s findings.

In the last part of the paper, I formulate a simple model to illustrate formally
why limited attention can lead to the observed patterns in the data. In the model,
doctors decide about discharge based on two inputs: Firstly, the signal they get about
the newborn’s health status when examining her and, secondly, her age. The age is
relevant because the same health signal provides different information about the true
underlying health status depending on the newborn’s age. If some doctors pay limited
attention to the exact age—that is, if they only take into account the discrete number
of days a newborn has been alive, but not the exact number of hours—, the model
predicts jumps at midnight as observed in the data. The estimated results from the
data imply that at least 30% of newborns are treated by a doctor who only pays
attention to the discrete number of days the newborn has been in the hospital, but
not the precise number of hours.

Limited attention on the patient side with respect to the newborn’s precise age
could in principle generate the same results. This mechanism seems much less
plausible though, since the parents do presumably remember the time of birth very
well and because in practice it is predominantly the doctors deciding about the
discharge date.

I conducted qualitative interviews with three doctors who have treated newborns
in German hospitals. They do not have any alternative explanation for my results and
two of them even suggested the limited attention mechanism without me suggesting
it to them first.

This paper’s results imply that researchers should be careful when using cutoffs
for identification if these cutoffs do not only affect the treatment of interest, but
could also trigger a different rule of thumb. For instance, the discontinuous change in Medicare eligibility at 65 years is a popular RDD threshold in the health economics literature. If doctors used rule-of-thumb age cutoffs to guide treatment decisions, the exclusion restriction would possibly be violated. A documented example of cutoff-dependent rule-of-thumb behavior leading to controversial RDD evidence is the debate around the moral hazard impact of loan securitization on lender screening. Keys et al (2010) had used the discontinuous increase in securitization ease at certain credit score thresholds to establish the causal effect of securitization on lender screening intensity. Bubb and Kaufman (2014), however, contested their findings, demonstrating that lender screening intensity follows rule of thumbs depending on the credit score, hence leading to jumps in screening intensity independent of the ease of securitization.

This paper also relates to the literature on doctor decision making and behavioral biases by professionals. While there is abundant evidence that traditional economic incentives matter in health care provider behavior (see e.g. the literature on the effect of fees for physicians on treatment choices, for instance Gruber et al 1999, Clemens and Gottlieb 2014), other research shows that rules of thumb (Almond et al 2010, Bharadwaj et al 2013) and behavioral biases (e.g. LeBlanc et al 2002, Brewer et al 2007, Tannenbaum et al 2015) also have a role in doctors’ decision making. There is some evidence suggesting that optimization failures among physicians decline with growing experience (Marmede et al 2010). To what degree behavioral biases and optimization failures vanish in high stakes situations or with the professional experience of the agent has been debated (e.g. List 2003, List 2011, Seru et al 2010, Feng and Seasholes 2005, Lacetera et al 2012, Foellmi et al 2016). This paper contributes to this literature by providing evidence of limited attention in a high-stakes sector by professionals.

Outline

84
I start by describing additional institutional background, the data and sample selection in Section 2.2. The empirical strategy and main results are discussed in Section 2.3. In Section 2.4, I confront the different possible explanations for the observed results with the data. Section 2.5 develops a model to illustrate how limited attention can lead to the observed empirical findings. Section 2.6 discusses and concludes.

2.2 Institutional Setting, Data and Sample

Institutional Background

The vast majority of German newborns (more than 98% in 2013) are delivered in a hospital. The mother and child are typically not treated by their ambulatory gynecologist who took care of them during the pregnancy, but by the doctors and midwives present in the hospital at the time of delivery. Within the first 30 minutes after birth the so-called U1 check is performed. It involves examining vital body function of the newborn.

The next check is the so-called U2. During the U2, the newborn is checked for further possible innate diseases such as heart defects or hearing loss. The U2 is usually still performed in the hospital (but it can also be done ambulatory) and after the U2 has been performed the newborn and the mother are often discharged if no problems were detected. The discharge decision is normally made jointly by the doctor responsible for the mother and the doctor responsible for the newborn, since if both are healthy they are typically released at the same time. The U2 guidelines ask for the U2 to take place between the third and the tenth day of the newborn’s life (that is, after two to nine midnights in the hospital). I show in detail later that the timing recommendation for the U2 cannot explain the observed results.
Germany has near universal health insurance coverage (in 2015, about 0.1% had no insurance). A newborn’s initial hospital stay is covered by the mother’s health insurer if the newborn is healthy. Otherwise the newborn’s health insurer is responsible (which normally is the health insurer of the parent with the higher income). About 90% of the German population are members of one of the public health insurers. They have no financial stake in how long they stay in the hospital after childbirth, except for a subset of patients who are subject to a small copay of 10 EUR a day after the sixth night is reached (in 2013, less than 3% of newborns in my window around midnight stayed 6 nights or more). The 10% of Germans who are insured with a private health insurer typically have a deductible before health insurance coverage begins. Those patients may have a financial stake in how long they remain hospitalized — however, their financial incentives are exactly inverse to those of the hospital and can therefore not explain the observed jump at midnight either. In my data, I am not able to distinguish between privately and publicly insured patients.

In 2013, there were 1,995 hospitals in Germany, with 596 being public, 706 non-profit and 693 for-profit. The reimbursement for hospitals is federally regulated and independent of the patient’s health insurer. It varies across hospitals only in the form of proportional shift factors. Until 2010 these shift factors were hospital-specific and they have been uniform at the state-level since then.

In the case of privately insured patients (or publicly insured patients who are willing to pay extra money in order to be treated by the head physician), the head physician can charge additionally per service. Typically, these additional charges are then shared with the other doctors in her department. German doctors working in hospitals are salaried and unionized, except for the head physicians whose pay is individually contracted and does often depend on economic outcomes in her department.

Data
I use administrative data from the Federal Statistical Agency in Germany.\textsuperscript{1} It covers the universe of in-patient hospitalizations from 2005 - 2013 and includes variables such as baseline patient characteristics like sex, age and region as well as case characteristics like diagnoses, procedures, length of stay, hospital identifier and admission and discharge date and time. All newborns that are born in the hospital and remain in the hospital beyond the delivery room qualify as an in-patient hospitalization and show up in my data. In 2012, more than 93\% of all births in Germany are covered by my data, the other 7\% being home births and newborns who do not stay in the hospital beyond the delivery room. Overall, my data includes more than 800,000 births in the two hour window around midnight.

All hospitals are required by law to report all of the previous year’s hospitalizations until March 31st to the Federal Statistical Agency. The data the hospitals send to the agency is based on the data generated for billing purposes and hence of very high quality. Among the hospitalizations of interest —those coded as births in the hospital— there are no cases in which the key variables length of stay, discharge date and time, admission date and time, birth weight or discharge reason are missing.

There are two potential sources of measurement error in length of stay. First, audits by health insurers and, second, readmissions. Audits could potentially induce some measurement error into the length of stay variable, because the billed number of nights could be changed relative to the actual nights the newborn stayed. As discussed in more detail in Schlockermann (2018), a significant fraction of successful audits would show up in my data in the form of a divergence between the billed number of days and the difference between discharge and admission date. Readmissions could also induce measurement error, because if a readmission takes place within a certain number of days, the hospitalizations have to be merged to one case. In the data,

\textsuperscript{1}The official source for all reported results in this paper is: 'FDZ der Statistischen Ämter des Bundes und der Länder, DRG-Statistik 2005-2013, eigene Berechnungen' (Translation: 'Research Data Center of the Federal and State Statistical Agencies, DRG-Statistics 2005-2013, own calculations')
I would then only see the one merged observation with the billed number of days being the sum of the billed days of the merged hospitalizations, the admission date being the admission date of the first hospitalization and the discharge date being the discharge date of the last hospitalization. As in the case of a successful audit, the billed number of days and the difference between discharge and admission date would diverge. Since for my analysis sample of healthy newborns these two measures differ in less than 0.01% of all cases, I conclude that measurement error in length of stay does not pose a problem for this paper. I conduct all of my analysis using the difference between discharge and admission time as my measure of length of stay.

**Analysis Sample**

For my main analysis sample, I focus on healthy and normal-weight newborns, since those have a qualitatively uniform payoff schedule which makes the hospitals’ financial incentives easy to understand. In particular, the analysis sample is restricted to singletons above 2500g birthweight without any major medical conditions. I also exclude newborns who die during their initial hospital stay or who are transferred to another hospital, since those newborns are subject to special reimbursement schedules. In the two hour window around midnight, the analysis sample makes up about 80% of all hospital births. In the results section, I demonstrate that my sample choice does not induce any selection at midnight.

Table 3.1 shows descriptive statistics for my main analysis sample as well as the excluded births in the two-hour window around midnight. Since the excluded births are those with severe illnesses and lower birth weight, there is a strong difference in average birth weight, average length of stay as well as diagnoses and procedures. The probability of death and transfer to a different hospital are zero in the analysis sample by construction.

---

2The implied number of readmissions is so small, because readmitted newborns have some serious medical condition and are therefore not part of my analysis sample as discussed below.
Table 2.1: Summary statistics - analysis sample vs excluded births

<table>
<thead>
<tr>
<th></th>
<th>Analysis Sample</th>
<th>Excluded Births</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Share Female</td>
<td>0.49 (0.5)</td>
<td>0.484 (0.5)</td>
<td>0.007 (0.001)</td>
</tr>
<tr>
<td>Birth Weight</td>
<td>3445.49 (433.22)</td>
<td>3004.49 (815.48)</td>
<td>441 (1.48)</td>
</tr>
<tr>
<td>Length of Stay in hours</td>
<td>85.33 (41.68)</td>
<td>203.88 (352.06)</td>
<td>-118.55 (0.45)</td>
</tr>
<tr>
<td>Number of Diagnoses</td>
<td>0.399 (0.9506)</td>
<td>2.3522 (3.8737)</td>
<td>-1.9532 (0.0054)</td>
</tr>
<tr>
<td>Number of Procedures</td>
<td>1.5044 (0.7831)</td>
<td>2.7887 (3.5437)</td>
<td>-1.2843 (0.0048)</td>
</tr>
<tr>
<td>Probability Death</td>
<td>0 (0)</td>
<td>0.0102 (0.1003)</td>
<td>-0.0102 (0.0001)</td>
</tr>
<tr>
<td>Probability Transfer</td>
<td>0 (0)</td>
<td>0.1022 (0.3029)</td>
<td>-0.1022 (0.0004)</td>
</tr>
<tr>
<td>N</td>
<td>644469</td>
<td>161119</td>
<td>805588</td>
</tr>
</tbody>
</table>

Means and standard errors for different variables for all births in the +/- 2 hour window around midnight - separately for the analysis sample and the excluded births.

Table 2.2 shows descriptive statistics for the newborns in the 10pm to 2am window compared to newborns in the remaining 20 hours. The differences are small overall, except for a longer average length of stay for newborns born between 2am and 10pm. The longer length of stay is, presumably, reflective of the higher number of (scheduled) C-Sections during the day (the distribution of C-Sections across time of the day is discussed in more detail later).

Table 2.2: Summary statistics — 10p-2a vs 2a-10p

<table>
<thead>
<tr>
<th></th>
<th>10p-2a</th>
<th>2a-10p</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obeys Sample Selection Criteria</td>
<td>0.8 (0.4)</td>
<td>0.7719 (0.4196)</td>
<td>0.0281 (0.0005)</td>
</tr>
<tr>
<td>Birth Weight</td>
<td>3357.3 (560.6)</td>
<td>3342.9 (575.7)</td>
<td>14.4 (0.7)</td>
</tr>
<tr>
<td>Probability Female</td>
<td>0.4891 (0.4999)</td>
<td>0.4889 (0.4999)</td>
<td>0.0002 (0.0006)</td>
</tr>
<tr>
<td>Length of Stay in hours</td>
<td>109.0382 (168.6039)</td>
<td>119.1144 (182.3892)</td>
<td>-10.0762 (0.2173)</td>
</tr>
<tr>
<td>Number of Diagnoses</td>
<td>0.7897 (2.082)</td>
<td>0.8394 (2.2185)</td>
<td>-0.0497 (0.0026)</td>
</tr>
<tr>
<td>Number of Procedures</td>
<td>1.7612 (1.8072)</td>
<td>1.7966 (1.9242)</td>
<td>-0.0354 (0.0023)</td>
</tr>
<tr>
<td>Probability Death</td>
<td>0.002 (0.045)</td>
<td>0.0019 (0.044)</td>
<td>0.0001 (0.0001)</td>
</tr>
<tr>
<td>Probability Transfer</td>
<td>0.0204 (0.1415)</td>
<td>0.0215 (0.145)</td>
<td>-0.001 (0.0002)</td>
</tr>
<tr>
<td>Number Departments</td>
<td>1.1051 (0.3887)</td>
<td>1.1106 (0.4014)</td>
<td>-0.0054 (0.0005)</td>
</tr>
<tr>
<td>N</td>
<td>805588</td>
<td>4802951</td>
<td>5608539</td>
</tr>
</tbody>
</table>

Means and standard errors for different variables for all births in the 10p-2a hour window around midnight compared with the births between 2a and 10p. ‘Obeys Sample Selection Criteria’ refers to whether the newborn would have been selected for the main analysis sample (i.e. whether the newborn is healthy with birth weight above 2500g). ‘Number Departments’ refers to the number of different departments that the newborn was hospitalized in during her stay.
2.3 Empirical Strategy and Results

Frequency of Births Around Midnight

I start by showing the distribution of births across time around midnight. The identifying assumption for the RDD is that parents and doctors have no tight control about the minute of birth, implying a smooth distribution of births at midnight. Figure 2.1 shows the total number of births observed for each 5 minute bin in the two hour window around midnight. It is shown separately for the analysis sample as well as for the excluded births to demonstrate that there is no discontinuous change in the share of cases that belong to the analysis sample.

There is no jump in the density at midnight. There appears to be some manipulation in the 5 minute window, however. Specifically, there is an excess mass of births in the 5 minutes before midnight and a missing mass in the 5 minutes after. The excess mass and the missing mass are about equal in size. Apparently, hospitals manipulate the birth time of newborns born in the 5 minutes after midnight by documenting them as having been born in the 5 minutes prior to midnight. Possibly, this is due to the fact that documenting the time of birth as having occurred before midnight increases the number of billed days and, therefore, might increase pay to the hospital (this will be clear in the later section when the hospitals’ financial incentives are discussed in detail).

Since the manipulation of birth weight is so very local, it does not pose a major problem for my results. All the binned scatter plots will be presented including the newborns in the 5 minute window and it will be clear that the jumps are present with or without taking the group into account. The next subsection will also demonstrate that there is no jump in birthweight or sex composition in the 5 minute window. I present all regressions using the full analysis sample as well as using a donut estimator excluding the observations in the 5 minute range. Moreover, I show the results for an estimator excluding the 5 minute window and all multiples of 5 minutes, because
—presumably, due to rounding by the midwives recording the birth data— there is some excess mass of births at multiples of 5 minutes. The differences turn out to be small.

![Figure 2.1: Frequency of births](image)

Each dot shows the total number of births in the corresponding 5 minute bin - either for the analysis sample or for the excluded births. There is some manipulation by in the +/- 5 minute window - it is unproblematic as discussed and shown.

I test for a discontinuity in the number of births by estimating the following kind of specifications

\[
\text{count}_t = \beta_0 + \beta_1 \cdot I\{t \geq 0\} \cdot t + \beta_2 \cdot I\{t < 0\} \cdot t + \gamma \cdot I\{t \geq 0\} + u_t \quad (2.1)
\]

with \( t \) denoting minutes relative to midnight (i.e. 11.55pm would correspond to \( t = -5 \)), \( \text{count}_t \) the number of births at minute \( t \) and \( I \{ \cdot \} \) the indicator function. The specification allows for different slopes on both sides of midnight and \( \gamma \) is the coefficient of interest. Table 2.3 presents the results for the analysis sample, Table
2.5 in the appendix for the excluded births. There are no significant jumps in the frequency regardless of the bandwidth or whether we use the donut estimator.

Table 2.3: Frequency of births (for the analysis sample)

<table>
<thead>
<tr>
<th></th>
<th>30min window</th>
<th>60min window</th>
<th>90min window</th>
<th>120min window</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full analysis sample</td>
<td>-144.73 (288.57)</td>
<td>-77.46 (181.33)</td>
<td>-92.05 (136.49)</td>
<td>-74.96 (116.2)</td>
</tr>
<tr>
<td>Obs</td>
<td>60</td>
<td>120</td>
<td>180</td>
<td>240</td>
</tr>
<tr>
<td>MeanDep</td>
<td>2724.43</td>
<td>2711.37</td>
<td>2699.71</td>
<td>2685.29</td>
</tr>
<tr>
<td>Donut estimator</td>
<td>262.65 (313)</td>
<td>91.22 (184.37)</td>
<td>5.36 (135.15)</td>
<td>-2.67 (115.46)</td>
</tr>
<tr>
<td>Obs</td>
<td>50</td>
<td>110</td>
<td>170</td>
<td>230</td>
</tr>
<tr>
<td>MeanDep</td>
<td>2717.86</td>
<td>2707.19</td>
<td>2696.32</td>
<td>2682.16</td>
</tr>
<tr>
<td>Donut and no 5 min mult</td>
<td>101.22 (169.22)</td>
<td>22.08 (101.21)</td>
<td>-16.66 (73.25)</td>
<td>-22.39 (63.87)</td>
</tr>
<tr>
<td>Obs</td>
<td>40</td>
<td>88</td>
<td>136</td>
<td>184</td>
</tr>
<tr>
<td>MeanDep</td>
<td>2535.23</td>
<td>2529.31</td>
<td>2521.87</td>
<td>2508.49</td>
</tr>
</tbody>
</table>

Shows results for regressing the count per minute on an indicator for whether midnight has been crossed controlling for a linear term (not weighted, i.e. plain OLS) in minute of birth the slope of which is allowed to change at the cutoff. The sample consists of the analysis sample. The columns correspond to different time windows around midnight. 'Full analysis sample', 'Donut estimator' and 'Donut and no 5 min mult' correspond to the estimated coefficient and standard error from the full analysis sample, a donut estimator excluding the 5 minute window around midnight and a donut estimator that excludes the 5 minute window around midnight as well as all multiples of 5 minutes. 'Obs' and 'MeanDep' correspond to the number of observations and the mean of the dependent variable. *: p<.1, **: p<.05, ***: p<.01

**Smoothness of Covariates Around Midnight**

The identifying assumption of imperfect control of time of birth has a second testable implication: predetermined covariates should evolve smoothly at midnight.

Figure 2.2 shows how birth weight, gender composition, the number of diagnoses and the number of medical procedures evolve around midnight. All four covariates
show no sign of a discontinuity at midnight. Note that the number of diagnoses and the number procedures are in principle endogenous with respect to length of stay, so had there been a jump it would not have made the RDD invalid.

Figure 2.2: Covariates around midnight

Birth weight is well-documented to be a strong predictor of length of stay, mortality and even outcomes in adult life (Currie 2009). It seems very unlikely for there to be selection taking place at midnight with respect to some unobservable characteristic that has a strong influence on length of stay yet does not correlate in any way with birth weight. Therefore, I take Figure 2.2 as providing strong evidence in favor of the validity of the RDD. The fact that average birth weight is almost flat over the entire two hour window suggests that newborns born around 10pm are not that different after all from newborns born around 2am. Therefore, we can put some faith even
in the RDDs with a relatively large bandwidth (which have the advantage of higher precision at the cost of some bias).

To econometrically test for discontinuities in covariates, I run the following kind of specifications

\[ y_i = \beta_0 + \beta_1 \cdot I \{ t_i \geq 0 \} \cdot t_i + \beta_2 \cdot I \{ t_i < 0 \} \cdot t_i + \gamma \cdot I \{ t_i \geq 0 \} + u_i \]  

(2.2)

with \( y_i \) denoting the covariate of interest (the birth weight, an indicator for female, the number of diagnoses or the number of medical procedures) for newborn \( i \) and \( t_i \) being \( i \)'s minute of birth relative to midnight. \( \gamma \) is the coefficient of interest.

In the appendix, Tables 2.6, 2.7, 2.8 and 2.9 show the results for birth weight, sex composition, the number of diagnoses and the number of medical procedures. The regressions confirm the graphical impression. For the share of female births there are significant coefficients (of very small magnitude) for the full sample which disappear once we use the donut estimator.

**Outcome Variables Around Midnight**

**Length of Stay**

The last two subsections confirmed that in terms of the underlying characteristics of the newborns there is no reason to expect any discontinuity in length of stay at midnight. Figure 2.3 shows average length of stay in hours for 5 minute bins for the two hour window around midnight. Clearly, length of stay is evolving very smoothly and almost linearly to the left and right of midnight. The most pronounced feature of the graph is the jump by more than 6 hours at midnight.
Next, I present the econometric results corresponding to the presented graph. I am estimating specifications of the form

$$y_i = \beta_0 + \beta_1 I\{t_i \geq 0\} \cdot t_i + \beta_2 I\{t_i < 0\} \cdot t_i + \gamma \cdot I\{t_i \geq 0\} + \beta_3 X_i + u_i \quad (2.3)$$

with $y_i$ now denoting an outcome variable of interest for newborn $i$, $t_i$ newborn $i$’s minute of birth relative to midnight and $X_i$ possible controls such as birth weight, sex and hospital fixed effects. Table 2.4 presents the results. The effect is very precisely estimated and robust to controls and changing windows.
Table 2.4: Length of stay

<table>
<thead>
<tr>
<th></th>
<th>30min window</th>
<th>60min window</th>
<th>90min window</th>
<th>120min window</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full</strong></td>
<td>6.93*** (0.5)</td>
<td>7.02*** (0.32)</td>
<td>6.84*** (0.25)</td>
<td>6.77*** (0.22)</td>
</tr>
<tr>
<td><strong>Cluster</strong></td>
<td>60</td>
<td>120</td>
<td>180</td>
<td>240</td>
</tr>
<tr>
<td><strong>Obs</strong></td>
<td>163466</td>
<td>325364</td>
<td>485948</td>
<td>644469</td>
</tr>
<tr>
<td><strong>MeanDep</strong></td>
<td>84.73</td>
<td>84.69</td>
<td>84.65</td>
<td>84.58</td>
</tr>
<tr>
<td><strong>Donut</strong></td>
<td>6.27*** (0.45)</td>
<td>6.83*** (0.3)</td>
<td>6.69*** (0.24)</td>
<td>6.65*** (0.22)</td>
</tr>
<tr>
<td><strong>Cluster</strong></td>
<td>50</td>
<td>110</td>
<td>170</td>
<td>230</td>
</tr>
<tr>
<td><strong>Obs</strong></td>
<td>135893</td>
<td>297791</td>
<td>458375</td>
<td>616896</td>
</tr>
<tr>
<td><strong>MeanDep</strong></td>
<td>84.76</td>
<td>84.7</td>
<td>84.66</td>
<td>84.58</td>
</tr>
<tr>
<td>+ 5 min mult</td>
<td>6.34*** (0.6)</td>
<td>6.96*** (0.37)</td>
<td>6.76*** (0.29)</td>
<td>6.75*** (0.25)</td>
</tr>
<tr>
<td><strong>Cluster</strong></td>
<td>40</td>
<td>88</td>
<td>136</td>
<td>184</td>
</tr>
<tr>
<td><strong>Obs</strong></td>
<td>101409</td>
<td>222579</td>
<td>342974</td>
<td>461563</td>
</tr>
<tr>
<td><strong>MeanDep</strong></td>
<td>84.82</td>
<td>84.84</td>
<td>84.79</td>
<td>84.73</td>
</tr>
<tr>
<td>+ controls</td>
<td>6.25*** (0.59)</td>
<td>6.9*** (0.35)</td>
<td>6.7*** (0.27)</td>
<td>6.76*** (0.23)</td>
</tr>
<tr>
<td><strong>Cluster</strong></td>
<td>40</td>
<td>88</td>
<td>136</td>
<td>184</td>
</tr>
<tr>
<td><strong>Obs</strong></td>
<td>101409</td>
<td>222579</td>
<td>342974</td>
<td>461563</td>
</tr>
<tr>
<td><strong>MeanDep</strong></td>
<td>84.82</td>
<td>84.84</td>
<td>84.79</td>
<td>84.73</td>
</tr>
</tbody>
</table>

Shows results for regressions using the analysis sample of length of stay on an indicator for whether midnight has been crossed controlling for a linear term (not weighted, i.e. plain OLS) in minute of birth that is allowed to change at the cutoff. Standard errors are clustered at the minute level. The columns correspond to different time windows around midnight. 'Full', 'Donut', '+ 5 min mult' and '+controls' correspond to the estimated coefficient and standard error from the full sample, a donut estimator excluding the 5 minute window around midnight, a donut estimator that in addition also excludes all multiples of 5 and the donut estimator excluding the 5 minute multiples plus controls, that is a quadratic in birth weight, an indicator for female, and year and hospital fixed effects. 'Cluster', 'Obs' and 'MeanDep' correspond to the number of clusters, number of observations and the mean of the dependent variable. *: p<.1, **: p<.05, ***: p<.01

**Margin of Adjustment**

Next, I explore the margin of adjustment, that is the question whether the jump in length of stay is driven by all newborns being released seven hours later on the same day or rather by 30% of newborns staying another night. Figure 2.4 shows the average number of additional midnights that the newborns stay (excluding the
original midnight that defines the threshold for the RDD) as well as the average time of the day at discharge. It is clear that the margin of adjustment is along the additional midnights, because for this variable we see a jump by about 0.3, while time of the day at discharge is smooth at midnight. I conclude that the treatment effect on length of stay is zero for about 70% of newborns, while it is a whole day for the remaining 30%. Tables 2.10 and 2.11 in the appendix present the results for estimating equation 2.3 for the number of additional midnights and the time of day at discharge. The estimation results confirm the graphical impressions.\footnote{To conclude this one needs to make some assumptions on the treatment effects: First, I assume that the treatment effect on time of the day at discharge is zero for all newborns and not just washing out on average. Second, I assume that the treatment effect on additional midnights is either zero or one. While there is no plausible reason for it to be negative, it could be that in some cases the treatment effect amounts to several additional midnight — e.g. if limited attention causes newborn \( i \) to stay another night. Then in some cases there might be a health shock occurring during that night so that the 'initial treatment effect' of one day ultimately becomes several days. So in that case the share of newborns with a positive treatment effect would be a bit less than the \( \approx 0.3 \), but the average size of this treatment effect (conditional on it being positive) would be a bit larger than one.}

\footnote{Figure 2.4 also sheds some light on the downward trend in length of stay that can be observed on both sides of the cutoff in Figure 2.3. Neither the number of additional midnights nor the time of the day at discharge are on an increasing trend away from the cutoff at midnight. Naturally, a later admission time together with a stable or even earlier time of discharge is going to lead to a decreasing length of stay.}
To shed some more light on how the adjustment is taking place, Figure 2.5 shows the cumulative share of already discharged newborns plotted against the number of additional midnights — separately for newborns born just before and newborns born just after midnight. That is, the graph shows the predicted cumulative fraction discharged from a specification such as in equation 2.3 with the cumulative fraction discharged on the left hand side. First, when taking the limit from below midnight and, second, when taking the limit from above midnight. Clearly, the cumulative shares diverge strongly after two additional midnights and converge afterwards only slowly. All visually present differences are also statistically significant.
Equivalently, one can consider hazard rates. Table 2.12 in the appendix presents the regression results for specifications as in equation 2.3 with the hazard rates after zero, one, two, three, etc additional midnights on the left hand side. Note that the significantly positive coefficients on the midnight indicator at 5 and 7 additional midnights are not in contradiction with this paper’s story. There is no contradiction, because going into the 5th additional midnight, the newborns to the left and to the right of the midnight cutoff are already differentially selected, because of the jumps in the hazard rates the nights before (see also the discussion in the model section later in the paper).
2.4 Testing Further Potential Explanations for the Results

The last section’s results demonstrate that the jump is midnight is present in the German setting where there is no maximum number of health insurance coverage, hence putting doubt on the original interpretation of Almond and Doyle as moral hazard. Before I formally illustrate the inattention mechanism in a simple model, I test and reject further possible explanations for the observed jump at midnight.

Hospital Financial Incentives

The number of billable nights is defined as the number of midnights that the newborn has been present in the hospital. The payment schedule for healthy singletons in my analysis sample in 2013 in the state of Berlin (as mentioned before, in other states the pay is shifted vertically by a proportional factor) is depicted in Figure 2.6. The pay increases until two nights are reached and then remains flat until the sixth night. After the newborn has been in the hospital for six nights, the payment to the hospital increases again linearly. This extra payment for newborns staying more than six nights affects less than 3% of cases and is unlikely to be of much importance. The qualitative shape of the schedule is constant over my sample period except that in earlier years the schedule is flat until even the seventh or eighth day.
To what degree can the financial incentives explain the jump in length of stay at midnight? Consider the following example: There are two types of newborns. Johns, born Monday at 11.59pm, and Maynards, born two minutes later, Tuesday 12.01am. The payoff schedule for a John and a Maynard as a function of the discharge day is depicted in Figure 2.7. After one additional midnight (excluding the midnight that defines the RDD threshold) — i.e. on Wednesday — , the marginal pay of staying another night for Johns is zero, because they already have two billable nights. For Maynards, however, the marginal pay for another night is positive, since they only have one billable night so far. Therefore, on Wednesday it is more lucrative to discharge Johns than to discharge Maynards. Hence, financial incentives can explain a higher cumulative fraction of discharged Johns relative to Maynards on Wednesday, after one additional midnight. After two additional midnights — on Thursday — ,
however, there is no difference in marginal pay. So all the Maynards that were kept in the hospital for an extra day to get the additional pay can now be discharged. Hence, the financial incentives predict that the cumulate fraction of discharged newborns diverge after one additional midnight, but converge again after the second.

Return to Figure 2.5 which shows —separately for newborns born just before (Johns) and born just after (Maynards) midnight— the fraction of already discharged newborns as a function of additional midnights passed. While the cumulative fractions diverge slightly at one additional midnight —as predicted by the financial incentives—, they do not converge the next night at all. In fact, they further diverge. Therefore, financial incentives cannot explain the result.

**Administrative Rules**
Another possible explanation for the result is that doctors have to keep newborns in for a certain number of midnights due to an administrative rule. Can such rules rationalize this paper’s findings? First, I specifically asked the doctors whom I interviewed whether such rules exist and all of them confirmed that —besides the U2 check which I discuss below— there is no rule present that requires some minimum or maximum length of stay for some newborns.

Second, the presence of such rules has predictions which are testable in the data. I turn to those next.

**Nationwide Rules**

There cannot be a fixed nationwide minimum length of stay for newborns for the simple reason that hospitals discharge nearly 10% of patients after just zero or one midnight. Could it be that there is a fixed minimum for a subset of patients? I am aware (and this is confirmed by the interviews with the doctors) of only one actual administrative rule which is defined via the number of nights. The so-called U2 check (a standardized health check for newborns, which I discuss in the institutional background section earlier in the paper) —which often times is done while still in the hospital— is supposed to be performed between the third and the 10th day of a newborn’s life. The first day in a newborn’s life is defined to be the day of birth. Hence, the U2 can be performed once the newborn has passed two midnights in the hospital, since then it is her third day of life. So just like the financial incentives, the U2 cannot rationalize the findings, since it would only explain a deviation in the cumulative fraction after one additional midnight (since then the third day in the life of Johns is reached), but not afterwards.

Could there be an unknown rule that requires three nights for some newborns? In this case, we would expect the cumulative fraction of already discharged Johns and Maynards to diverge after two additional midnights, since the Johns would already have been present for three midnights. After three additional midnight, however, the
cumulative fractions should converge again, since the rule is no longer binding. This pattern is not present in Figure 2.5 and, hence, such a rule is not consistent with the data.

**Hospital-specific Rules**

The arguments that I just used when discussing nationwide rules also hold true within hospitals. To show this, I estimated the RDD specifications for each hospital separately.

First, the results cannot be due to hospital-specific strict minima of midnights for newborns, since even when conditioning on hospitals with a significant hospital-specific jump of length of stay at midnight, more than 5% of newborns leave after just zero or one midnight.

Second, hospital specific rules that apply for a subset of newborns. The logic used for the nationwide case holds true at the hospital level as well: Consider those hospitals for which there is a significant jump in the cumulative fraction of already discharged newborns after **two** additional midnights. If this finding were driven by these hospitals having a hospital-specific rule requiring three nights for some newborns, then —using the same logic as before for the nationwide rules— these hospitals should not have a significant hospital-specific RDD coefficient once we put the cumulative fraction of discharged newborns after **three** additional midnights on the left hand side. But for nearly 50% of these hospitals there also is a significant jump in the cumulative fraction after **three** additional midnights.

To confirm this graphically, consider Figure 2.8. For this Figure, I put the hospitals into bins based on their coefficient in the hospital-specific RDD with the cumulative fraction of discharged newborns after **two** additional midnights on the left hand side. Figure 2.8 shows —for each of these bins— the patient weighted fraction of significant hospital-specific RDD coefficients after **three** additional midnights on the vertical axis. Clearly, hospitals with a downward jump for the RDD after **two** additional
midnights tend to also show a significant downward jump for the RDD with three additional midnights.

Figure 2.8: Fraction of significant hospital-specific RDD coefficients for the cumulative fraction of discharged newborns after three additional midnights plotted against the (binned) hospital-specific RDD coefficient for the cumulative fraction after two add. midnights

To summarize: Even within hospitals, the divergence between the cumulative fractions of discharged Johns and Maynards persists for several days in contrast to what a (potentially hospital-specific) fixed administrative rule would predict.

C-Section Probability Around Midnight

One covariate the smoothness of which I cannot directly observe is whether a newborn was delivered via C-Section. However, for mothers who were admitted to the hospital we can see whether they go on to give vaginal birth or have a C-Section (but we cannot link these mothers to their newborns). To get at least some sense of
whether there is some dramatic shift in the C-Section probability around midnight, Figure 2.9 shows the total number of admitted women who go on to become mothers and the share of these admitted mothers who end up having a C-Section plotted against the mother’s time of admittance (note that we cannot distinguish between the analysis sample and excluded births here, since the distinction depends on the newborn’s characteristics). Since the mothers are necessarily admitted prior to the child’s births (since our sample of newborns is restricted to newborns born in the hospital), any kind of discontinuity in the C-Section probability would show up some time before midnight.

Figure 2.9: C-Section probability of mothers admitted around midnight

The graph shows a downward trend in the C-Section probability over the course of the night and no sudden upward jump. There is a slight excess mass of C-Sections for mothers admitted exactly at midnight which is probably due to a rounding effect.
But this slight excess mass can impossibly explain the jump in length of stay at midnight, for three reasons. First, any effect of the excess C-Sections would show sometime later than midnight for the newborns, since the mothers are admitted some time prior to the actual time of birth (especially in the case of C-Sections). Second, the magnitude would be much too small to explain the shift in length of stay even if all C-Sections stayed twice as long. Third, the slight excess mass at midnight could only explain a short spike in length of stay, but not a sustained upward shift in length of stay for all the births past midnight.

**Changes of Shift**

Another concern for identification is that doctors may be having changes of shift at midnight. If this were the driving force behind the results, one would expect similar jumps at other times of the day. Figure 2.10 presents the number of births as well as the average length of stay in 5 minute bins around the clock from 11pm to 11pm (I chose to break the graph at 11pm, since we already know that there is no jump at 11pm from the earlier graphs). The only clear jump in length of stay occurs at midnight.

Figure 2.11 zooms in on the time period around 8am. A little after 8am in the morning there is a very steep yet smooth increase in the number of births as well as in length of stay. The rise in the number of births and length of stay is most likely due to scheduled C-sections taking place from around that time on (and most often around that time) and continuing throughout the day.
Figure 2.10: Frequency of births and average length of stay around the clock

Each dot shows the total number of births in the corresponding 5 minute bin for the analysis sample.

Each dot shows the average length of stay in hours in the corresponding 5 minute bin for the analysis sample. Length of stay is winsorized at the 1st and 99th percentile.
Each dot shows the total number of births in the corresponding 5 minute bin for the analysis sample.

Length of stay is winsorized at the 1st and 99th percentile.
The interpretation of scheduled C-Sections taking place in the morning is also supported by Figure 2.12. It shows the total number of admitted women who go on to become mothers and the share of these admitted mothers who end up having a C-Section plotted against the mother’s time of admittance to the hospital. Clearly, there is a strong uptick in the overall number of admitted mothers and the share of C-Section starting about two hours before the uptick in the number of births and average length of stay. This is consistent with mothers being admitted for scheduled C-Sections in the early morning and the actual births taking place about two hours later on average.

Figure 2.12: Count and C-Section probability of mothers admitted in the morning.
2.5 A Model of Limited Attention

This model serves an illustrative purpose. It formally demonstrates the intuitive logic that limited attention leads to the observed jumps in hazard rates.

Consider a number of newborns who were born around midnight between Monday and Tuesday (so the Johns and Maynards from the earlier example are part of the group). For simplicity, assume that these newborns can either be discharged at \( t_1 \) (Wednesday 9am), \( t_2 \) (Thursday 9am) or at \( t_3 \) (Friday 9am). Imagine a doctor being on her ward round at \( t_1 \) on Wednesday morning, examining a newborn \( i \). She needs to decide about whether the newborn can be discharged already. The newborn’s true health status is \( h_i \). The doctor forms an estimate \( \hat{h}_i \) and her decision rule is to discharge if the estimated health status is beyond a certain cutoff, i.e. whenever \( \hat{h}_i > \bar{h} \).

To form her estimate, the doctor examines the newborn. The examination yields health signal \( s_i \). The signal is a function of the true health status \( h_i \) and the age of the newborn \( age_i \) (one could add some noise term, but it would only complicate notation without adding anything of substance). The signal depends on the age of the newborn, because the signal conveys different information depending on the age. For instance, whether the digestion of the newborn functions properly can only be examined after the newborn ingested some food and had some time to digest. Hence, two (otherwise identical) newborns one of which has a digestion disease would show the same signal until a certain age is reached even though their underlying health status is different. Formally, \( s_i = s(age_i, h_i) \).

The doctor uses the signal \( s_i \) and the perceived age \( \hat{age}_i \) (which is a function of the true age and which might—in the case of an inattentive doctor—deviate from the true age) to make an estimate of \( h_i \). That is, \( \hat{h}_i = f(\hat{age}_i, s_i) \) for some function
\[ \hat{h}_i = f(a\gamma_i, s_i) \]
\[ = f(a\gamma_i, s(age_i, h_i)) \]
\[ \equiv \hat{h}(a\gamma_i, age_i, h_i) \]

and the newborn is discharged if \( \hat{h}_i = \hat{h}(a\gamma_i, age_i, h_i) > \bar{h} \). I assume \( \hat{h}(a\gamma_i, age_i, h_i) \) is smooth.

Let \( \tau_i \) denote time of birth for newborn \( i \). Moreover, let \( h_i \) be discrete and take on values \( h_1, ..., h_H \). \( \pi(h_i|\tau_i) \) denotes the probability mass function of types \( h_i \) conditional on time of birth. The identification assumption for the RDD amounts to \( \pi(h_i|\tau_i) \) being continuous in \( \tau_i \) at midnight.

Assume a unit mass of newborns is born at each point in time. The hazard rate at examination time \( t_1 \) (that is, Wednesday, at 9am) for newborns with time of birth \( \tau_i \) is

\[
hazard(t_1|\tau_i) = \sum_{h_1}^{h_H} I\left\{ \hat{h}(a\gamma_i, age_i, h_i) > \bar{h} \right\} \pi(h_i|\tau_i) \]
\[
= \sum_{h_1}^{h_H} I\left\{ \hat{h}(a\gamma_i, t - \tau_i, h_i) > \bar{h} \right\} \pi(h_i|\tau_i) \]

with \( I\{\cdot\} \) denoting the indicator function. \( a\gamma_i \) is a function of \( t \) and \( \tau_i \), but I omit the arguments here to simplify notation.

The decision cutoff for the estimated health is \( \bar{h} \). This maps into a decision cutoff for the true health status. Define the relevant true health cutoff for discharge \( \hat{h}(t, \tau_i) \) implicitly as \( \hat{h}(a\gamma_i, t - \tau_i, \hat{h}(t, \tau_i)) = \bar{h} \). That is, newborn \( i \) that is born at \( \tau_i \) is discharged at \( t_1 \) if \( h_i > \hat{h}(t_1, \tau_i) \), since \( \hat{h}(\cdot) \) is monotonically increasing in its third argument.
To investigate the presence of a jump in the hazard at $t_1$ for newborns born around midnight, we need to compare the limiting hazards from below and above.

\[
\lim_{\tau \uparrow \text{midnight}} \text{hazard }(t_1 | \tau_i) = \sum_{h_1} \pi \left( h \left( \lim_{\tau \uparrow \text{midnight}} \hat{a} \hat{g} e_i, t_1 - \text{midnight}, h_i \right) > h \right) \pi (h_i | \text{midnight})
\]

\[
= \sum_{h_1} \pi \left( h \left( \lim_{\tau \uparrow \text{midnight}} \hat{a} \hat{g} e_i, 33 \text{hours}, h_i \right) > h \right) \pi (h_i | \text{midnight})
\]

due to the smoothness of $\hat{h}$ and $\pi$, the fact that $\hat{a} \hat{g} e_i = t_1 - \tau_i$ and because $t_1 - \text{midnight}$ is 33 hours in our example with the first examination taking place at $t_1$, Wednesday at 9am. So what matters for whether the two limiting hazard rates are different is whether doctor’s perceived age of the newborn $\hat{a} \hat{g} e_i$ is different depending on whether the newborn was born just before or just after midnight.

**Attentive Doctor**

If the doctor pays full attention, she will use the true (continuous) measure of age to make her estimate. In that case at $t_1$

\[
\lim_{\tau \uparrow \text{midnight}} \hat{a} \hat{g} e_i = 33 \text{hours} = \lim_{\tau \downarrow \text{midnight}} \hat{a} \hat{g} e_i
\]
and, hence,

\[
\lim_{\tau \uparrow \text{midnight}} \text{hazard} (t_1 | \tau_i) = \sum_{h_1}^h I \left\{ \hat{h} (33\text{hours}, 33\text{hours}, h_i) > \bar{h} \right\} \pi (h_i | \text{midnight}) \\
= \lim_{\tau \downarrow \text{midnight}} \text{hazard} (t_1 | \tau_i)
\]

Thus, no jump in hazard at \( t_1 \) is expected under the identifying smoothness assumptions regarding \( \pi (h_i | \tau_i) \). Following the same logic, there also are no jumps in the hazard rates at later points in time.

**Inattentive Doctor**

An inattentive doctor might only consider the number of days (which is counted as the number of midnights crossed) as the newborn’s age. In that case when deciding about discharge at \( t_1 \)

\[
\lim_{\tau \uparrow \text{midnight}} \text{age}_i = 48\text{hours}
\]

and

\[
\lim_{\tau \downarrow \text{midnight}} \text{age}_i = 24\text{hours}
\]

, since the perceived ages are two days, respectively, one day.

Hence,

\[
\lim_{\tau \uparrow \text{midnight}} \text{hazard} (t_1 | \tau_i) = \sum_{h_1}^h I \left\{ \hat{h} (48\text{hours}, 33\text{hours}, h_i) > \bar{h} \right\} \pi (h_i | \text{midnight}) \\
\neq \sum_{h_1}^h I \left\{ \hat{h} (24\text{hours}, 33\text{hours}, h_i) > \bar{h} \right\} \pi (h_i | \text{midnight}) \\
= \lim_{\tau \downarrow \text{midnight}} \text{hazard} (t | \tau_i)
\]

and so we expect a jump in hazard at midnight at \( t_1 \).

In the data, \( \lim_{\tau \uparrow \text{midnight}} \text{hazard} (t_1 | \tau_i) \) \( > \) \( \lim_{\tau \downarrow \text{midnight}} \text{hazard} (t | \tau_i) \), implying that perceived age has a positive effect on perceived health conditional on true age and true health (i.e. conditional on the health signal \( s_i \)). As discussed before, this implication
is intuitive, since certain diseases only manifest in a negative health signal after a certain time (e.g. digestion diseases). Hence, conditional on the health signal, the perceived age should enter the estimated true health status positively.

At \( t_2 \), there are now two effects that can give rise to differential hazard rates. First, the ages are still perceived differently and, second, the newborns are now differentially selected, i.e. the health cutoffs at \( t_1 \) \( \lim_{\tau \uparrow \text{midnight}} \tilde{h}(t_1, \tau_i) \) and \( \lim_{\tau \downarrow \text{midnight}} \tilde{h}(t_1, \tau_i) \) were not the same, since the decision rules at \( t_1 \) \( \hat{h}(48\text{hours}, 33\text{hours}, h_i) > \bar{h} \) and \( \hat{h}(24\text{hours}, 33\text{hours}, h_i) > \bar{h} \) were different.

At \( t_2 \), the limiting hazards are

\[
\lim_{\tau \uparrow \text{midnight}} \text{hazard}(t_2 | \tau_i) = \sum_{h_1} \left\{ \hat{h}(72\text{hours}, 57\text{hours}, h_i) > \bar{h} \right\} \\
\cdot \pi\left( h_i \mid \text{midnight}, h_i \leq \lim_{\tau \uparrow \text{midnight}} \tilde{h}(t_1, \tau_i) \right)
\]

\[
\lim_{\tau \downarrow \text{midnight}} \text{hazard}(t_2 | \tau_i) = \sum_{h_1} \left\{ \hat{h}(48\text{hours}, 57\text{hours}, h_i) > \bar{h} \right\} \\
\cdot \pi\left( h_i \mid \text{midnight}, h_i \leq \lim_{\tau \downarrow \text{midnight}} \tilde{h}(t_1, \tau_i) \right)
\]

On the one hand, \( \lim_{\tau \uparrow \text{midnight}} \text{hazard}(t_2 | \tau_i) \) should be larger than \( \lim_{\tau \downarrow \text{midnight}} \text{hazard}(t_2 | \tau_i) \), because perceived health depends positively on perceived age. On the other hand, \( \lim_{\tau \uparrow \text{midnight}} \tilde{h}(t_1, \tau_i) < \lim_{\tau \downarrow \text{midnight}} \tilde{h}(t_1, \tau_i) \) and, thus, the pool of newborns from before midnight is negatively selected. In the data, the first effect dominates for the most part, since (see Table 2.12) the hazard rates jump down at midnight at one, two, three and four additional midnights. At five and seven additional midnights, however, the hazard rates jump up, implying that the selection effect dominates the effect of perceived age.

**Implications for the share of inattentive doctors**
The results in Tables 2.10 and 2.11 in the appendix show that newborns born just after midnight stay stay about 0.3 nights longer on average while being discharged around the same time of the day. That is, about 30% of the newborns born around midnight are staying an additional night in the hospital as a result of being born just after midnight instead of just before. This implies that in at least 30% of cases the doctors are not paying attention to the newborn’s precise age. The 30% is a lower bound on the share of cases with inattentive doctors, because the inattention is necessary but not sufficient for being born before or after midnight to be relevant for the length of stay. For being born before or after midnight to be relevant, the doctor needs to be inattentive and the resulting difference in the perceived age of the newborn needs to trigger a different discharge decision.

If, for instance, the newborn’s age did not affect the estimate the doctor makes of her true health — i.e. \( \hat{h}_i = f(s_i) \) instead of \( \hat{h}_i = f(\hat{\text{age}}_i, s_i) \) — we would not see any effect of crossing midnight on the newborn’s length of stay even for inattentive doctors. If, conversely, the estimated health depended solely on the newborn’s age (that is, if the doctors used a simple rule of thumb after how many days to discharge, irrespective of any health signal \( s_i \)) — i.e. \( \hat{h}_i = f(\hat{\text{age}}_i) \) and not \( \hat{h}_i = f(\hat{\text{age}}_i, s_i) \) — an inattentive doctor would always result in the newborn staying one additional night.

Therefore, the 30% is a lower bound on the share of inattentive doctors as well as a lower bound for the share of cases in which the perceived age of the newborn is pivotal for the decision.

2.6 Conclusion

There are two main conclusions from this paper.

First, the interpretation of the results in Almond and Doyle (2011) needs to be reconsidered. Almond and Doyle find a jump in length of stay for Californian
newborns born around midnight and attribute it to moral hazard with respect to the 24 hours of additional insurance coverage that being born after midnight generates in the U.S.. Since this paper finds a similarly sized jump for German newborns for whom there is no jump in insurance coverage, the original interpretation of the Almond and Doyle paper is questionable. As a general lesson, I conclude that researchers should be careful when using sharp cutoffs for identification if these cutoffs could also result in a change in outcomes by triggering different rule of thumb behavior.

Second, this paper provides evidence of limited attention by professionals in a high-stakes situation, contributing to the literature studying whether the well-established phenomena from behavioral economics remain present in such settings. Specifically, my results imply that at least 30% of newborns are treated by doctors who pay only limited attention to the newborn’s precise age. One policy implication is that institutions (such as documentation standards) should be designed in a way that helps doctors overcome biases and limited attention. The results also suggest that doctors may be affected by nudges and that this might be an underutilized tool to direct health care decisions in a desired direction.

References


Appendix

2.A Further Tables
Table 2.5: Frequency of births (for the excluded births)

<table>
<thead>
<tr>
<th></th>
<th>30min window</th>
<th>60min window</th>
<th>90min window</th>
<th>120min window</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full analysis sample</td>
<td>-40.48 (71.9)</td>
<td>-24.81 (44.99)</td>
<td>-20.97 (34.1)</td>
<td>-21.75 (29.17)</td>
</tr>
<tr>
<td>Obs</td>
<td>60</td>
<td>120</td>
<td>180</td>
<td>240</td>
</tr>
<tr>
<td>MeanDep</td>
<td>681.28</td>
<td>674.62</td>
<td>674.01</td>
<td>671.33</td>
</tr>
<tr>
<td>Donut estimator</td>
<td>34.86 (82.39)</td>
<td>7.15 (46.55)</td>
<td>-1.05 (34.08)</td>
<td>-7.76 (29.27)</td>
</tr>
<tr>
<td>Obs</td>
<td>50</td>
<td>110</td>
<td>170</td>
<td>230</td>
</tr>
<tr>
<td>MeanDep</td>
<td>677.5</td>
<td>672.3</td>
<td>672.46</td>
<td>670.07</td>
</tr>
<tr>
<td>Donut and no 5 min mult</td>
<td>-2.19 (46.65)</td>
<td>-2.78 (27.26)</td>
<td>-7.06 (20.05)</td>
<td>-14.06 (16.92)</td>
</tr>
<tr>
<td>Obs</td>
<td>40</td>
<td>88</td>
<td>136</td>
<td>184</td>
</tr>
<tr>
<td>MeanDep</td>
<td>633.7</td>
<td>629.08</td>
<td>628.41</td>
<td>625.6</td>
</tr>
</tbody>
</table>

Shows results for regressing the count per minute on an indicator for whether midnight has been crossed controlling for a linear term (not weighted, i.e. plain OLS) in minute of birth the slope of which is allowed to change at the cutoff. The sample consists of the excluded births. The columns correspond to different time windows around midnight. ‘Full analysis sample’, ‘Donut estimator’ and ‘Donut and no 5 min mult’ correspond to the estimated coefficient and standard error from the full analysis sample, a donut estimator excluding the 5 minute window around midnight and a donut estimator that excludes the 5 minute window around midnight as well as all multiples of 5 minutes. ‘Obs’ and ‘MeanDep’ correspond to the number of observations and the mean of the dependent variable. *: p<.1, **: p<.05, ***: p<.01
Table 2.6: Birth weight

<table>
<thead>
<tr>
<th></th>
<th>30min window</th>
<th>60min window</th>
<th>90min window</th>
<th>120min window</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>5.02 (3.81)</td>
<td>0.55 (2.95)</td>
<td>-2.49 (2.57)</td>
<td>-0.98 (2.27)</td>
</tr>
<tr>
<td>Cluster</td>
<td>60</td>
<td>120</td>
<td>180</td>
<td>240</td>
</tr>
<tr>
<td>Obs</td>
<td>163466</td>
<td>325364</td>
<td>485948</td>
<td>644469</td>
</tr>
<tr>
<td>MeanDep</td>
<td>3443.42</td>
<td>3443.25</td>
<td>3443.08</td>
<td>3443.57</td>
</tr>
<tr>
<td>Donut</td>
<td>1.71 (5.94)</td>
<td>-2.26 (3.67)</td>
<td>-4.93* (2.95)</td>
<td>-2.43 (2.52)</td>
</tr>
<tr>
<td>Cluster</td>
<td>50</td>
<td>110</td>
<td>170</td>
<td>230</td>
</tr>
<tr>
<td>Obs</td>
<td>135893</td>
<td>297791</td>
<td>458375</td>
<td>616896</td>
</tr>
<tr>
<td>MeanDep</td>
<td>3443.25</td>
<td>3443.16</td>
<td>3443</td>
<td>3443.53</td>
</tr>
<tr>
<td>+ 5 min mult</td>
<td>-2.95 (6.54)</td>
<td>-2.38 (4.12)</td>
<td>-3.19 (3.41)</td>
<td>-1.69 (2.86)</td>
</tr>
<tr>
<td>Cluster</td>
<td>40</td>
<td>88</td>
<td>136</td>
<td>184</td>
</tr>
<tr>
<td>Obs</td>
<td>101409</td>
<td>222579</td>
<td>342974</td>
<td>461563</td>
</tr>
<tr>
<td>MeanDep</td>
<td>3442.15</td>
<td>3442.4</td>
<td>3442.39</td>
<td>3442.79</td>
</tr>
</tbody>
</table>

Shows results for regressions using the analysis sample of birth weight on an indicator for whether midnight has been crossed controlling for a linear term (not weighted, i.e. plain OLS) in minute of birth that is allowed to change at the cutoff. Standard errors are clustered at the minute level. The columns correspond to different time windows around midnight. 'Full', 'Donut' and '+ 5 min mult' correspond to the estimated coefficient and standard error from the full sample, a donut estimator excluding the 5 minute window around midnight and a donut estimator that in addition also excludes all multiples of 5. 'Cluster', 'Obs' and 'MeanDep' correspond to the number of clusters, number of observations and the mean of the dependent variable. *: p<.1, **: p<.05, ***: p<.01
Table 2.7: Share female

<table>
<thead>
<tr>
<th></th>
<th>30min window</th>
<th>60min window</th>
<th>90min window</th>
<th>120min window</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>0.0118***</td>
<td>0.0066**</td>
<td>0.0064**</td>
<td>0.0045*</td>
</tr>
<tr>
<td></td>
<td>(0.0038)</td>
<td>(0.003)</td>
<td>(0.0025)</td>
<td>(0.0023)</td>
</tr>
<tr>
<td>Cluster</td>
<td>60, 120, 180</td>
<td>180</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>Obs</td>
<td>163466</td>
<td>325364</td>
<td>485948</td>
<td>644469</td>
</tr>
<tr>
<td>MeanDep</td>
<td>0.49</td>
<td>0.4908</td>
<td>0.4908</td>
<td>0.4905</td>
</tr>
<tr>
<td>Donut</td>
<td>0.0061</td>
<td>0.0025</td>
<td>0.0039</td>
<td>0.0023</td>
</tr>
<tr>
<td></td>
<td>(0.0057)</td>
<td>(0.0035)</td>
<td>(0.0028)</td>
<td>(0.0025)</td>
</tr>
<tr>
<td>Cluster</td>
<td>50, 110, 170</td>
<td>170</td>
<td>230</td>
<td></td>
</tr>
<tr>
<td>Obs</td>
<td>135893</td>
<td>297791</td>
<td>458375</td>
<td>616896</td>
</tr>
<tr>
<td>MeanDep</td>
<td>0.4904</td>
<td>0.491</td>
<td>0.491</td>
<td>0.4906</td>
</tr>
<tr>
<td>+ 5 min mult</td>
<td>0.0048</td>
<td>0.0019</td>
<td>0.0024</td>
<td>0.0017</td>
</tr>
<tr>
<td></td>
<td>(0.0076)</td>
<td>(0.0045)</td>
<td>(0.0035)</td>
<td>(0.0031)</td>
</tr>
<tr>
<td>Cluster</td>
<td>40, 88, 136</td>
<td>136</td>
<td>184</td>
<td></td>
</tr>
<tr>
<td>Obs</td>
<td>101409</td>
<td>222579</td>
<td>342974</td>
<td>461563</td>
</tr>
<tr>
<td>MeanDep</td>
<td>0.4904</td>
<td>0.4913</td>
<td>0.4909</td>
<td>0.4903</td>
</tr>
</tbody>
</table>

Shows results for regressions using the analysis sample of an indicator for a female newborn on an indicator for whether midnight has been crossed controlling for a linear term (not weighted, i.e. plain OLS) in minute of birth that is allowed to change at the cutoff. Standard errors are clustered at the minute level. The columns correspond to different time windows around midnight. 'Full', 'Donut' and '+ 5 min mult' correspond to the estimated coefficient and standard error from the full sample, a donut estimator excluding the 5 minute window around midnight and a donut estimator that in addition also excludes all multiples of 5. 'Cluster', 'Obs' and 'MeanDep' correspond to the number of clusters, number of observations and the mean of the dependent variable. *: p<.1, **: p<.05, ***: p<.01
Table 2.8: Number of Diagnoses

<table>
<thead>
<tr>
<th></th>
<th>30min window</th>
<th>60min window</th>
<th>90min window</th>
<th>120min window</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full</strong></td>
<td>0.008 (0.008)</td>
<td>0.005 (0.006)</td>
<td>0.003 (0.005)</td>
<td>0.006 (0.005)</td>
</tr>
<tr>
<td><strong>Cluster</strong></td>
<td>60</td>
<td>120</td>
<td>180</td>
<td>240</td>
</tr>
<tr>
<td><strong>Obs</strong></td>
<td>163466</td>
<td>325364</td>
<td>485948</td>
<td>644469</td>
</tr>
<tr>
<td><strong>MeanDep</strong></td>
<td>0.382</td>
<td>0.379</td>
<td>0.378</td>
<td>0.379</td>
</tr>
<tr>
<td><strong>Donut</strong></td>
<td>0.002 (0.01)</td>
<td>0.002 (0.007)</td>
<td>0.001 (0.006)</td>
<td>0.005 (0.005)</td>
</tr>
<tr>
<td><strong>Cluster</strong></td>
<td>50</td>
<td>110</td>
<td>170</td>
<td>230</td>
</tr>
<tr>
<td><strong>Obs</strong></td>
<td>135893</td>
<td>297791</td>
<td>458375</td>
<td>616896</td>
</tr>
<tr>
<td><strong>MeanDep</strong></td>
<td>0.381</td>
<td>0.379</td>
<td>0.378</td>
<td>0.379</td>
</tr>
<tr>
<td><strong>+ 5 min mult</strong></td>
<td>0.011 (0.007)</td>
<td>0.004 (0.006)</td>
<td>0 (0.005)</td>
<td>0.003 (0.007)</td>
</tr>
<tr>
<td><strong>Cluster</strong></td>
<td>40</td>
<td>88</td>
<td>136</td>
<td>184</td>
</tr>
<tr>
<td><strong>Obs</strong></td>
<td>101409</td>
<td>222579</td>
<td>342974</td>
<td>461563</td>
</tr>
<tr>
<td><strong>MeanDep</strong></td>
<td>0.378</td>
<td>0.376</td>
<td>0.375</td>
<td>0.377</td>
</tr>
</tbody>
</table>

Shows results for regressions using the analysis sample of the number of diagnoses on an indicator for whether midnight has been crossed controlling for a linear term (not weighted, i.e. plain OLS) in minute of birth that is allowed to change at the cutoff. Standard errors are clustered at the minute level. The columns correspond to different time windows around midnight. ‘Full’, ‘Donut’ and ‘+ 5 min mult’ correspond to the estimated coefficient and standard error from the full sample, a donut estimator excluding the 5 minute window around midnight and a donut estimator that in addition also excludes all multiples of 5. ‘Cluster’, ‘Obs’ and ‘MeanDep’ correspond to the number of clusters, number of observations and the mean of the dependent variable. *: p<.1, **: p<.05, ***: p<.01
<table>
<thead>
<tr>
<th></th>
<th>30min window</th>
<th>60min window</th>
<th>90min window</th>
<th>120min window</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full</strong></td>
<td>0.007 (0.007)</td>
<td>0.01** (0.005)</td>
<td>0.008** (0.004)</td>
<td>0.006 (0.004)</td>
</tr>
<tr>
<td><strong>Cluster</strong></td>
<td>60</td>
<td>120</td>
<td>180</td>
<td>240</td>
</tr>
<tr>
<td><strong>Obs</strong></td>
<td>163466</td>
<td>325364</td>
<td>485948</td>
<td>644469</td>
</tr>
<tr>
<td><strong>MeanDep</strong></td>
<td>1.488</td>
<td>1.488</td>
<td>1.489</td>
<td>1.489</td>
</tr>
<tr>
<td><strong>Donut</strong></td>
<td>-0.001 (0.013)</td>
<td>0.008 (0.007)</td>
<td>0.006 (0.005)</td>
<td>0.005 (0.004)</td>
</tr>
<tr>
<td><strong>Cluster</strong></td>
<td>50</td>
<td>110</td>
<td>170</td>
<td>230</td>
</tr>
<tr>
<td><strong>Obs</strong></td>
<td>135893</td>
<td>297791</td>
<td>458375</td>
<td>616896</td>
</tr>
<tr>
<td><strong>MeanDep</strong></td>
<td>1.488</td>
<td>1.488</td>
<td>1.489</td>
<td>1.489</td>
</tr>
<tr>
<td>+ 5 min mult</td>
<td>0.008 (0.006)</td>
<td>0.008 (0.005)</td>
<td>0.003 (0.005)</td>
<td>-0.001 (0.006)</td>
</tr>
<tr>
<td><strong>Cluster</strong></td>
<td>40</td>
<td>88</td>
<td>136</td>
<td>184</td>
</tr>
<tr>
<td><strong>Obs</strong></td>
<td>101409</td>
<td>222579</td>
<td>342974</td>
<td>461563</td>
</tr>
<tr>
<td><strong>MeanDep</strong></td>
<td>1.487</td>
<td>1.486</td>
<td>1.488</td>
<td>1.488</td>
</tr>
</tbody>
</table>

Shows results for regressions using the analysis sample of the number of procedures on an indicator for whether midnight has been crossed controlling for a linear term (not weighted, i.e. plain OLS) in minute of birth that is allowed to change at the cutoff. Standard errors are clustered at the minute level. The columns correspond to different time windows around midnight. 'Full', 'Donut' and '+ 5 min mult' correspond to the estimated coefficient and standard error from the full sample, a donut estimator excluding the 5 minute window around midnight and a donut estimator that in addition also excludes all multiples of 5. 'Cluster', 'Obs' and 'MeanDep' correspond to the number of clusters, number of observations and the mean of the dependent variable. *: p<.1, **: p<.05, ***: p<.01
Table 2.10: Number of additional midnights

<table>
<thead>
<tr>
<th></th>
<th>30min window</th>
<th>60min window</th>
<th>90min window</th>
<th>120min window</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>0.3*** (0.02)</td>
<td>0.3*** (0.01)</td>
<td>0.29*** (0.01)</td>
<td>0.29*** (0.01)</td>
</tr>
<tr>
<td>Cluster</td>
<td>60</td>
<td>120</td>
<td>180</td>
<td>240</td>
</tr>
<tr>
<td>Obs</td>
<td>163466</td>
<td>325364</td>
<td>485948</td>
<td>644469</td>
</tr>
<tr>
<td>MeanDep</td>
<td>3.48</td>
<td>3.47</td>
<td>3.47</td>
<td>3.47</td>
</tr>
<tr>
<td>Donut</td>
<td>0.27*** (0.02)</td>
<td>0.29*** (0.01)</td>
<td>0.29*** (0.01)</td>
<td>0.29*** (0.01)</td>
</tr>
<tr>
<td>Cluster</td>
<td>50</td>
<td>110</td>
<td>170</td>
<td>230</td>
</tr>
<tr>
<td>Obs</td>
<td>135893</td>
<td>297791</td>
<td>458375</td>
<td>616896</td>
</tr>
<tr>
<td>MeanDep</td>
<td>3.47</td>
<td>3.47</td>
<td>3.47</td>
<td>3.46</td>
</tr>
<tr>
<td>+ 5 min mult</td>
<td>0.28*** (0.03)</td>
<td>0.3*** (0.02)</td>
<td>0.29*** (0.01)</td>
<td>0.29*** (0.01)</td>
</tr>
<tr>
<td>Cluster</td>
<td>40</td>
<td>88</td>
<td>136</td>
<td>184</td>
</tr>
<tr>
<td>Obs</td>
<td>101409</td>
<td>222579</td>
<td>342974</td>
<td>461563</td>
</tr>
<tr>
<td>MeanDep</td>
<td>3.47</td>
<td>3.48</td>
<td>3.47</td>
<td>3.47</td>
</tr>
<tr>
<td>+ controls</td>
<td>0.27*** (0.02)</td>
<td>0.3*** (0.01)</td>
<td>0.29*** (0.01)</td>
<td>0.29*** (0.01)</td>
</tr>
<tr>
<td>Cluster</td>
<td>40</td>
<td>88</td>
<td>136</td>
<td>184</td>
</tr>
<tr>
<td>Obs</td>
<td>101409</td>
<td>222579</td>
<td>342974</td>
<td>461563</td>
</tr>
<tr>
<td>MeanDep</td>
<td>3.47</td>
<td>3.48</td>
<td>3.47</td>
<td>3.47</td>
</tr>
</tbody>
</table>

Shows results for regressions using the analysis sample of additional midnights (excluding the one defining the RDD threshold) in hospital on an indicator for whether midnight has been crossed controlling for a linear term (not weighted, i.e. plain OLS) in minute of birth that is allowed to change at the cutoff. Standard errors are clustered at the minute level. The columns correspond to different time windows around midnight. 'Full', 'Donut', '+ 5 min mult' and '+controls' correspond to the estimated coefficient and standard error from the full sample, a donut estimator excluding the 5 minute window around midnight, a donut estimator that in addition also excludes all multiples of 5 and the donut estimator excluding the 5 minute multiples plus controls, that is a quadratic in birth weight, an indicator for female, and year and hospital fixed effects. 'Cluster', 'Obs' and 'MeanDep' correspond to the number of clusters, number of observations and the mean of the dependent variable. *: p<.1, **: p<.05, ***: p<.01
<table>
<thead>
<tr>
<th></th>
<th>30min window</th>
<th>60min window</th>
<th>90min window</th>
<th>120min window</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>-2.096 (1.553)</td>
<td>-0.769 (1.128)</td>
<td>-0.022 (0.936)</td>
<td>0.192 (0.862)</td>
</tr>
<tr>
<td>Obs</td>
<td>163466</td>
<td>325364</td>
<td>485948</td>
<td>644469</td>
</tr>
<tr>
<td>MeanDep</td>
<td>813.951</td>
<td>814.001</td>
<td>814.084</td>
<td>813.937</td>
</tr>
<tr>
<td>Cluster</td>
<td>60</td>
<td>120</td>
<td>180</td>
<td>240</td>
</tr>
<tr>
<td>Donut</td>
<td>-2.863 (2.23)</td>
<td>-0.517 (1.363)</td>
<td>0.318 (1.062)</td>
<td>0.476 (0.952)</td>
</tr>
<tr>
<td>Cluster</td>
<td>50</td>
<td>110</td>
<td>170</td>
<td>230</td>
</tr>
<tr>
<td>Obs</td>
<td>135893</td>
<td>297791</td>
<td>458375</td>
<td>616896</td>
</tr>
<tr>
<td>MeanDep</td>
<td>814.121</td>
<td>814.084</td>
<td>814.143</td>
<td>813.974</td>
</tr>
<tr>
<td>+ 5 min mult</td>
<td>-2.993 (2.641)</td>
<td>-1.614 (1.48)</td>
<td>-0.178 (1.17)</td>
<td>-0.117 (1.012)</td>
</tr>
<tr>
<td>Cluster</td>
<td>40</td>
<td>88</td>
<td>136</td>
<td>184</td>
</tr>
<tr>
<td>Obs</td>
<td>101409</td>
<td>222579</td>
<td>342974</td>
<td>461563</td>
</tr>
<tr>
<td>MeanDep</td>
<td>814.856</td>
<td>814.938</td>
<td>814.894</td>
<td>814.807</td>
</tr>
<tr>
<td>+ controls</td>
<td>-2.874 (2.569)</td>
<td>-1.167 (1.499)</td>
<td>-0.14 (1.161)</td>
<td>-0.234 (0.977)</td>
</tr>
<tr>
<td>Cluster</td>
<td>40</td>
<td>88</td>
<td>136</td>
<td>184</td>
</tr>
<tr>
<td>Obs</td>
<td>101409</td>
<td>222579</td>
<td>342974</td>
<td>461563</td>
</tr>
<tr>
<td>MeanDep</td>
<td>814.856</td>
<td>814.938</td>
<td>814.894</td>
<td>814.807</td>
</tr>
</tbody>
</table>

Shows results for regressions using the analysis sample of discharge time of the day in minutes on an indicator for whether midnight has been crossed controlling for a linear term (not weighted, i.e. plain OLS) in minute of birth that is allowed to change at the cutoff. Standard errors are clustered at the minute level. The columns correspond to different time windows around midnight. 'Full', 'Donut', '+ 5 min mult' and '+controls' correspond to the estimated coefficient and standard error from the full sample, a donut estimator excluding the 5 minute window around midnight, a donut estimator that in addition also excludes all multiples of 5 and the donut estimator excluding the 5 minute multiples plus controls, that is a quadratic in birth weight, an indicator for female, and year and hospital fixed effects. 'Cluster', 'Obs' and 'MeanDep' correspond to the number of clusters, number of observations and the mean of the dependent variable. *: p<.1, **: p<.05, ***: p<.01
<table>
<thead>
<tr>
<th>Outcome</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>Cluster</th>
<th>Obs</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hazard after 0 add. midnights</td>
<td>-0.00096</td>
<td>0.00116</td>
<td>240</td>
<td>644425</td>
<td>0.0595</td>
</tr>
<tr>
<td>Hazard after 1 add. midnights</td>
<td>-0.01109***</td>
<td>0.00121</td>
<td>240</td>
<td>606091</td>
<td>0.063</td>
</tr>
<tr>
<td>Hazard after 2 add. midnights</td>
<td>-0.12461***</td>
<td>0.00269</td>
<td>240</td>
<td>567891</td>
<td>0.3248</td>
</tr>
<tr>
<td>Hazard after 3 add. midnights</td>
<td>-0.02598***</td>
<td>0.00371</td>
<td>240</td>
<td>383415</td>
<td>0.5106</td>
</tr>
<tr>
<td>Hazard after 4 add. midnights</td>
<td>-0.04188***</td>
<td>0.00447</td>
<td>240</td>
<td>187661</td>
<td>0.5227</td>
</tr>
<tr>
<td>Hazard after 5 add. midnights</td>
<td>0.0305***</td>
<td>0.00711</td>
<td>240</td>
<td>89565</td>
<td>0.5315</td>
</tr>
<tr>
<td>Hazard after 6 add. midnights</td>
<td>-0.02366**</td>
<td>0.00962</td>
<td>240</td>
<td>41959</td>
<td>0.5212</td>
</tr>
<tr>
<td>Hazard after 7 add. midnights</td>
<td>0.09103***</td>
<td>0.01421</td>
<td>240</td>
<td>20091</td>
<td>0.4941</td>
</tr>
</tbody>
</table>

Table 2.12: Hazard rates

Shows results for regressing the hazard rate after a certain number of additional midnights (excluding the midnight defining the RDD threshold) in the hospital on an indicator for whether the newborn’s time of birth was after midnight, controlling for a linear term (not weighted, i.e. plain OLS) in minute of birth that is allowed to change at the midnight cutoff. The sample is the analysis sample in a +/- 120 minute bandwidth around midnight. Standard errors are clustered at the minute level. 'Cluster', 'Obs' and 'Mean' correspond to the number of clusters, number of observations and the mean of the dependent variable. *: p<.1, **: p<.05, ***: p<.01
Chapter 3

Optimal Unemployment Insurance if the Tax System is Progressive

3.1 Introduction

The so-called Baily-Chetty formula — introduced by Baily (1978) — is the standard framework for thinking about optimal unemployment insurance. The formula expresses the optimal benefit level in terms of three estimable parameters: Risk aversion, the relative drop in consumption when becoming unemployed and the elasticity of average unemployment duration with respect to the benefit level.\footnote{This is under a quadratic approximation of the utility function and under the assumption that the probability of becoming unemployed in the first place is not affected by the generosity of the UI system — this assumption finds support in a regression discontinuity design for Germany (see Schmieder, von Wachter, and Bender 2012) as well as a regression kink design for the US (see Landais 2015)}

Chetty (2006) and Chetty (2008) show that the formula is robust to various extensions (e.g. arbitrary borrowing constraints) and that even with heterogeneity in utility functions and wages, the elasticity of average unemployment duration w.r.t. UI benefits is a sufficient statistic and no other information about the behavioral responses is needed. In this paper I show that the elasticity of average unemployment...
duration ceases to be sufficient once one allows for a realistic tax system, i.e. progressive labor income taxes with increasing average tax rates. In this more general environment the individual elasticities have to be weighted by the tax amount the individual would pay if she were employed. That is, for a given degree of risk aversion and a given unemployment consumption drop, the Baily-Chetty formula overstates (understates) the optimal benefit level if the high earners’ durations react more (less) elastically than the low earners’.

Intuitively, the social welfare costs of having one month less of employment depend on how much surplus this month of working would have created. Since the higher earners pay higher taxes, there is bigger wedge between the marginal social returns (the additional production) and the marginal private costs (less leisure) of employment. Hence, at the margin, one less month of employment for high earners reduces social welfare by more than one less month of employment for low earners. Therefore, the elasticity of unemployment durations has to be weighted by the labor income tax.

In the empirical section I investigate the degree of heterogeneity in the elasticity of unemployment duration across income groups, using U.S. data and exploiting within state variation over time in the generosity of the unemployment insurance system. In particular, I instrument for individual benefits using leave-one-out averages\(^2\) for each state-year-cell in order to isolate variation in benefits that is driven by state-level changes in UI generosity.

Using the described IV design, I find modestly sized behavioral responses and little heterogeneity. In the main specification, the estimated elasticity of unemployment durations with respect to the benefit level is 0.22 for high earners and 0.15 for low earners. The size of these estimates is smaller than what has previously been found

\(^2\)leave-one-out average meaning the average benefit in the individuals state-year-cell leaving out the benefit of the individual whose benefit I instrument for.
in the literature. The theoretical part of this paper is closely related to two other recent papers analyzing optimal unemployment insurance.

Lawson (2014) extends the Baily-Chetty framework by allowing for a government that collects taxes not only to finance unemployment benefits, but also for other government expenditures. He demonstrates that the optimal level of unemployment benefits is smaller when taking into account the existence of taxes for other purposes. I build on his key insight, introducing heterogeneous (i.e. progressive) tax rates. I demonstrate that it is the average labor taxes of the unemployed (that is, the average taxes they would pay if they had a job), not the average labor taxes across the population that is important for the interaction between the UI system and the labor tax. Since low earners tend to be unemployed more often, this reduces the labor tax rate that one should use to calibrate Lawson (2014).

Andrews and Miller (2013) study the implications of heterogeneity on the optimal level of social insurance. They focus on heterogeneity in risk aversion and find that the covariance between risk aversion and the unemployment consumption drop is crucial to the optimal design of unemployment insurance. They focus on the impacts of heterogeneous preferences on the consumption smoothing benefit, while I study the implications of heterogeneous tax rates on the fiscal costs. Therefore, this paper is highly complementary to theirs. In two of the cases they study they find that heterogeneity in the behavioral response matters. While related, the heterogeneity matters for conceptually different reasons than what this paper focuses on.3

The empirical part of this paper is closely related to Chetty (2008). Like my paper, he uses data from the Survey of Income and Program Participation (SIPP). His

---

3The first of their cases in which heterogeneity in elasticities matters is when UI taxes are set fairly within an individual — that is, higher risk individuals pay higher taxes and individual elasticities matter. This is not a scenario I consider and conceptually different from the issues arising due to the progressivity of the tax system. The second case is when they consider proportional UI benefits and taxes and analyze optimality using a small policy change in relative not absolute benefits. In the main part I also study the case of proportional benefits and show that the presence of a progressive tax systems further strengthens their effect.
identification strategy is different from this paper’s, because he proxies for individual benefits using three different variables\(^4\) instead of instrumenting. Moreover, his focus is on heterogeneity across liquidity groups instead of income groups. Interestingly, he finds bigger elasticities than I do. Moreover, he finds that the elasticity is negatively correlated with assets, while I find no such pattern.

### 3.2 Theory

I present a derivation for the static case. Under the regularity conditions from Chetty (2006) the formulas also apply when one allows for dynamics (including self-insurance and borrowing constraints).\(^5\)

The sufficient statistics approach to optimal UI requires to restrict the set of policies that is optimized over. First, I consider the case of uniform benefits and UI taxes across people (but allowing for general labor income taxation). Second, I study benefits and UI taxes that are proportional to previous wages.

**Uniform benefits and UI taxes**

Consider an economy with a continuum of agents with mass one. There are different types of agents indexed by \(i = 1, \ldots, I\). The mass of type \(i\) is denoted by \(Pr[i]\). An agent of type \(i\) has utility function \(v_i(c_i)\) when employed and \(u_i(c_i)\) when unemployed with \(c_i\) denoting consumption.

Agent \(i\) chooses search effort \(p_i\) and pays search costs \(\psi_i(p_i)\). \(p_i\) is the probability of \(i\) finding a job. If \(i\) finds a job, she gets wage \(w_i\) and pays labor income taxes \(T_i\) and UI

\(^4\)The three proxies are average benefit in the state-year-cell, the maximum allowed benefit amount in the state-year-cell and the individuals predicted benefits from a UI calculator.

\(^5\)The key condition is that consumption and taxes have to enter all constraints in the same way. This is true e.g. for budget and borrowing constraints. It is necessary so that one can express the welfare effects of tax/benefit changes in terms of marginal consumption utilities. To derive the formula in the dynamic environment one considers the steady state of the dynamic economy in which a certain fraction of the people is unemployed and a certain fraction employed (but there are stochastics and dynamics at the individual level, i.e. the economy looks like an Aiyagari model. Then the problem is basically static again and the steps to derive the solution look very similar to the static model.
taxes $\tau$ — thus, she consumes $w_i - T_i - \tau$. If agent $i$ does not find a job, she receives unemployment benefit $b$ and consumes $b$. The agent maximizes expected utility. Hence, the problem of agent $i$ reads

$$\max_{p_i} p_i v_i (w_i - T_i - \tau) + (1 - p_i) u_i (b) - \psi_i (p_i)$$

taking $w_i$, $T_i$, $\tau$ and $b$ as given.

The government is utilitarian. The social welfare function therefore reads

$$SW = \sum_i Pr [i] \max_{p_i} p_i v_i (w_i - T_i - \tau) + (1 - p_i) u_i (b) - \psi_i (p_i).$$

The government has to obey the budget constraint

$$\sum_i Pr [i] (p_i (T_i + \tau) - (1 - p_i) b) = G$$

with $G$ being exogenously given. To derive a formula for the optimal $(\tau, b)$ (taking the labor income tax system as given), I consider a small balanced budget policy reform $(d\tau, db)$. The balanced budget requirement determines by how much taxes have to increase ($d\tau$) when increasing benefits by $db$:

$$d\tau = db \frac{1}{\sum_i Pr [i] p_i} \left[ \sum_i Pr [i] (1 - p_i) + \sum_i Pr [i] (T_i + \tau + b) \frac{d(1 - p_i)}{db} \right]$$

Here one can already see the main mechanism I consider: The behavioral responses $\frac{d(1 - p_i)}{db}$ are multiplied by $T_i$. That is, the behavioral response of high earners has a bigger impact on the increase in the UI tax. Using the envelope condition, the small policy reform causes a change in social welfare

$$dSW = -d\tau \left( \sum_i Pr [i] p_i v'_i \right) + db \left( \sum_i Pr [i] (1 - p_i) u'_i \right)$$

At the optimal $(\tau, b)$ the marginal change in social welfare has to be zero. Using $dSW = 0$, the expression for $d\tau$ as a function of $db$ and a bit of algebra one can derive that at the optimal $(\tau, b)$ it has to be that

$$\frac{E [u'_i] - E [v'_i]}{E [v'_i]} = \frac{\sum_i Pr [i] (1 - p_i) \frac{T_i + \tau + b}{b} \epsilon_i}{\sum_i Pr [i] (1 - p_i)}$$

(3.1)
with $E[u'_i]$ and $E[v'_i]$ denoting the average marginal utilities of the unemployed and the employed and $\varepsilon_i \equiv \frac{d(1-p_i)}{db} \frac{b}{1-p_i}$ denoting the elasticity of unemployment w.r.t. the benefit level for type $i$.

One can express $\tau$ as a function of $b$ and the exogenous parameters $T_i$ and $G$. With heterogeneous agents, however, that makes the expression just less transparent. For the whole paper I will focus on the RHS in the presence of heterogeneity and labor taxes. I will stay agnostic about how to calibrate the LHS — for a detailed discussion of how to think about $\frac{E[u'_i] - E[v'_i]}{E[v'_i]}$ when agents are heterogeneous see Andrews and Miller (2013).

$\tau$ is a unique function (defined by the government budget constraint) of $b$ and all exogenous parameters. The LHS is decreasing in $b$, while the RHS is increasing. Thus, for any set of exogenous parameters there is unique optimal $b$ (and, thus, $\tau$) implicitly defined by (3.1). Comparative statics can be performed in the following way: Any change in exogenous parameters that raises the LHS (e.g., curvature of the utility function) increases optimal benefit size, while any parameter change that raises the RHS (e.g., bigger behavioral responses) decreases optimal benefits (if both sides are affected, one has to use the implicit function theorem). Key point of this theoretical section is to demonstrate that — because of the progressive labor income tax system — the elasticity of average unemployment duration is not a sufficient statistic for the behavioral responses. It can be seen in (3.1), that heterogeneity in the individual elasticities $\varepsilon_i$ matters whenever $T_i$ is not constant across $i$. If $T_i = T$, however, the formula reduces to

$$\frac{E[u'_i] - E[v'_i]}{E[v'_i]} = \frac{T + \tau + b}{b} \varepsilon$$
with $\varepsilon \equiv \sum_i Pr[i](1-p_i) \varepsilon_i$. Hence, $\varepsilon$ becomes a sufficient statistic. $\varepsilon$ is the elasticity of aggregate unemployment w.r.t. benefits (one has to weight the $\varepsilon_i$ by the share of type $i$ among the unemployed which is $\frac{Pr[i](1-p_i)}{\sum_i Pr[i](1-p_i)}$).

The formula reduces to Lawson (2014) if there is only one type of agent. Then the formula becomes

$$\frac{E[u'] - E[v']}{E[v']} = \frac{T + \tau + b}{b} \varepsilon$$

Note that there is still an expectations operator, since if we allow for dynamics and stochastics the single agent can have different marginal utilities at different points in his life.

In this single agent case it is transparent to plug in the government budget constraint to get Lawsons main result

$$\frac{E[u'] - E[v']}{E[v']} = \frac{G + b}{b} \frac{1}{p} \varepsilon$$

This further reduces to the familiar Chetty (2006) result

$$\frac{E[u'] - E[v']}{E[v']} = \frac{1}{p} \varepsilon$$

if one assumes $G = 0$.

**Proportional benefits and UI taxes**

Now let $\tau$ and $b$ denote the tax and the replacement rate. This setup is identical to the constant benefit case except that UI taxes $\tau$ are replaced by $\tau w_i$ and UI benefits $b$

---

6This tells us that we should actually weight the behavioral response more heavily for those $i$ that stay in unemployment longer anyway. A representative sample of job losers might not give us the elasticity of average duration if baseline duration and $\varepsilon_i$ are correlated. I ignore this aspect in the empirical section — for the case of $i$ denoting income groups, it does not really matter, since the different income groups have similar average durations.
by \( bw_i \). Following the same steps as before one can derive that at the optimal policy

\[
\frac{E_{w_i} [u_i'] - E_{w_i} [v'_i]}{E_{w_i} [v'_i]} = \frac{\sum_i Pr [i] (1 - p_i) w_i \frac{T_i + r + b}{w_i} \varepsilon_i}{\sum_i Pr [i] (1 - p_i) w_i} \tag{3.2}
\]

with \( E_{w_i} [\cdot] \) denoting \( w_i \)-weighted expectations.

Again, \( \varepsilon \) is not a sufficient statistic, but heterogeneity matters. In the case of proportional benefits and taxes there are two factors causing the elasticity of high-earners to be more important than the elasticity of low earners.

First, even if \( T_i = 0 \) — that is, even if we ignore the labor income tax — the elasticities have to be weighted by \( w_i \). Second, the progressivity of the income tax implies that \( T_i \) is increasing in \( w_i \). The first effect has been studied before in Andrews and Miller (2013). Allowing for a progressive labor tax, however, adds the second effect, further strengthening the conclusion that the elasticity of the high earners is more important.

Intuitively, increasing benefits by one dollar has two effects which have to be traded off against each other.

First, increasing benefits means transferring one dollar of consumption from a state of high consumption (employment) to a state of low consumption (unemployment). If people are risk averse, this is enhancing social welfare, because marginal utility is higher in the unemployed state.

Second, however, higher unemployment benefits cause people to stay in unemployment longer. If a person stays in unemployment one month longer, this means one additional month of paying benefits and also one month less of getting labor and UI tax revenue. Because high earners pay higher labor income taxes when employed, the revenue leakage is bigger for high earners.

For the budget to remain balanced, the government has to increase UI taxes in order to compensate for the second effect. If the elasticity of average unemployment
Table 3.1: Summary Statistics — Above and Below Median Income

<table>
<thead>
<tr>
<th></th>
<th>Income Below Median</th>
<th>Income Above Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income</td>
<td>1205.2</td>
<td>3826.41</td>
</tr>
<tr>
<td>Benefit</td>
<td>710.76</td>
<td>1061.58</td>
</tr>
<tr>
<td>Age</td>
<td>36.44</td>
<td>41.74</td>
</tr>
<tr>
<td>Asset Income</td>
<td>10.44</td>
<td>36.72</td>
</tr>
<tr>
<td>Spell Duration</td>
<td>7.4</td>
<td>8</td>
</tr>
<tr>
<td>N</td>
<td>3097</td>
<td>3097</td>
</tr>
</tbody>
</table>

Reported numbers are sample means for all job losers. All units are in months. Income refers to the monthly labor income prior to the job loss.

duration is mainly driven by the high earners, the second effect is bigger and therefore the necessary tax increase is larger. Thus, providing benefits is less attractive. If one abstracts from the existence of a progressive tax system, one fails to capture this effect.

### 3.3 Empirical Analysis

The last section demonstrated that heterogeneity in the behavioral response across income groups is crucial for designing optimal UI. In this section I empirically investigate the quantitative importance of this heterogeneity.

**Data**

I use data from the Survey of Income and Program Participation (SIPP) spanning the years 1989 to 2012. I restrict my sample to all male job losers (since they are more attached to the labor force), for whom I observe pre-unemployment earnings and who report receiving unemployment benefits after losing their job. This leaves me with 6,194 unemployment spells. See Appendix for more details on the sample construction and on how I define unemployment spells, benefits, etc. Table 3.1 provides summary statistics.

**Specification**
Let $h_{i,t}$ denote the hazard rate from unemployment for individual $i$ in month $t$ of his spell (I only include $t \leq 12$). I estimate the following specification

$$h_{i,t} = \alpha_t + \beta_1 I \{y_i \leq \text{Median}\} I \{t \leq 6\} \ln b_i + \beta_2 I \{y_i \leq \text{Median}\} I \{t > 6\} \ln b_i + \beta_1 I \{y_i > \text{Median}\} I \{t \leq 6\} \ln b_i + \beta_2 I \{y_i > \text{Median}\} I \{t > 6\} \ln b_i + \gamma X_{i,t} + \varepsilon_{i,t}$$

with $\alpha_t$ being fixed effects for month $t$ of the spell, $I \{y_i \leq \text{Median}\}$ and $I \{y_i > \text{Median}\}$ being indicators for people with pre-unemployment income below or above median, $I \{t \leq 6\}$ and $I \{t > 6\}$ being indicators for the month of spell being earlier or later than month six, $\ln b_i$ the natural log of the unemployment benefit of $i$ and $X_{i,t}$ being individual and state level controls as well as state and year fixed effects.

That is, I estimate the effect of benefits on hazard rates separately for low and for high earners. Moreover, I allow the effect to be different for the later months of the spell. I choose a linear specification for the hazard so I can use an IV strategy.

**Identification Strategy**

I instrument for benefits with the average benefit in the state-year cell, excluding the benefit of the individual I instrument for. I use the IV procedure in order to isolate benefit variation driven by changes in the generosity of the UI system. There are threats to identification. For instance, a recession might change the composition of job losers which in turn changes average benefits. Since a recession also effects hazard rates via labor demand, the exclusion restriction could be violated. To the degree that the bias effects low and high earners similarly, however, this does not change my conclusions concerning heterogeneity. Moreover, I show that the qualitative conclusions are fairly robust to adding flexible individual and state level controls.

**Results**
Since I interact the endogenous variable ln $b_i$ with several of the covariates, I need four instruments. I achieve this by also interacting the instrument with these covariates in the same way. In all IV specifications I fully control for the indicator variables I interact the instrument with, since the indicators could be correlated with the error term.

All standard errors are clustered at the state level. In all specifications I control for the so-called seam effect: Because households are interviewed every four months, there is an unusually large number of changes reported at the beginning of the four month period they are interviewed about. I control for this by including a dummy that takes the value one whenever a month is the first month of the interview period.

Table 3.2 presents the OLS results, followed by the IV — once with and once without individual and state level controls. Individual controls include a 20-piece linear spline in previous earnings, number of children, indicators for marriage status, race and education. State level controls include the number of spells in the state-year cell as well as average pre-unemployment income of all new spells in the state-year cell.

All first stage regressions in all IV specifications have F statistics far beyond 10. The number of observations in the regression table is not the number of spells, but refers to the overall number of months in which I observe an unemployed individual either exiting or not exiting.

The OLS results in the first column suggest significant negative effects of benefits on the probability of exiting unemployment in the first six months after losing a job. Note that the OLS specification controls flexibly for individual determinants of benefit size like previous earnings, number of children, etc. The estimates for the effect of benefits are smaller for low earners, but equality of coefficients cannot be rejected.
Table 3.2: Response of Unemployment Hazards to Benefits by Previous Income

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>IV</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Inc x 0-6 Months x ln(b)</td>
<td>-0.0253***</td>
<td>-0.0394**</td>
<td>-0.0396**</td>
</tr>
<tr>
<td></td>
<td>(-4.84)</td>
<td>(-2.97)</td>
<td>(-2.80)</td>
</tr>
<tr>
<td>Low Inc x 7-12 Months x ln(b)</td>
<td>0.0128</td>
<td>0.0110</td>
<td>0.0114</td>
</tr>
<tr>
<td></td>
<td>(2.00)</td>
<td>(0.52)</td>
<td>(0.55)</td>
</tr>
<tr>
<td>High Inc x 0-6 Months x ln(b)</td>
<td>-0.035***</td>
<td>-0.040*</td>
<td>-0.047**</td>
</tr>
<tr>
<td></td>
<td>(-4.59)</td>
<td>(-2.28)</td>
<td>(-2.68)</td>
</tr>
<tr>
<td>High Inc x 7-12 Months x ln(b)</td>
<td>0.0044</td>
<td>0.0037</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td>(0.41)</td>
<td>(0.15)</td>
<td>(-0.02)</td>
</tr>
<tr>
<td>Ind. and State Controls</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>State and Year Fixed Effects</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Month of Spell Fixed Effects</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>N</td>
<td>31221</td>
<td>31221</td>
<td>31221</td>
</tr>
</tbody>
</table>

* t stats in parantheses. Observation count is the number of months for which I observe an individual either exiting or not exiting unemployment. Robust standard errors clustered at the state level. Individual controls include a 20-piece linear spline in previous earnings, number of children, indicators for marriage status, race and education. State level controls include the number of spells in the state-year cell as well as average pre-unemployment income of all new spells in the state-year cell. All specifications include controls for the seam effect as well as full controls for low/high income and first/last six months. *: p < 0.05, **: p < 0.01, ***: p < 0.001

The IV specifications in the next two columns confirm this. The IV results are bigger than the OLS results, suggesting some bias and/or measurement error in the OLS specification. Adding controls to the IV specification changes the coefficients somewhat, but the qualitative conclusions remain. The results of the last specification (IV with controls) imply an elasticity of unemployment duration w.r.t. benefit level of .15 for low earners and .22 for high earners. This is smaller than what has previously been found in the literature.

The fact that low earners are not reacting more strongly than high earners is puzzling given the analysis from Chetty (2008) who finds that individuals with little assets react more strongly. Since assets and income are correlated, one would expect those with lower incomes to react more strongly. For Table 3.3 I split the sample into those reporting positive asset income and those reporting zero asset income. Otherwise the analysis is the same as before.
<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>IV</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No Asset x 0-6 Months x ln(b)</strong></td>
<td>-0.0256***</td>
<td>-0.0344**</td>
<td>-0.0419**</td>
</tr>
<tr>
<td></td>
<td>(-5.52)</td>
<td>(-2.94)</td>
<td>(-3.22)</td>
</tr>
<tr>
<td><strong>No Asset x 7-12 Months x ln(b)</strong></td>
<td>0.0067</td>
<td>0.0157</td>
<td>0.0063</td>
</tr>
<tr>
<td></td>
<td>(1.00)</td>
<td>(0.79)</td>
<td>(0.30)</td>
</tr>
<tr>
<td><strong>Asset x 0-6 Months x ln(b)</strong></td>
<td>-0.036***</td>
<td>-0.038*</td>
<td>-0.043*</td>
</tr>
<tr>
<td></td>
<td>(-5.00)</td>
<td>(-2.08)</td>
<td>(-2.33)</td>
</tr>
<tr>
<td><strong>Asset x 7-12 Months x ln(b)</strong></td>
<td>0.014</td>
<td>0.001</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td>(1.55)</td>
<td>(0.04)</td>
<td>(-0.05)</td>
</tr>
<tr>
<td><strong>Ind. and State Controls</strong></td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td><strong>State and Year Fixed Effects</strong></td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td><strong>Month of Spell Fixed Effects</strong></td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>31221</td>
<td>31221</td>
<td>31221</td>
</tr>
</tbody>
</table>

* t stats in parentheses. Observation count is the number of months for which I observe an individual either exiting or not exiting unemployment. Robust standard errors clustered at the state level. Individual controls include a 20-piece linear spline in previous earnings, number of children, indicators for marriage status, race and education. State level controls include the number of spells in the state-year cell as well as average pre-unemployment income of all new spells in the state-year cell. All specifications include controls for the seam effect as well as full controls for low/high income and first/last six months. 'No Asset' refers to that person having no asset income and 'Asset' to that person having asset income. *: p < 0.05, **: p < 0.01, ***: p < 0.001

Puzzlingly, I do not find that the group with zero asset income reacts more strongly to higher benefits. This result is unintuitive and in sharp contrast to Chetty (2008). Still, it is worth mentioning that this paper is not the only paper finding similar elasticities of unemployment w.r.t. UI for low and high earners. In Schmieder, von Wachter, and Bender (2012) the authors investigate in a RD design the effect of maximum UI duration on hazard rates and find that the effects are basically identical for people with low and high education.

**Connecting with the Theory**
The results of the previous section suggest that $\varepsilon_i = \varepsilon$ is a good approximation. This implies for the uniform benefit/tax case that at the optimal policy

$$
\frac{E[u_i'] - E[v_i']}{E[v_i']} = \varepsilon \frac{\sum_i Pr[i] (1 - p_i) \frac{T_i + \tau + b}{b}}{\sum_i Pr[i] (1 - p_i)}
$$

and for the proportional case

$$
\frac{E_{w_i}[u_i'] - E_{w_i}[v_i']}{E_{w_i}[v_i']} = \varepsilon \frac{\sum_i Pr[i] (1 - p_i) w_i \frac{T_i + \tau + b}{b}}{\sum_i Pr[i] (1 - p_i) w_i}
$$

(3.3)

Thus, the interaction between $\varepsilon_i$ and the tax system does not matter in practice due to the homogeneity in $\varepsilon_i$ across income groups. Despite the fact that the main mechanism I consider turns out to be unimportant empirically, the theoretical results are useful: The formula sheds new light on how to correctly calibrate the level of the labor tax that enters the optimal UI formula in Lawson (2014). He studies optimal UI in the presence of a labor income tax and one representative agent and calibrates the formula with the average labor tax in the population. My results show that the correct tax rate for Lawson’s formula is the average labor tax among the unemployed — i.e., the labor tax rate weighted by $Pr[i] (1 - p_i)$. Intuitively, if a certain group is never unemployed then their tax rate does not matter for the optimal UI system. Since high earners are less likely to be unemployed, taking the labor income tax into account should effect the optimal level of UI less than in Lawson (2014).

3.4 Conclusion

In this paper I derive a new formula for optimal UI with heterogeneous agents that remains valid in the presence of a progressive labor tax system. I show that with a progressive labor tax the elasticity of average unemployment duration with respect to
the benefit level seizes to be a sufficient statistic for the behavioral response. Instead, heterogeneity in the behavioral response across income groups matters.

I estimate the elasticity of unemployment duration for different income groups and find no support for meaningful heterogeneity. This implies that formulas with representative agents can provide good approximations. Still my derivations shed some light on how to calibrate those formulas correctly: When studying the interaction between the UI system and labor taxes, what matters is the labor taxes that the average unemployed would pay if she were employed — not the average labor taxes of the overall population.

References


Appendix

3.A Sample Construction

I use the SIPP Uniform extracts provided by the Center for Economic and Policy Research in Washington 2014 Version 2.1.7 I consider all male job losers in the SIPP between 1989 and 2012. A job loser is defined as someone who reports a change in his employment status from working for being with a job for at least a week (variable esr = 1 to 5) to not having a job (variable esr = 6 to 8). Moreover, I require the person to report earnings (variables earn1 and earn2) before the job loss and receipt of positive UI benefits (variable ui_a) afterwards. I define pre-unemployment earnings as the average earnings of the 4 months before the job loss (unless somebody reports less than 4 months of earnings, in which I case I take the average of the reported earnings).

I define UI benefits as the average of ui_a for the first six months of unemployment excluding the months in which the person reports zero UI benefit or is not unemployed for the full month. The qualitative results are robust to the details of the definitions.